First-order Methods for Affinely Constrained Composite Non-convex Non-smooth Problems: Lower Complexity Bound and Near-optimal Methods

Wei Liu · Qihang Lin · Yangyang Xu

July 18, 2023

Abstract Many recent studies on first-order methods (FOMs) focus on *composite non-convex non-smooth* optimization with linear and/or nonlinear function constraints. Upper (or worst-case) complexity bounds have been established for these methods. However, little can be claimed about their optimality as no lower bound is known, except for a few special *smooth non-convex* cases. In this paper, we make the first attempt to establish lower complexity bounds of FOMs for solving a class of composite non-convex non-smooth optimization with linear constraints. Assuming two different first-order oracles, we establish lower complexity bounds of FOMs to produce a (near) ϵ-stationary point of a problem (and its reformulation) in the considered problem class, for any given tolerance $\epsilon > 0$. In addition, we present an inexact proximal gradient (IPG) method by using the more relaxed one of the two assumed first-order oracles. The oracle complexity of the proposed IPG, to find a (near) ϵ-stationary point of the considered problem and its reformulation, matches our established lower bounds up to a logarithmic factor. Therefore, our lower complexity bounds and the proposed IPG method are almost non-improvable.

Keywords non-convex optimization, non-smooth optimization, first-order methods, proximal gradient method, information-based complexity, lower complexity bound, worst-case complexity

Mathematics Subject Classification 90C26, 90C06, 90C60, 49M37, 68Q25, 65Y20

1 Introduction

First-order methods (FOMs) have attracted increasing attention because of their efficiency in solving large-scale problems arising from machine learning and other areas. The recent studies on FOMs have focused on non-convex problems, and one of the actively studied topics is the oracle complexity for finding a near-stationary point under various assumptions. In this paper, we explore this topic for problems with a composite

W. Liu, Y. Xu
Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY
E-mail: {liuw16, xuy21}@rpi.edu
Q. Lin
Department of Business Analytics, University of Iowa, Iowa City
E-mail: qihanglin@uiowa.edu

non-convex non-smooth objective function and linear equality constraints, formulated as

$$\min_{\mathbf{x}\in\mathbb{R}^d} F_0(\mathbf{x}) := f_0(\mathbf{x}) + g(\mathbf{x}), \quad \text{s.t. } \mathbf{A}\mathbf{x} + \mathbf{b} = \mathbf{0}. \tag{P}$$

Here, $\mathbf{A} \in \mathbb{R}^{n\times d}$, $\mathbf{b} \in \mathbb{R}^n$, $f_0 : \mathbb{R}^d \to \mathbb{R}$ is smooth, and $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a proper lower semicontinuous convex function but potentially non-smooth. We assume the following properties throughout this paper.

Assumption 1 *The following statements hold.*

(a) ∇f_0 *is* L_f*-Lipschitz continuous, i.e.,* $\|\nabla f_0(\mathbf{x}) - \nabla f_0(\mathbf{x}')\| \le L_f \|\mathbf{x} - \mathbf{x}'\|, \quad \forall \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$.
(b) $\inf_{\mathbf{x}} F_0(\mathbf{x}) > -\infty$.
(c) $g(\mathbf{x}) = \bar{g}(\bar{\mathbf{A}}\mathbf{x} + \bar{\mathbf{b}})$, *where* $\bar{\mathbf{A}} \in \mathbb{R}^{\bar{n}\times d}$, $\bar{\mathbf{b}} \in \mathbb{R}^{\bar{n}}$ *and* $\bar{g} : \mathbb{R}^{\bar{n}} \to \mathbb{R} \cup \{+\infty\}$ *is a proper lower semicontinuous convex function but potentially non-smooth.*

Due to non-convexity, computing or even approximating a global optimal solution for problem (P) is intractable in general [25]. Hence, we focus on using an FOM to find a (near) ϵ-stationary point of problem (P) for a given tolerance $\epsilon > 0$; see Definition 1. An FOM finds a (near) ϵ-stationary point by querying information from some oracles, which typically dominates the runtime of the method, so its efficiency can be measured by the number of oracles it queries, which is defined as the method's *oracle complexity*. The goal of this paper is to establish a lower bound for the oracle complexity of a class of FOMs to find a (near) ϵ-stationary point of problem (P) and, additionally, to present an FOM that can nearly achieve this lower complexity bound.

Undoubtedly, an algorithm's oracle complexity depends on what oracle information the algorithm can utilize and what operations it can perform by using the oracle. Since we focus on FOMs, we assume that a first-order oracle is accessible and each generated iterate is a certain combination of the oracle information. More specifically, we make the following assumption that is standard for an FOM to solve problem (P).

Assumption 2 *There is an oracle such that given any* $\mathbf{x}$ *and* $\boldsymbol{\xi}$ *in* $\mathbb{R}^d$*, it can return* $(\mathbf{A}^\top\mathbf{b}, \nabla f_0(\mathbf{x}), \mathbf{A}^\top\mathbf{A}\mathbf{x})$ *and* $\mathbf{prox}_{\eta g}(\boldsymbol{\xi})$ *for any* $\eta > 0$*. In addition, the sequence* $\left\{\mathbf{x}^{(t)}\right\}_{t=0}^{\infty}$ *generated by the underlying algorithm satisfies that, for any* $t \ge 1$, $\mathbf{x}^{(t)} \in \mathbf{span}\left(\left\{\boldsymbol{\xi}^{(t)}, \boldsymbol{\zeta}^{(t)}\right\}\right)$*, where*

$$\boldsymbol{\xi}^{(t)} \in \mathbf{span}\left(\left\{\mathbf{A}^\top\mathbf{b}\right\} \bigcup \cup_{s=0}^{t-1} \left\{\mathbf{x}^{(s)}, \nabla f_0(\mathbf{x}^{(s)}), \mathbf{A}^\top\mathbf{A}\mathbf{x}^{(s)}\right\}\right), \ \boldsymbol{\zeta}^{(t)} \in \left\{\mathbf{prox}_{\eta g}(\boldsymbol{\xi}^{(t)}) \mid \eta > 0\right\}.$$

Here, $\mathbf{span}(\mathcal{S})$ represents the set of all linear combinations of finitely many vectors in a set $\mathcal{S}$. Under the linear-span assumption above, an algorithm can access ∇f_0 at any historical solutions and can compute matrix-vector multiplications with $\mathbf{A}$ and $\mathbf{A}^\top$ as well as the proximal mapping of g, i.e.,

$$\mathbf{prox}_{\eta g}(\mathbf{x}) := \arg\min_{\mathbf{x}'} \left\{ g(\mathbf{x}') + \tfrac{1}{2\eta}\|\mathbf{x}' - \mathbf{x}\|^2 \right\} \tag{1.1}$$

for any $\mathbf{x} \in \mathbb{R}^d$ and $\eta > 0$. We say $\mathbf{x}^{(t)}$ is generated by the *t-th iteration*[1] of the algorithm.

It should be noted that the above linear-span assumption allows the algorithm to compute the proximal mapping of g but does not permit the projection onto the affine set $\{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{A}\mathbf{x} + \mathbf{b} = \mathbf{0}\}$. It is also worth noting here that computing the proximal mapping of g is more difficult than computing a subgradient of g. In fact, if we disallow computing the proximal mapping of g but instead give the algorithm access to its

[1] Without further specification, one iteration of an FOM will call the oracle once. We will specify the type of iteration if this does not hold, e.g., by outer- or inner-iteration.

subgradients, the algorithm is limited to the class of subgradient methods for non-smooth optimization, in which case finding an ϵ-stationary point is impossible with finite oracle complexity [39].

FOMs under Assumption 2 have been developed with theoretically proved oracle complexity for finding an ϵ-stationary point of problem (P) under different settings. These results are stated as *upper bounds* for the maximum number of oracles those algorithms require to reach an ϵ-stationary point. In two special cases of problem (P), some existing algorithms' oracle complexity is known to be non-improvable (also called optimal) because the complexity matches, up to constant factors, a theoretical *lower bound* of oracle complexity [27], which is the minimum number of oracles an algorithm in a class needs to find an ϵ-stationary point. These two cases are as follows.

1. When the linear constraint $\mathbf{A}\mathbf{x} + \mathbf{b} = \mathbf{0}$ does not exist, problem (P) becomes a composite non-smooth non-convex optimization problem
$$\min_{\mathbf{x}} f_0(\mathbf{x}) + g(\mathbf{x}), \tag{1.2}$$
for which the proximal gradient method finds an ϵ-stationary point by $O(L_f\epsilon^{-2})$ first-order oracles [28]. This oracle complexity cannot be improved because it matches the lower bound provided in [4,5].
2. When $g \equiv 0$, problem (P) becomes a linear equality-constrained smooth non-convex optimization problem
$$\min_{\mathbf{x}} f_0(\mathbf{x}), \quad \text{s.t. } \mathbf{A}\mathbf{x} + \mathbf{b} = \mathbf{0}. \tag{1.3}$$
It is known that, if exact projection onto the feasible set $\{\mathbf{x} \mid \mathbf{A}\mathbf{x}+\mathbf{b} = \mathbf{0}\}$ is allowed, the projected gradient method can find an ϵ-stationary point in $O(L_f\epsilon^{-2})$ iterations. When exact projection is prohibited, one can perform inexact projection through the matrix-vector multiplications with $\mathbf{A}$ and $\mathbf{A}^\top$, which are allowed under Assumption 2. This way, with $O(\kappa(\mathbf{A})\log(\epsilon^{-1}))$ multiplications, one can project any point to the feasible set with a poly(ϵ) error[2]. Here, $\kappa(\mathbf{A})$ is the condition number of $\mathbf{A}$ defined as
$$\kappa(\mathbf{A}) := \sqrt{\frac{\lambda_{\max}(\mathbf{A}\mathbf{A}^\top)}{\lambda_{\min}(\mathbf{A}\mathbf{A}^\top)}},$$
where $\lambda_{\min}(\mathbf{A}\mathbf{A}^\top)$ and $\lambda_{\max}(\mathbf{A}\mathbf{A}^\top)$ are the smallest positive and the largest eigenvalues of $\mathbf{A}\mathbf{A}^\top$, respectively. Using this inexact projection as a subroutine, one can easily develop an inexact projected gradient method to this special case of (P) with oracle complexity of $O(\kappa(\mathbf{A})\log(\epsilon^{-1})L_f\epsilon^{-2})$. This complexity matches the lower bound $O(\kappa(\mathbf{A})L_f\epsilon^{-2})$ in [35] up to a logarithmic factor and thus is nearly optimal.

These two examples suggest that a lower bound of oracle complexity is valuable as it informs algorithm designers which algorithms can be potentially improved for higher efficiency and which are non-improvable without additional assumptions. In literature, there exist multiple algorithms with the theoretically proved oracle complexity for problem (P) or a more general problem. We refer readers to [2,12,16,17,24,31,40,41] for problems with affine constraints and [6,7,10,18,20,34,36,37] for problems with nonlinear constraints. However, to the best of our knowledge, there is no lower bound of the oracle complexity for finding a (near) ϵ-stationary point of problem (P) under Assumption 2. Although there are lower bounds established for decentralized smooth optimization [35], linearly constrained smooth optimization [35], and non-convex strongly-concave min-max problems [19,43], these lower bounds either cannot be applied to problem (P) or can only be applied to a special case of problem (P), so they will not be tight enough as we will show later in this paper. To fill this gap in literature, we pose the following question:

What is the minimum oracle complexity for a first-order method satisfying a linear-span assumption, e.g., Assumption 2, to find a (near) ϵ-stationary point of problem (P) *that satisfies Assumption 1?*

[2] poly(ϵ) denotes a polynomial of ϵ.

1.1 Contributions

Our first major contribution is to provide an answer to the question above by establishing a lower bound of the oracle complexity of FOMs for finding a (near) ϵ-stationary point of problem (P). This is achieved by adapting the worst-case instance in [35] for affinely constrained smooth optimization to the affinely constrained structured non-smooth problem (P). In particular, we use a subset of the affine constraints of the instance in [35] to design the regularized term $g(\mathbf{x})$ in problem (P) and leave the remaining constraints in that instance as affine constraints in problem (P). We show that under Assumption 2, any algorithm needs at least $O(\kappa([\bar{\mathbf{A}};\mathbf{A}])L_f\Delta_{F_0}\epsilon^{-2})$ iterations/oracles to find a (near) ϵ-stationary point of the instance we design; see Theorem 2.1. Here

$$\Delta_{F_0} := F_0(\mathbf{x}^{(0)}) - \inf_{\mathbf{x}} F_0(\mathbf{x}),\ [\bar{\mathbf{A}};\mathbf{A}] := \begin{bmatrix}\bar{\mathbf{A}}\\ \mathbf{A}\end{bmatrix} \text{ and } \kappa([\bar{\mathbf{A}};\mathbf{A}]) := \sqrt{\frac{\lambda_{\max}([\bar{\mathbf{A}};\mathbf{A}][\bar{\mathbf{A}};\mathbf{A}]^\top)}{\lambda_{\min}([\bar{\mathbf{A}};\mathbf{A}][\bar{\mathbf{A}};\mathbf{A}]^\top)}}. \tag{1.4}$$

Our lower complexity bound for (P) can be viewed as a generalization of the lower bound $O(\kappa(\mathbf{A})L_f\epsilon^{-2})$ for problem (1.3) in [35]. Our result provides a new insight that the difficulty of finding a (near) ϵ-stationary point of problem (P) depends on the interaction between the affine constraints and the non-smooth regularization term characterized by $\kappa([\bar{\mathbf{A}};\mathbf{A}])$.

Under Assumption 2, an algorithm is allowed to call $\mathbf{prox}_{\eta g}(\cdot)$, which may not be easy to compute, especially when $\bar{\mathbf{A}}$ in Assumption 1 is a generic matrix without a special structure. On the contrary, when $\bar{g}$ in Assumption 1 is simple enough so that $\mathbf{prox}_{\eta\bar{g}}(\cdot)$ can be computed easily, one can apply a variable splitting technique such as the one used in the alternating direction method of multipliers (ADMM) to find a (near) ϵ-stationary point of problem (P). To do so, we introduce a new variable $\mathbf{y}\in\mathbb{R}^{\bar{n}}$ and reformulate problem (P) to its Splitting Problem

$$\min_{\mathbf{x}\in\mathbb{R}^d,\mathbf{y}\in\mathbb{R}^{\bar{n}}} F(\mathbf{x},\mathbf{y}) := f_0(\mathbf{x}) + \bar{g}(\mathbf{y}), \quad \text{s.t.} \quad \mathbf{A}\mathbf{x}+\mathbf{b}=0,\ \mathbf{y}=\bar{\mathbf{A}}\mathbf{x}+\bar{\mathbf{b}}. \tag{SP}$$

For $t=0,1,\ldots$, let $(\mathbf{x}^{(t)},\mathbf{y}^{(t)})$ be the solution generated in the t-th iteration by an algorithm for solving (SP). We assume they satisfy the following linear-span assumption (see [4, 30]), instead of Assumption 2.

Assumption 3 *There is an oracle such that given any $\mathbf{x}\in\mathbb{R}^d$ and $\mathbf{y},\boldsymbol{\xi}$ in $\mathbb{R}^{\bar{n}}$, it can return $(\bar{\mathbf{b}},\mathbf{A}^\top\mathbf{b},\bar{\mathbf{A}}^\top\bar{\mathbf{b}})$, $(\nabla f_0(\mathbf{x}),\bar{\mathbf{A}}\mathbf{x},\mathbf{A}^\top\mathbf{A}\mathbf{x},\bar{\mathbf{A}}^\top\bar{\mathbf{A}}\mathbf{x},\bar{\mathbf{A}}^\top\mathbf{y})$, and $\mathbf{prox}_{\eta\bar{g}}(\boldsymbol{\xi})$ for any $\eta>0$. In addition, the sequence $\left\{(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\right\}_{t=0}^{\infty}$ generated by the underlying algorithm satisfies that, for all $t\geq 1$,*

$$\begin{aligned}
&\mathbf{x}^{(t)} \in \mathbf{span}\left(\left\{\mathbf{A}^\top\mathbf{b},\bar{\mathbf{A}}^\top\bar{\mathbf{b}}\right\}\bigcup\cup_{s=0}^{t-1}\left\{\mathbf{x}^{(s)},\nabla f_0(\mathbf{x}^{(s)}),\mathbf{A}^\top\mathbf{A}\mathbf{x}^{(s)},\bar{\mathbf{A}}^\top\bar{\mathbf{A}}\mathbf{x}^{(s)},\bar{\mathbf{A}}^\top\mathbf{y}^{(s)}\right\}\right),\\
&\mathbf{y}^{(t)} \in \mathbf{span}\left(\left\{\boldsymbol{\xi}^{(t)},\boldsymbol{\zeta}^{(t)}\right\}\right), \ \textit{where}\\
&\boldsymbol{\xi}^{(t)} \in \mathbf{span}\left(\left\{\bar{\mathbf{b}}\right\}\bigcup\cup_{s=0}^{t-1}\left\{\mathbf{y}^{(s)},\bar{\mathbf{A}}\bar{\mathbf{A}}^\top\mathbf{y}^{(s)},\bar{\mathbf{A}}\mathbf{x}^{(s)}\right\}\right),\ \boldsymbol{\zeta}^{(t)}\in\left\{\mathbf{prox}_{\eta\bar{g}}(\boldsymbol{\xi}^{(t)})\mid\eta>0\right\}.
\end{aligned}$$

We say $(\mathbf{x}^{(t)},\mathbf{y}^{(t)})$ is generated by the t-th iteration of the algorithm. Our second major contribution is to show that under Assumption 3, the minimum oracle complexity is $O(\kappa([\bar{\mathbf{A}};\mathbf{A}])L_f\Delta_F\epsilon^{-2})$ for an algorithm to find an ϵ-stationary point of problem (SP) that satisfies Assumption 1(a,c) and the condition

$$\Delta_F := F(\mathbf{x}^{(0)},\mathbf{y}^{(0)}) - \inf_{\mathbf{x},\mathbf{y}} F(\mathbf{x},\mathbf{y}) < \infty.$$

The lower bound can be proved by using the same worst-case instance as that we will use to prove the lower bound for problem (P). Moreover, under Assumption 3, the minimum oracle complexity for an algorithm to find a near ϵ-stationary point of problem (P) is $O(\kappa([\bar{\mathbf{A}};\mathbf{A}])L_f\Delta_{F_0}\epsilon^{-2})$, the same as that we establish under Assumption 2 which allows the more expensive operator $\mathbf{prox}_{\eta g}$.

Given a lower bound of oracle complexity, a critical question is whether the bound is tight, or equivalently, whether it can be achieved by an algorithm that meets the underlying assumptions upon which the lower bound is established. To address this question and shed light on the tightness of our lower bounds, we make a third significant contribution by introducing a novel method called the *inexact proximal gradient* (IPG) method. This method is specifically designed to solve problem (SP), while satisfying the conditions outlined in Assumption 3. Remarkably, IPG is able to achieve an oracle complexity that matches our established lower bound $O(\kappa([\bar{\mathbf{A}};\mathbf{A}])L_f\Delta_F\epsilon^{-2})$ up to logarithmic factors for solving problem (SP). That is, the lower complexity bound is tight for problem (SP). Meanwhile, IPG is able to find a near ϵ-stationary point of problem (P) by $O(\kappa([\bar{\mathbf{A}};\mathbf{A}])L_f\Delta_{F_0}\epsilon^{-2})$ iterations/oracles, up to logarithmic factors. This means that the lower complexity bound is also tight for problem (P) under Assumption 3.

1.2 Related Work

The *proximal gradient* (PG) method can find an ϵ-stationary point of the composite non-smooth non-convex problem (1.2) within $O(L_f\epsilon^{-2})$ [28] iterations, which matches the lower bound in [4, 5]. When there is no affine constraints and $g \equiv 0$ in problem (P), our lower-bound complexity result is reduced to the lower bound in [4].

For the affinely constrained non-convex smooth problem (1.3), Sun and Hong [35] show a lower-bound complexity of $O(\kappa(\mathbf{A})L_f\epsilon^{-2})$ for finding an ϵ-stationary point. In the same paper, they give an FOM that achieves this lower bound. When $g \equiv 0$, our complexity lower bound for problem (P) is reduced to their lower bound.

Before our work, there only exist upper bounds of oracle complexity for finding a (near) ϵ-stationary point of (P) and (SP). For instance, Kong et al. [17] develop a quadratic-penalty accelerated inexact proximal point method that finds an ϵ-stationary point of problem (P) with oracle complexity $O(\epsilon^{-3})$. Lin et al. [20] study a method similar to [17] and show that oracle complexity of $O(\epsilon^{-5/2})$ is sufficient. The *augmented Lagrangian method* (ALM) is another effective approach for problem (P). The oracle complexity of ALM for problem (P) has been studied by [13, 14, 24, 41]. For example, the inexact proximal accelerated ALM by [24] achieves oracle complexity of $O(\epsilon^{-5/2})$ and, in a special case where $g(\mathbf{x})$ is the indicator function of a polyhedron, the smoothed proximal ALM by [41] improves the complexity to $O(\epsilon^{-2})$. ADMM is an effective algorithm for optimization with a separable structure like that in problem (SP). ADMM and its variants have been studied by [9, 15, 16, 23, 38, 40, 41] for constrained non-convex optimization problems including problem (SP). For example, it is shown by [9,16,38] that ADMM finds an ϵ-stationary point of problem (SP) with oracle complexity of $O(\epsilon^{-2})$.

The aforementioned methods for problem (P) all satisfy Assumption 2, and the aforementioned methods for problem (SP) all satisfy Assumption 3. However, the oracle complexity of those methods for finding an ϵ-stationary point either does not match or is not comparable with our lower-bound complexity for the corresponding problems. Specifically, the oracle complexity of ADMM in [38] for solving problem (SP) depends on the Kurdyka-Łojasiewicz (KŁ) coefficient, which is not directly comparable with our lower bound. The oracle complexity $O(\kappa^2([\mathbf{A};\bar{\mathbf{A}}])L_f^2\Delta_F\epsilon^{-2})$ of ADMM for solving problem (SP) is presented in [9], under the assumption that $[\bar{\mathbf{A}};\mathbf{A}]$ has a full-row rank. The results of [16] of ADMM are only applicable to problems (P) and (SP) with a separable structure and $\bar{\mathbf{A}} = \mathbf{I}_d$ or $g \equiv 0$, for which case their oracle complexity is $O(\kappa^2(\mathbf{A})L_f^2\varDelta\epsilon^{-2})$ with $\varDelta = \varDelta_{F_0}$ or $\varDelta_F$. Zhang and Luo in [41] study the complexity of ALM for problem (P) when g is the indicator function of a polyhedral set and the exact projection onto the polyhedral set can be computed. They obtained complexity of $O(\widehat{\kappa}^2 L_f^3 \varDelta_{F_0}\epsilon^{-2})$, where $\widehat{\kappa}$ is a joint condition number of the equality $\mathbf{Ax} + \mathbf{b} = \mathbf{0}$ and the inequality defining the polyhedral set. When $g \equiv 0$, their complexity is reduced to $O(\kappa^2(\mathbf{A})L_f^3\varDelta_{F_0}\epsilon^{-2})$. Zhang et al. [42] further extend [40,41] to problems with nonlinear convex inequality constraints. However, their algorithm requires exact projection to the set defined by the inequality constraints, which is impractical for many applications and does not satisfy Assumption 2.

1.3 Notations and Definitions

For any $a \in \mathbb{R}$, we use $\lceil a \rceil$ to denote the smallest integer that is no less than a. $\mathbf{Null}(\mathbf{H})$ represents the null space of a matrix $\mathbf{H}$. For any vector $\mathbf{z}$, $[\mathbf{z}]_j$ denotes its j-th coordinate. We denote $\mathbf{1}_p$ for an all-one vector in $\mathbb{R}^p$, and $\mathbf{0}$ to represent an all-zero vector when its dimension is clear from the context. $\mathbf{I}_p$ denotes a $p \times p$ identity matrix and

$$\mathbf{J}_p := \begin{bmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{bmatrix} \in \mathbb{R}^{(p-1)\times p}. \tag{1.5}$$

A vector $\mathbf{x}$ is said ω-close to another vector $\widehat{\mathbf{x}}$ if $\|\mathbf{x} - \widehat{\mathbf{x}}\| \le \omega$. For any set $\mathcal{X}$, we denote $\iota_{\mathcal{X}}$ as its indicator function, i.e., $\iota_{\mathcal{X}}(\mathbf{x}) = 0$ if $\mathbf{x} \in \mathcal{X}$ and $+\infty$ otherwise. We use $\otimes$ for the Kronecker product, $\mathbf{co}(\mathcal{S})$ for the convex hull of a set $\mathcal{S}$ and $\overline{\mathbf{co}}(\mathcal{S})$ for the closure of $\mathbf{co}(\mathcal{S})$. For a function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, its directional derivative at $\mathbf{x}$ along a direction $\mathbf{v} \in \mathbb{R}^d$ is defined as

$$f'(\mathbf{x};\mathbf{v}) = \lim_{s\downarrow 0} \frac{f(\mathbf{x} + s\mathbf{v}) - f(\mathbf{x})}{s}, \tag{1.6}$$

where $s \downarrow 0$ means $s \to 0$ and $s > 0$. ∂g denotes the subdifferental of a closed convex function g.

Definition 1 Given $\epsilon \ge 0$, a point $\mathbf{x}^*$ is called an ϵ-stationary point of (P) if for some $\boldsymbol{\gamma} \in \mathbb{R}^n$,

$$\max\left\{\operatorname{dist}\left(\mathbf{0}, \nabla f_0(\mathbf{x}^*) + \mathbf{A}^\top\boldsymbol{\gamma} + \partial g(\mathbf{x}^*)\right), \|\mathbf{Ax}^* + \mathbf{b}\|\right\} \le \epsilon, \tag{1.7}$$

and a point $(\mathbf{x}^*, \mathbf{y}^*)$ is called an ϵ-stationary point of (SP) if for some $\mathbf{z}_1 \in \mathbb{R}^{\bar{n}}$ and $\mathbf{z}_2 \in \mathbb{R}^n$,

$$\max\left\{\operatorname{dist}(\mathbf{0}, \partial\bar{g}(\mathbf{y}^*) - \mathbf{z}_1), \|\nabla f_0(\mathbf{x}^*) + \bar{\mathbf{A}}^\top\mathbf{z}_1 + \mathbf{A}^\top\mathbf{z}_2\|, \|\mathbf{y}^* - \bar{\mathbf{A}}\mathbf{x}^* - \bar{\mathbf{b}}\|, \|\mathbf{Ax}^* + \mathbf{b}\|\right\} \le \epsilon. \tag{1.8}$$

When $\epsilon = 0$, we simply call $\mathbf{x}^*$ and $(\mathbf{x}^*, \mathbf{y}^*)$ stationary points of problems (P) and (SP), respectively. We say that $\bar{\mathbf{x}}$ is a near ϵ-stationary point of (P) if it is ω-close to an ϵ-stationary point $\mathbf{x}^*$ of (P) with $\omega = O(\epsilon)$.

1.4 Organization

The rest of this paper is organized as follows. In Section 2, we present lower-bound complexity results for solving problem (P) of FOMs under Assumption 2. Then in Section 3, we show the lower-bound complexity results for solving problems (P) and (SP) of FOMs under Assumption 3. In Section 4, we propose an inexact proximal gradient algorithm for solving these two problems and present its worst-case oracle complexity, which shows the tightness of our lower complexity bounds. Concluding remarks are given in Section 5.

2 Lower Bound of Oracle Complexity for Problem (P) under Assumption 2

In this section, under Assumption 2, we derive a lower bound of the oracle complexity of an FOM to find a (near) ϵ-stationary point of problem (P) for a given $\epsilon > 0$. Motivated by [4, 35], our approach is to design a worst-case instance of (P) such that a large number of oracles satisfying Assumption 2 will be needed to find a desired solution.

2.1 A Challenging Instance $\mathcal{P}$ of Problem (P)

Let m_1 and m_2 be positive integers such that $m_1 m_2$ is even and $m_1 \geq 2$. Let $m = 3m_1 m_2$ and $\bar{d}$ be an odd positive integer such that $\bar{d} \geq 5$. Also, set $d = m\bar{d}$, and let

$$\mathbf{x} = \left(\mathbf{x}_1^\top, \ldots, \mathbf{x}_m^\top\right)^\top \in \mathbb{R}^d, \text{ with } \mathbf{x}_i \in \mathbb{R}^{\bar{d}}, \, i = 1, \ldots, m. \tag{2.1}$$

Moreover, we define a matrix $\mathbf{H} \in \mathbb{R}^{(m-1)\bar{d} \times m\bar{d}}$ by

$$\mathbf{H} := mL_f \cdot \mathbf{J}_m \otimes \mathbf{I}_{\bar{d}} = mL_f \cdot \underbrace{\begin{bmatrix} -\mathbf{I}_{\bar{d}} & \mathbf{I}_{\bar{d}} & & & \\ & -\mathbf{I}_{\bar{d}} & \mathbf{I}_{\bar{d}} & & \\ & & \ddots & \ddots & \\ & & & -\mathbf{I}_{\bar{d}} & \mathbf{I}_{\bar{d}} \end{bmatrix}}_{m \text{ blocks}} \left.\vphantom{\begin{bmatrix} a \\ a \\ a \\ a \end{bmatrix}}\right\} m-1 \text{ blocks}. \tag{2.2}$$

Define

$$\begin{gathered} \mathcal{M} := \{im_1 | i = 1, 2, \ldots, 3m_2 - 1\}, \quad \mathcal{M}^C := \{1, 2, \ldots, m-1\} \backslash \mathcal{M}, \\ n = (m - 3m_2)\bar{d}, \quad \bar{n} = (3m_2 - 1)\bar{d}, \end{gathered} \tag{2.3}$$

and let

$$\bar{\mathbf{A}} := mL_f \cdot \mathbf{J}_{\mathcal{M}} \otimes \mathbf{I}_{\bar{d}}, \quad \mathbf{A} := mL_f \cdot \mathbf{J}_{\mathcal{M}^C} \otimes \mathbf{I}_{\bar{d}}, \quad \bar{\mathbf{b}} = \mathbf{0} \in \mathbb{R}^{\bar{n}}, \quad \mathbf{b} = \mathbf{0} \in \mathbb{R}^n, \tag{2.4}$$

where $\mathbf{J}_{\mathcal{M}}$ and $\mathbf{J}_{\mathcal{M}^C}$ are the rows of $\mathbf{J}_m$ indexed by $\mathcal{M}$ and $\mathcal{M}^C$, respectively.

Furthermore, we define $\bar{g} : \mathbb{R}^{\bar{n}} \to \mathbb{R}$ and $g : \mathbb{R}^d \to \mathbb{R}$ as

$$\bar{g}(\mathbf{y}) := \frac{\beta}{mL_f} \|\mathbf{y}\|_1 = \max\left\{ \mathbf{u}^\top \mathbf{y} \,\middle|\, \|\mathbf{u}\|_\infty \leq \frac{\beta}{mL_f} \right\}, \tag{2.5}$$

and

$$g(\mathbf{x}) := \bar{g}(\bar{\mathbf{A}}\mathbf{x}) = \beta \sum_{i \in \mathcal{M}} \|\mathbf{x}_i - \mathbf{x}_{i+1}\|_1, \tag{2.6}$$

where $\mathbf{x}$ has the block structure in (2.1), $\mathcal{M}$ is defined in (2.3), and β is a constant satisfying

$$\beta > (50\pi + 1 + \|\mathbf{A}\|)\sqrt{m}\epsilon. \tag{2.7}$$

Finally, we design the smooth function f_0 to complete the instance of (P). Let $\Psi : \mathbb{R} \mapsto \mathbb{R}$ and $\Phi : \mathbb{R} \mapsto \mathbb{R}$ be defined as

$$\Psi(u) := \begin{cases} 0, & \text{if } u \leq 0, \\ 1 - e^{-u^2}, & \text{if } u > 0, \end{cases} \quad \text{and} \quad \Phi(v) := 4 \arctan v + 2\pi. \tag{2.8}$$

In addition, for each $j = 1, \ldots, \bar{d}$, we define function $\varphi(\cdot, j) : \mathbb{R}^{\bar{d}} \to \mathbb{R}$ as

$$\varphi(\mathbf{z}, j) := \begin{cases} -\Psi(1)\Phi([\mathbf{z}]_1), & \text{if } j = 1, \\ \Psi(-[\mathbf{z}]_{j-1})\Phi(-[\mathbf{z}]_j) - \Psi([\mathbf{z}]_{j-1})\Phi([\mathbf{z}]_j), & \text{if } j = 2, \ldots, \bar{d}, \end{cases} \tag{2.9}$$

and for $i = 1, \ldots, m$, we define $h_i : \mathbb{R}^{\bar{d}} \to \mathbb{R}$ as

$$h_i(\mathbf{z}) := \begin{cases} \varphi(\mathbf{z}, 1) + 3\sum_{j=1}^{\lfloor \bar{d}/2 \rfloor} \varphi(\mathbf{z}, 2j), & \text{if } i \in \left[1, \frac{m}{3}\right], \\ \varphi(\mathbf{z}, 1), & \text{if } i \in \left[\frac{m}{3} + 1, \frac{2m}{3}\right], \\ \varphi(\mathbf{z}, 1) + 3\sum_{j=1}^{\lfloor \bar{d}/2 \rfloor} \varphi(\mathbf{z}, 2j+1), & \text{if } i \in \left[\frac{2m}{3} + 1, m\right]. \end{cases} \tag{2.10}$$

Now for a given $\epsilon \in (0, 1)$ and $L_f > 0$, for $i = 1, \ldots, m$, we define $f_i : \mathbb{R}^{\bar{d}} \to \mathbb{R}$ as

$$f_i(\mathbf{z}) := \frac{300\pi\epsilon^2}{mL_f} h_i\left(\frac{\sqrt{m}L_f\mathbf{z}}{150\pi\epsilon}\right), \forall \mathbf{z} \in \mathbb{R}^{\bar{d}}, \tag{2.11}$$

and let $f_0 : \mathbb{R}^d \to \mathbb{R}$ be

$$f_0(\mathbf{x}) := \sum_{i=1}^{m} f_i(\mathbf{x}_i), \forall \mathbf{x} \in \mathbb{R}^d \text{ with the structure in (2.1).} \tag{2.12}$$

Putting all the components given above, we obtain a specific instance of (P). We formalize it in the following definition.

Definition 2 (instance $\mathcal{P}$) Given $\epsilon \in (0, 1)$ and $L_f > 0$, let m_1, m_2 and $\bar{d}$ be integers such that $m_1 \geq 2$ is even and $\bar{d} \geq 5$ is odd. We refer to as *instance* $\mathcal{P}$ the instance of problem (P) where f_0 is given in (2.12) with each f_i defined in (2.11), g is given in (2.6) with β satisfying (2.7), and $(\mathbf{A}, \bar{\mathbf{A}}, \mathbf{b}, \bar{\mathbf{b}})$ is given in (2.4).

2.2 Properties of Instance $\mathcal{P}$

In order to show the challenge of instance $\mathcal{P}$ for an algorithm under Assumption 2, we give a few facts and properties about $\mathcal{P}$. First, notice that $\bar{\mathbf{A}}$ and $\mathbf{A}$ in (2.4) are two block submatrices of $\mathbf{H}$ in rows. It is easy to obtain the following proposition.

Proposition 2.1 *By the definitions in (2.1) through (2.6), it holds*

(a) $\mathbf{x}_1 = \mathbf{x}_2 = \cdots = \mathbf{x}_m$ *if and only if* $\mathbf{H}\mathbf{x} = \mathbf{0}$*;*
(b) $\mathbf{x}_i = \mathbf{x}_{i+1}$ *for* $i \in \mathcal{M}$ *if and only if* $\bar{\mathbf{A}}\mathbf{x} = \mathbf{0}$ *or equivalently* $g(\mathbf{x}) = 0$*;*
(c) $\mathbf{x}_i = \mathbf{x}_{i+1}$ *for* $i \in \mathcal{M}^C$ *if and only if* $\mathbf{A}\mathbf{x} = \mathbf{0}$*;*
(d) $\mathbf{x}_1 = \mathbf{x}_2 = \cdots = \mathbf{x}_m$ *if and only if* $\mathbf{A}\mathbf{x} = \mathbf{0}$ *and* $g(\mathbf{x}) = 0$.

Second, it is straightforward to have the partial derivatives $\left\{\frac{\partial h_i(\mathbf{z})}{\partial [\mathbf{z}]_j}\right\}$ based on three cases of j.

Case i) When $j = 1$,

$$\frac{\partial h_i(\mathbf{z})}{\partial [\mathbf{z}]_j} = \begin{cases} -\Psi(1)\Phi'([\mathbf{z}]_1) + 3\left[-\Psi'(-[\mathbf{z}]_1)\Phi(-[\mathbf{z}]_2) - \Psi'([\mathbf{z}]_1)\Phi([\mathbf{z}]_2)\right], & \text{if } i \in \left[1, \frac{m}{3}\right], \\ -\Psi(1)\Phi'([\mathbf{z}]_1), & \text{if } i \in \left[\frac{m}{3} + 1, m\right]. \end{cases} \tag{2.13}$$

Case ii) When j is even,

$$\frac{\partial h_i(\mathbf{z})}{\partial[\mathbf{z}]_j} = \begin{cases} 3\left[-\Psi(-[\mathbf{z}]_{j-1})\Phi'(-[\mathbf{z}]_j) - \Psi([\mathbf{z}]_{j-1})\Phi'([\mathbf{z}]_j)\right], & \text{if } i \in \left[1, \frac{m}{3}\right], \\ 0, & \text{if } i \in \left[\frac{m}{3}+1, \frac{2m}{3}\right], \\ 3\left[-\Psi'(-[\mathbf{z}]_j)\Phi(-[\mathbf{z}]_{j+1}) - \Psi'([\mathbf{z}]_j)\Phi([\mathbf{z}]_{j+1})\right], & \text{if } i \in \left[\frac{2m}{3}+1, m\right]. \end{cases} \tag{2.14}$$

Case iii) When j is odd and $j \neq 1$,

$$\frac{\partial h_i(\mathbf{z})}{\partial[\mathbf{z}]_j} = \begin{cases} 3\left[-\Psi'(-[\mathbf{z}]_j)\Phi(-[\mathbf{z}]_{j+1}) - \Psi'([\mathbf{z}]_j)\Phi([\mathbf{z}]_{j+1})\right], & \text{if } i \in \left[1, \frac{m}{3}\right], \\ 0, & \text{if } i \in \left[\frac{m}{3}+1, \frac{2m}{3}\right], \\ 3\left[-\Psi(-[\mathbf{z}]_{j-1})\Phi'(-[\mathbf{z}]_j) - \Psi([\mathbf{z}]_{j-1})\Phi'([\mathbf{z}]_j)\right], & \text{if } i \in \left[\frac{2m}{3}+1, m\right]. \end{cases} \tag{2.15}$$

Thirdly, instance $\mathcal{P}$ satisfies Assumption 1. By (2.6), Assumption 1(c) clearly holds. To verify Assumptions 1(a) and 1(b), we need the next lemma.

Lemma 2.1 *Let Ψ, Φ and h_i be given in (2.8) and (2.10). The following statements hold:*

(a) *$\Psi(u) = 0$ and $\Psi'(u) = 0$ for all $u \leq 0$, and $0 \leq \Psi(u) < 1$, $0 \leq \Psi'(u) \leq \sqrt{2/e}$, $0 < \Phi(v) < 4\pi$, and $0 < \Phi'(v) \leq 4$ for all $v \in \mathbb{R}$.*
(b) *$\Psi(u)\Phi'(v) > 1$ for all u and v satisfying $u \geq 1$ and $|v| < 1$.*
(c) *$h_i(\mathbf{0}) - \inf_{\mathbf{z}} h_i(\mathbf{z}) \leq 10\pi\bar{d}$ for $i = 1, 2, \ldots, m$.*
(d) *∇h_i is 75π-Lipschitz continuous for $i = 1, 2, \ldots, m$.*
(e) *h_i is $25\pi\sqrt{\bar{d}}$-Lipschitz continuous for $i = 1, 2, \ldots, m$.*

Proof. (a)-(d) are directly from [35, Lemma 3.1]. By derivatives (2.13)–(2.15), we obtain that, for any $\mathbf{z} \in \mathbb{R}^{\bar{d}}$,

$$\left|\frac{\partial h_i(\mathbf{z})}{\partial[\mathbf{z}]_j}\right| \leq \max\left\{\sup_u |\Psi(1)\Phi'(u)| + 6\sup_u |\Psi'(u)| \sup_v |\Phi(v)|, 6\sup_u |\Psi(u)| \sup_v |\Phi'(v)|\right\} < 25\pi, \forall\, i, j, \tag{2.16}$$

where the second inequality is from Lemma 2.1(a) and the fact that $\sup_v |\Psi(1)\Phi'(v)| \leq 4(1-e^{-1}) < \pi$ by the definition of Φ. Hence, $\|\nabla h_i(\mathbf{z})\| \leq 25\pi\sqrt{\bar{d}}$. Thus (e) holds, and we complete the proof. □

From the definition of f_i in (2.11) and the chain rule, we have

$$[\nabla f_i(\mathbf{z})]_j = \frac{2\epsilon}{\sqrt{m}}\left[\nabla h_i\left(\frac{\sqrt{m}L_f\mathbf{z}}{150\pi\epsilon}\right)\right]_j, \quad \forall j = 1, \ldots, m, \tag{2.17}$$

which together with Lemma 2.1 clearly indicates the properties below about $\{f_i\}_{i=0}^m$.

Lemma 2.2 *Let $\{f_i\}_{i=0}^m$ be defined in (2.11) and (2.12). Then*

(a) *$f_i(\mathbf{0}) - \inf_{\mathbf{z}} f_i(\mathbf{z}) \leq 3000\pi^2\bar{d}\epsilon^2/mL_f$ for $i = 1, \ldots, m$, and $f_0(\mathbf{0}) - \inf_{\mathbf{x}} f_0(\mathbf{x}) \leq 3000\pi^2\bar{d}\epsilon^2/L_f$.*
(b) *∇f_i is L_f-Lipschitz continuous for $i = 0, 1, \ldots, m$.*
(c) *f_i is $\frac{50\pi\epsilon\sqrt{\bar{d}}}{\sqrt{m}}$-Lipschitz continuous for $i = 1, \ldots, m$, and f_0 is $50\pi\epsilon\sqrt{m\bar{d}}$-Lipschitz continuous.*

By Lemma 2.2 and $\bar{g} \geq 0$, we immediately have the following proposition.

Proposition 2.2 *Instance $\mathcal{P}$ given in Definition 2 satisfies Assumption 1.*

The following lemma characterizes the joint condition number of $\bar{\mathbf{A}}$ and $\mathbf{A}$ defined in (1.4), which will appear in our lower bound of oracle complexity.

Lemma 2.3 *Let* $\mathbf{A}$ *and* $\bar{\mathbf{A}}$ *be given in* (2.4)*. Then* $\frac{m}{4} \leq \kappa([\bar{\mathbf{A}}; \mathbf{A}]) = \kappa(\mathbf{H}) = \frac{\sin(\frac{(3m_1m_2-1)\pi}{6m_1m_2})}{\sin(\frac{\pi}{6m_1m_2})} < m$.

Proof. The first equality holds because $\mathbf{H}$ are split into $\bar{\mathbf{A}}$ and $\mathbf{A}$ in rows. Let

$$\bar{\mathbf{H}} = mL_f \begin{bmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix} \in \mathbb{R}^{(m-1)\times m}.$$

We then have $\mathbf{H}\mathbf{H}^\top = (\bar{\mathbf{H}} \otimes \mathbf{I}_{\bar{d}})(\bar{\mathbf{H}} \otimes \mathbf{I}_{\bar{d}})^\top = \left(\bar{\mathbf{H}}\bar{\mathbf{H}}^\top\right) \otimes \left(\mathbf{I}_{\bar{d}}\mathbf{I}_{\bar{d}}^\top\right) = \left(\bar{\mathbf{H}}\bar{\mathbf{H}}^\top\right) \otimes \mathbf{I}_{\bar{d}}$. Let $\lambda_i(\bar{\mathbf{H}}\bar{\mathbf{H}}^\top)$ be the i-th largest eigenvalue of $\bar{\mathbf{H}}\bar{\mathbf{H}}^\top$. Since $\bar{\mathbf{H}}\bar{\mathbf{H}}^\top$ is tridiagonal and Toeplitz, its eigenvalues have closed forms [11]:

$$\lambda_i(\bar{\mathbf{H}}\bar{\mathbf{H}}^\top) = 4m^2L_f^2 \sin^2\left(\frac{i\pi}{6m_1m_2}\right), \forall\, i = 1, 2, \ldots, m-1.$$

It then yields that $\kappa([\bar{\mathbf{A}}; \mathbf{A}]) = \kappa(\mathbf{H}) = \kappa(\bar{\mathbf{H}}) = \frac{\sin(\frac{(3m_1m_2-1)\pi}{6m_1m_2})}{\sin(\frac{\pi}{6m_1m_2})}$. Because $\sin(z) \leq 1, \forall\, z$, $\sin(z) \geq \frac{2z}{3}$ for $z \in [0, \pi/12]$, and $m_1m_2 \geq 2$, we have

$$\frac{\sin(\frac{(3m_1m_2-1)\pi}{6m_1m_2})}{\sin(\frac{\pi}{6m_1m_2})} \leq \frac{1}{\frac{\pi}{9m_1m_2}} = \frac{3m}{\pi} < m.$$

Also, because $z \geq \sin(z) \geq \frac{z}{2}$ for $z \in [0, \pi/2]$ and $m_1m_2 \geq 2$, we have

$$\frac{\sin(\frac{(3m_1m_2-1)\pi}{6m_1m_2})}{\sin(\frac{\pi}{6m_1m_2})} \geq \frac{\frac{(m-1)\pi}{4m}}{\frac{\pi}{2m}} = \frac{m-1}{2} \geq \frac{m}{4}.$$

Hence, we have obtained all desired results and complete the proof. □

2.3 An Auxiliary Problem and Its Properties

To establish a lower bound of the oracle complexity for solving (P) under Assumptions 1 and 2, we consider an auxiliary problem of instance $\mathcal{P}$ in this subsection and analyze its properties. The Auxiliary Problem is given as follows:

$$\min_{\mathbf{x}\in\mathbb{R}^d} f_0(\mathbf{x}), \quad \text{s.t. } \mathbf{H}\mathbf{x} = \mathbf{0}, \tag{AP}$$

where $\mathbf{H}$ and f_0 are defined in (2.12) and (2.2), respectively. An ϵ-stationary point $\mathbf{x}^*$ of problem (AP) satisfies

$$\max\left\{\|\mathbf{H}\mathbf{x}^*\|, \min_{\boldsymbol{\gamma}'\in\mathbb{R}^{(m-1)\bar{d}}} \left\|\nabla f_0(\mathbf{x}^*) + \mathbf{H}^\top\boldsymbol{\gamma}'\right\|\right\} \leq \epsilon. \tag{2.18}$$

The next lemma characterizes the relationship between the (near-)stationary points of instance $\mathcal{P}$ and the auxiliary problem (AP). The proof techniques have been utilized in [21, 22].

Lemma 2.4 *Let $\epsilon > 0$ be given in Definition 2 for instance $\mathcal{P}$. Then for any $\widehat{\epsilon} \in [0, \epsilon]$, an $\widehat{\epsilon}$-stationary point of instance $\mathcal{P}$ is also an $\widehat{\epsilon}$-stationary point of the auxiliary problem (AP).*

Proof. Suppose $\mathbf{x}^*$ is an $\widehat{\epsilon}$-stationary point of instance $\mathcal{P}$, i.e., for some $\boldsymbol{\gamma} \in \mathbb{R}^n$ and $\boldsymbol{\xi} \in \partial g(\mathbf{x}^*)$ such that

$$\|\nabla f_0(\mathbf{x}^*) + \boldsymbol{\xi} + \mathbf{A}^\top \boldsymbol{\gamma}\| \le \widehat{\epsilon} \quad \text{and} \quad \|\mathbf{A}\mathbf{x}^*\| \le \widehat{\epsilon}. \tag{2.19}$$

By the definitions of g in (2.5) and (2.6), there exists $\mathbf{u} \in \mathbb{R}^{\bar{n}}$ such that $\boldsymbol{\xi} = \bar{\mathbf{A}}^\top \mathbf{u}$ and thus the first condition in (2.19) becomes

$$\|\nabla f_0(\mathbf{x}^*) + \bar{\mathbf{A}}^\top \mathbf{u} + \mathbf{A}^\top \boldsymbol{\gamma}\| \le \widehat{\epsilon}. \tag{2.20}$$

Hence, in order to show that $\mathbf{x}^*$ is an $\widehat{\epsilon}$-stationary point of problem (AP), we only need to prove $\|\mathbf{H}\mathbf{x}^*\| \le \widehat{\epsilon}$. To show this, we first notice from the definition in (1.6) that, for any $\mathbf{v} \in \mathbf{Null}(\mathbf{A})$,

$$F_0'(\mathbf{x}^*; \mathbf{v}) = \mathbf{v}^\top \nabla f_0(\mathbf{x}^*) + g'(\mathbf{x}^*; \mathbf{v}) \ge \mathbf{v}^\top \nabla f_0(\mathbf{x}^*) + \mathbf{v}^\top \boldsymbol{\xi} = \mathbf{v}^\top \left(\nabla f_0(\mathbf{x}^*) + \boldsymbol{\xi} + \mathbf{A}^\top \boldsymbol{\gamma}\right) \ge -\widehat{\epsilon}\|\mathbf{v}\|, \tag{2.21}$$

where the first inequality follows from [8]

$$g'(\mathbf{x}; \mathbf{v}) = \sup_{\boldsymbol{\xi}' \in \partial g(\mathbf{x})} \mathbf{v}^\top \boldsymbol{\xi}', \forall\, \mathbf{x} \in \operatorname{dom}(g), \forall\, \mathbf{v}, \tag{2.22}$$

and the second inequality is by (2.19) and the Cauchy-Schwarz inequality.

Second, we claim $\bar{\mathbf{A}}\mathbf{x}^* = \mathbf{0}$, namely, $\mathbf{x}_i^* = \mathbf{x}_{i+1}^*$ for all $i \in \mathcal{M}$, where $\mathcal{M}$ is defined in (2.3). Suppose this claim is not true. Then for some $\bar{i} \in \mathcal{M}$, it holds $\mathbf{x}_{\bar{i}}^* \ne \mathbf{x}_{\bar{i}+1}^*$. We let

$$\bar{\boldsymbol{\xi}} := \mathbf{x}_{\bar{i}+1}^* - \mathbf{x}_{\bar{i}}^* \ne \mathbf{0}, \tag{2.23}$$

and $\mathbf{v}^* = (\mathbf{v}_1^{*\top}, \mathbf{v}_2^{*\top}, \ldots, \mathbf{v}_m^{*\top})^\top$ where each $\mathbf{v}_i^* \in \mathbb{R}^{\bar{d}}$ is defined as

$$\mathbf{v}_i^* = \bar{\boldsymbol{\xi}}, \text{ if } i \le \bar{i}, \text{ and } \mathbf{v}_i^* = \mathbf{0}, \text{ if } i > \bar{i}.$$

It is easy to see that $\mathbf{v}_i^* = \mathbf{v}_{i+1}^*$ for any $i \ne \bar{i}$. Thus by Proposition 2.1(b), $\mathbf{A}\mathbf{v}^* = \mathbf{0}$ and, for any $s \in (0, 1)$,

$$\begin{aligned} g(\mathbf{x}^* + s\mathbf{v}^*) =& \beta \sum_{i<\bar{i}} \left\|\mathbf{x}_i^* + s\bar{\boldsymbol{\xi}} - \mathbf{x}_{i+1}^* - s\bar{\boldsymbol{\xi}}\right\|_1 + \beta\|\mathbf{x}_{\bar{i}}^* + s\bar{\boldsymbol{\xi}} - \mathbf{x}_{\bar{i}+1}^*\|_1 + \beta \sum_{i>\bar{i}} \left\|\mathbf{x}_i^* - \mathbf{x}_{i+1}^*\right\|_1 \\ =& \beta \sum_{i<\bar{i}} \left\|\mathbf{x}_i^* - \mathbf{x}_{i+1}^*\right\|_1 + \beta(1-s)\|\mathbf{x}_{\bar{i}}^* - \mathbf{x}_{\bar{i}+1}^*\|_1 + \beta \sum_{i>\bar{i}} \left\|\mathbf{x}_i^* - \mathbf{x}_{i+1}^*\right\|_1 = g(\mathbf{x}^*) - s\beta\|\bar{\boldsymbol{\xi}}\|_1, \end{aligned} \tag{2.24}$$

where the first and last equalities follow from (2.6) and the definition of $\mathbf{v}^*$, and the second one is by (2.23) and $s \in (0, 1)$. In addition, from (2.11) and (2.16), we have that, for any $\mathbf{z} \in \mathbb{R}^{\bar{d}}$,

$$\|\nabla f_i(\mathbf{z})\|_\infty = \frac{2\epsilon}{\sqrt{m}} \left\|\nabla h_i\left(\frac{\sqrt{m} L_f \mathbf{z}}{150\pi\epsilon}\right)\right\|_\infty \le \frac{50\pi\epsilon}{\sqrt{m}}. \tag{2.25}$$

Moreover, by the definition of $\mathbf{v}_i^*$ for $i = 1, \ldots, m$, we have

$$f_i(\mathbf{x}_i^* + s\mathbf{v}_i^*) - f_i(\mathbf{x}_i^*) = s\nabla f_i(\mathbf{x}_i^* + s'\mathbf{v}_i^*)^\top \mathbf{v}_i^* \le s\|\nabla f_i(\mathbf{x}_i^* + s'\mathbf{v}_i^*)\|_\infty \|\mathbf{v}_i^*\|_1 \overset{(2.25)}{\le} \frac{50 s\pi\epsilon}{\sqrt{m}} \|\bar{\boldsymbol{\xi}}\|_1,$$

where the equality holds from the mean value theorem for some $s' \in (0, s)$. The inequality above, together with (2.12) and (2.24), implies

$$\begin{aligned}
&\frac{1}{s}\Big(F_0(\mathbf{x}^* + s\mathbf{v}^*) - F_0(\mathbf{x}^*)\Big) = \frac{1}{s}\Big(f_0(\mathbf{x}^* + s\mathbf{v}^*) - f_0(\mathbf{x}^*) + g(\mathbf{x}^* + s\mathbf{v}^*) - g(\mathbf{x}^*)\Big)\\
=&\frac{1}{s}\Big(\sum_{i=1}^{m}\big(f_i(\mathbf{x}_i^* + s\mathbf{v}_i^*) - f_i(\mathbf{x}_i^*)\big) + g(\mathbf{x}^* + s\mathbf{v}^*) - g(\mathbf{x}^*)\Big)\\
\leq&\frac{1}{s}\left(50\pi s\epsilon\sqrt{m}\|\bar{\boldsymbol{\xi}}\|_1 - \beta s\|\bar{\boldsymbol{\xi}}\|_1\right) = \left(50\pi\epsilon\sqrt{m} - \beta\right)\|\bar{\boldsymbol{\xi}}\|_1.
\end{aligned}$$

Taking the limit of the left-hand side of the inequality above as s approaching zero, we have

$$\left(50\pi\epsilon\sqrt{m} - \beta\right)\|\bar{\boldsymbol{\xi}}\|_1 \geq F_0'(\mathbf{x}^*; \mathbf{v}).$$

Thus by (2.21) and the choice of $\mathbf{v}^*$, we have

$$\left(50\pi\epsilon\sqrt{m} - \beta\right)\|\bar{\boldsymbol{\xi}}\|_1 \geq -\widehat{\epsilon}\|\mathbf{v}^*\| \geq -\widehat{\epsilon}\sqrt{\bar{i}}\|\bar{\boldsymbol{\xi}}\| \geq -\widehat{\epsilon}\sqrt{m}\|\bar{\boldsymbol{\xi}}\|_1.$$

This leads to a contradiction as $\beta > (50\pi + 1)\epsilon\sqrt{m}$ from (2.7) and $\widehat{\epsilon} \leq \epsilon$. Therefore, the claim $\bar{\mathbf{A}}\mathbf{x}^* = \mathbf{0}$ is true. Thus the second condition in (2.19) indicates $\|\mathbf{H}\mathbf{x}^*\| \leq \widehat{\epsilon}$, and we complete the proof. □

By Lemma 2.4, if $\mathbf{x}^*$ is not an ϵ-stationary point of the auxiliary problem (AP), it cannot be an ϵ-stationary point of instance $\mathcal{P}$. In other words, the number of oracles needed to find an ϵ-stationary point of $\mathcal{P}$ is at least the number of oracles needed to find an ϵ-stationary point of problem (AP). Note that the auxiliary problem (AP) of instance $\mathcal{P}$ is the worst-case instance used in [35] to establish the lower-bound complexity for affinely constrained smooth optimization. In fact, according to [35], any algorithm that can access ∇f_0 and matrix-vector multiplication with $\mathbf{H}$ and $\mathbf{H}^\top$ at any historical solutions needs at least $\Theta(\kappa(\mathbf{H})L_f\Delta_{f_0}\epsilon^{-2})$ oracles to find an ϵ-stationary point of (AP), where $\Delta_{f_0} := f_0(\mathbf{x}^{(0)}) - \inf_{\mathbf{x}} f_0(\mathbf{x})$. However, we cannot directly apply the lower bound here because the problem (P) that we consider has both affine constraints and a non-smooth term, and our algorithms cannot apply $\mathbf{H}$ and $\mathbf{H}^\top$ for matrix-vector multiplications but instead can use $\mathbf{A}$ and $\mathbf{A}^\top$ as well as the proximal mapping of g (see Assumption 2). Next, we show that any algorithm under Assumption 2 also needs at least $\Theta(\kappa(\mathbf{H})L_f\Delta_{f_0}\epsilon^{-2})$ oracles to find an ϵ-stationary point of problem (AP).

To do so, we need the following lemma, which is a direct consequence of [35, Lemma 3.3] and the proof is provided for the sake of completeness.

Lemma 2.5 *Let $\{f_i\}_{i=1}^m$ be defined in (2.11) with $\epsilon > 0$. For any $\mathbf{z} \in \mathbb{R}^{\bar{d}}$, if $|[\mathbf{z}]_{\bar{j}}| < \frac{150\pi\epsilon}{\sqrt{m}L_f}$ for some $\bar{j} \in \{1, 2, \ldots, \bar{d}\}$, then $\left\|\frac{1}{m}\sum_{i=1}^m \nabla f_i(\mathbf{z})\right\| > \frac{2\epsilon}{\sqrt{m}}$.*

Proof. Let $\bar{\mathbf{z}} = \frac{\sqrt{m}L_f\mathbf{z}}{150\pi\epsilon}$. We first consider the case where $|[\mathbf{z}]_j| < \frac{150\pi\epsilon}{\sqrt{m}L_f}$ for all $j = 1, 2, \ldots, \bar{j}$. In this case, $|[\bar{\mathbf{z}}]_1| = \frac{\sqrt{m}L_f|[\mathbf{z}]_1|}{150\pi\epsilon} < 1$. By (2.11), we have

$$\left\|\frac{1}{m}\sum_{i=1}^m \nabla f_i(\mathbf{z})\right\| \geq \left|\frac{1}{m}\sum_{i=1}^m [\nabla f_i(\mathbf{z})]_1\right| = \left|\frac{2\epsilon}{m\sqrt{m}}\sum_{i=1}^m [\nabla h_i(\bar{\mathbf{z}})]_1\right|. \tag{2.26}$$

In addition, according to (2.13), we have

$$\frac{1}{m}\sum_{i=1}^{m}\left[\nabla h_i\left(\bar{\mathbf{z}}\right)\right]_1 = -\Psi(1)\Phi'([\bar{\mathbf{z}}]_1) + [-\Psi'(-[\bar{\mathbf{z}}]_1)\Phi(-[\bar{\mathbf{z}}]_2) - \Psi'([\bar{\mathbf{z}}]_1)\Phi([\bar{\mathbf{z}}]_2)] \leq -\Psi(1)\Phi'([\bar{\mathbf{z}}]_1) < -1, \tag{2.27}$$

where the first inequality comes from the non-negativity of Ψ' and Φ by Lemma 2.1(a), and the second inequality is by Lemma 2.1(b) and $|\bar{\mathbf{z}}_1| < 1$. Combing (2.26) and (2.27) yields the desired inequality.

Second, we consider the case where there exists $j \in \{2, \ldots, \bar{j}\}$ such that $|[\mathbf{z}]_j| < \frac{150\pi\epsilon}{\sqrt{m}L_f} \leq |[\mathbf{z}]_{j-1}|$. In this case, $|[\bar{\mathbf{z}}]_j| < 1 \leq |[\bar{\mathbf{z}}]_{j-1}|$. By (2.11) again, we have

$$\left\|\frac{1}{m}\sum_{i=1}^{m}\nabla f_i(\mathbf{z})\right\| \geq \left|\frac{1}{m}\sum_{i=1}^{m}\left[\nabla f_i(\mathbf{z})\right]_j\right| = \left|\frac{2\epsilon}{m\sqrt{m}}\sum_{i=1}^{m}\left[\nabla h_i\left(\bar{\mathbf{z}}\right)\right]_j\right|. \tag{2.28}$$

According to (2.14) and (2.15), we have

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}\left[\nabla h_i\left(\bar{\mathbf{z}}\right)\right]_j =& -\Psi(-[\bar{\mathbf{z}}]_{j-1})\Phi'(-[\bar{\mathbf{z}}]_j) - \Psi([\bar{\mathbf{z}}]_{j-1})\Phi'([\bar{\mathbf{z}}]_j) - \Psi'(-[\bar{\mathbf{z}}]_j)\Phi(-[\bar{\mathbf{z}}]_{j+1}) - \Psi'([\bar{\mathbf{z}}]_j)\Phi([\bar{\mathbf{z}}]_{j+1})\\ \leq& -\Psi(-[\bar{\mathbf{z}}]_{j-1})\Phi'(-[\bar{\mathbf{z}}]_j) - \Psi([\bar{\mathbf{z}}]_{j-1})\Phi'([\bar{\mathbf{z}}]_j) = -\Psi(|[\bar{\mathbf{z}}]_{j-1}|)\Phi'([\bar{\mathbf{z}}]_j) < -1,\end{aligned} \tag{2.29}$$

where the first inequality comes from the nonnegativity of Ψ' and Φ by Lemma 2.1(a), the second equality holds by the fact that $\Phi'(u) = \Phi'(-u)$ and $\Psi(u) = 0$ for all $u \leq 0$ from (2.8) and Lemma 2.1(a), and the second inequality is by Lemma 2.1(b) and the fact that $|[\bar{\mathbf{z}}]_{j-1}| \geq 1$ and $|[\bar{\mathbf{z}}]_j| < 1$. Combining (2.28) and (2.29) yields the desired inequality and completes the proof. □

The following lemma provides a lower bound to the stationarity measure of a point $\mathbf{x}$ as a solution to problem (AP).

Lemma 2.6 *Let $\mathbf{x} \in \mathbb{R}^d$ be given in (2.1), $\mathbf{H}$ in (2.2), and $\{f_i\}_{i=0}^m$ in (2.11) and (2.12). Then*

$$\max\left\{\|\mathbf{H}\mathbf{x}\|, \min_{\boldsymbol{\gamma}}\left\|\nabla f_0(\mathbf{x}) + \mathbf{H}^\top\boldsymbol{\gamma}\right\|\right\} \geq \frac{\sqrt{m}}{2}\left\|\frac{1}{m}\sum_{i=1}^{m}\nabla f_i(\bar{\mathbf{x}})\right\|, \text{ with } \bar{\mathbf{x}} := \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i.$$

Proof. By simple calculation and the fact $\mathbf{Null}(\mathbf{H}) = \left\{\mathbf{1}_m \otimes \mathbf{u} : \mathbf{u} \in \mathbb{R}^{\bar{d}}\right\}$, we obtain

$$\min_{\boldsymbol{\gamma}}\left\|\nabla f_0(\mathbf{x}) + \mathbf{H}^\top\boldsymbol{\gamma}\right\|^2 = \left\|\mathbf{Proj}_{\mathbf{Null}(\mathbf{H})}(\nabla f_0(\mathbf{x}))\right\|^2 = \frac{1}{m}\left\|\sum_{i=1}^{m}\nabla f_i\left(\mathbf{x}_i\right)\right\|^2, \tag{2.30}$$

$$\|\mathbf{H}\mathbf{x}\|^2 = m^2 L_f^2\sum_{i=1}^{m-1}\left\|\mathbf{x}_i - \mathbf{x}_{i+1}\right\|^2. \tag{2.31}$$

By the L_f-Lipschitz continuity of ∇f_0, we have

$$\frac{1}{2}\left\|\frac{1}{m}\sum_{i=1}^{m}\nabla f_i(\bar{\mathbf{x}})\right\|^2 - \left\|\frac{1}{m}\sum_{i=1}^{m}\nabla f_i\left(\mathbf{x}_i\right)\right\|^2 \leq \left\|\frac{1}{m}\sum_{i=1}^{m}\left(\nabla f_i(\bar{\mathbf{x}}) - \nabla f_i\left(\mathbf{x}_i\right)\right)\right\|^2$$

$$\leq \frac{1}{m}\sum_{i=1}^{m}\left\|\nabla f_i\left(\bar{\mathbf{x}}\right)-\nabla f_i\left(\mathbf{x}_i\right)\right\|^2 \leq \frac{1}{m}\sum_{i=1}^{m} L_f^2\left\|\bar{\mathbf{x}}-\mathbf{x}_i\right\|^2 \overset{(a)}{\leq} \frac{L_f^2}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\left\|\mathbf{x}_j-\mathbf{x}_i\right\|^2$$

$$\overset{(b)}{\leq} \frac{L_f^2}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\left[|j-i|\sum_{k=\min\{i,j\}}^{\max\{i,j\}-1}\left\|\mathbf{x}_k-\mathbf{x}_{k+1}\right\|^2\right] \overset{(c)}{\leq} \frac{L_f^2}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} m\sum_{k=1}^{m-1}\left\|\mathbf{x}_k-\mathbf{x}_{k+1}\right\|^2 = mL_f^2\sum_{i=1}^{m-1}\left\|\mathbf{x}_i-\mathbf{x}_{i+1}\right\|^2,$$

where (a) comes from $\|\bar{\mathbf{x}}-\mathbf{x}_i\|^2 = \frac{1}{m^2}\|\sum_{j=1}^{m}\mathbf{x}_j-\mathbf{x}_i\|^2 \leq \frac{1}{m}\sum_{j=1}^{m}\|\mathbf{x}_j-\mathbf{x}_i\|^2$, (b) results from the fact that $\|\mathbf{x}_j-\mathbf{x}_i\|^2 \leq |j-i|\sum_{k=\min\{i,j\}}^{\max\{i,j\}-1}\|\mathbf{x}_k-\mathbf{x}_{k+1}\|^2$, (c) holds by $|j-i|<m$, $\max\{i,j\}\leq m$ and $\min\{i,j\}\geq 1$.

Hence, by (2.30) and (2.31) and the fact $a+b\leq 2\max\{a,b\}$ for any $a,b\in\mathbb{R}$, we obtain the desired result from the inequality above and complete the proof. □

The previous two lemmas imply that if there exists $\bar{j}\in\{1,2,\ldots,\bar{d}\}$ such that $[\bar{\mathbf{x}}]_{\bar{j}}=0$, where $\bar{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i$. Then $\mathbf{x}$ cannot be an ϵ-stationary point of the auxiliary problem (AP) of instance $\mathcal{P}$.

2.4 Lower Bound of Oracle Complexity

In this subsection, we provide a lower bound of the oracle complexity of FOMs for solving problem (P) under Assumptions 1 and 2, by showing that a large number of oracles will be needed to find a (near) ϵ-stationary point of instance $\mathcal{P}$.

For any integer $t\geq 0$, let

$$\mathbf{x}^{(t)} = (\mathbf{x}_1^{(t)\top},\ldots,\mathbf{x}_m^{(t)\top})^\top \text{ with each } \mathbf{x}_i^{(t)}\in\mathbb{R}^{\bar{d}}, \text{ and } \bar{\mathbf{x}}^{(t)} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i^{(t)} \tag{2.32}$$

be the t-th iterate of an algorithm and the block average. Without loss of generality, we assume $\mathbf{x}^{(0)}=\mathbf{0}$. Otherwise, we can change variable[3] $\mathbf{x}$ in instance $\mathcal{P}$ to $\mathbf{x}-\mathbf{x}^{(0)}$, and process the rest of the proof with this new resulting instance. Our lower bound will be established based on the fact that, if t is not large enough, $\mathrm{supp}(\bar{\mathbf{x}}^{(t)})\subset\{1,\ldots,\bar{j}-1\}$ for some $\bar{j}\in\{1,2,\ldots,\bar{d}\}$. Hence, $[\bar{\mathbf{x}}^{(t)}]_{\bar{j}}=0$ and $\mathbf{x}^{(t)}$ cannot be a (near) ϵ-stationary point due to Lemmas 2.5 and 2.6. According to Assumption 2, $\mathrm{supp}(\bar{\mathbf{x}}^{(t)})$ is influenced by $\mathrm{supp}(\nabla f_i(\mathbf{x}_i^{(t)}))$, which is characterized by the following lemma.

Lemma 2.7 *Let $\{f_i\}_{i=1}^m$ be defined in (2.11). Given any $\bar{j}\in\{1,\ldots,\bar{d}\}$ and $\mathbf{z}\in\mathbb{R}^{\bar{d}}$ with $\mathrm{supp}(\mathbf{z})\subset\{1,\ldots,\bar{j}-1\}$[4], it holds that*

1. *When $\bar{j}=1$, $\mathrm{supp}(\nabla f_i(\mathbf{z}))\subset\{1\}$, for any $i\in[1,m]$;*
2. *When $\bar{j}$ is even,*

$$\mathrm{supp}(\nabla f_i(\mathbf{z}))\subset\begin{cases}\{1,\ldots,\bar{j}\}, & \text{if } \ i\in\left[1,\frac{m}{3}\right],\\ \{1,\ldots,\bar{j}-1\}, & \text{if } \ i\in\left[\frac{m}{3}+1,m\right];\end{cases}$$

[3] This involves changing functions $f(\mathbf{x})$ to $f(\mathbf{x}-\mathbf{x}^{(0)})$, $g(\mathbf{x})$ to $g(\mathbf{x}-\mathbf{x}^{(0)})$, and $\mathbf{A}\mathbf{x}=\mathbf{b}$ to $\mathbf{A}(\mathbf{x}-\mathbf{x}^{(0)})=\mathbf{b}$ in instance $\mathcal{P}$. Then, the resulted instance becomes $\min_{\mathbf{x}\in\mathbb{R}^d} F_0(\mathbf{x}) := f_0(\mathbf{x}-\mathbf{x}^{(0)})+g(\mathbf{x}-\mathbf{x}^{(0)})$, s.t. $\mathbf{A}(\mathbf{x}-\mathbf{x}^{(0)})+\mathbf{b}=\mathbf{0}$.

[4] When $\bar{j}=1$, this means $\mathrm{supp}(\mathbf{z})=\emptyset$ and $\mathbf{z}=\mathbf{0}$.

3. When $\bar{j}$ is odd and $\bar{j} \neq 1$,

$$\operatorname{supp}(\nabla f_i(\mathbf{z})) \subset \begin{cases} \{1, \ldots, \bar{j}-1\}, & \textit{if} \;\; i \in \left[1, \frac{2m}{3}\right] \\ \{1, \ldots, \bar{j}\}, & \textit{if} \;\; i \in \left[\frac{2m}{3}+1, m\right]. \end{cases}$$

Proof. Denote $\bar{\mathbf{z}} = \frac{\sqrt{m} L_f \mathbf{z}}{150\pi\epsilon}$ and recall definition (2.17). By Lemma 2.1(a), we have

$$\Psi\left(-[\bar{\mathbf{z}}]_j\right) = \Psi\left([\bar{\mathbf{z}}]_j\right) = \Psi'\left(-[\bar{\mathbf{z}}]_j\right) = \Psi'\left([\bar{\mathbf{z}}]_j\right) = 0, \forall\, j \geq \bar{j}. \tag{2.33}$$

Therefore, by definitions (2.13)–(2.15), the support of $\mathbf{z}$ leads to the following structure of $[\nabla f_i(\mathbf{z})]_j$ for $j \geq \bar{j}$:

- If $j = 1$, $[\nabla f_i(\mathbf{z})]_j = -\Psi(1)\Phi'\left([\bar{\mathbf{z}}]_j\right)$, for any $i \in [1, m]$;
- If j is even,

$$[\nabla f_i(\mathbf{z})]_j = \begin{cases} -\frac{6\epsilon}{\sqrt{m}}\left[\Psi\left(-[\bar{\mathbf{z}}]_{j-1}\right)\Phi'\left(-[\bar{\mathbf{z}}]_j\right) + \Psi\left([\bar{\mathbf{z}}]_{j-1}\right)\Phi'\left([\bar{\mathbf{z}}]_j\right)\right], & \text{for } i \in \left[1, \frac{m}{3}\right], \\ 0, & \text{for } i \in \left[\frac{m}{3}+1, m\right]; \end{cases}$$

- If j is odd and $j \neq 1$,

$$[\nabla f_i(\mathbf{z})]_j = \begin{cases} 0, & \text{for } i \in \left[1, \frac{2m}{3}\right], \\ -\frac{6\epsilon}{\sqrt{m}}\left[\Psi\left(-[\bar{\mathbf{z}}]_{j-1}\right)\Phi'\left(-[\bar{\mathbf{z}}]_j\right) + \Psi\left([\bar{\mathbf{z}}]_{j-1}\right)\Phi'\left([\bar{\mathbf{z}}]_j\right)\right], & \text{for } i \in \left[\frac{2m}{3}+1, m\right]. \end{cases}$$

Since $\Psi\left(-[\bar{\mathbf{z}}]_{j-1}\right) = \Psi\left([\bar{\mathbf{z}}]_{j-1}\right) = 0$ for any $j > \bar{j}$, the structures above imply $[\nabla f_i(\mathbf{z})]_j = 0, \forall\, j > \bar{j}$ and thus give the desired claims. □

According to the structure of $\mathbf{A}$ given in (2.4), $\operatorname{supp}((\mathbf{A}^\top \mathbf{A}\mathbf{x})_i)$ is determined by $\operatorname{supp}(\mathbf{x}_{i-1})$, $\operatorname{supp}(\mathbf{x}_i)$ and $\operatorname{supp}(\mathbf{x}_{i+1})$. Also, $\operatorname{supp}(\mathbf{prox}_{\eta g}(\mathbf{x}))$ has a similar property according to the definition of g in (2.6). These properties are formally stated in the following lemma.

Lemma 2.8 *Let $\mathbf{x}$ be the structured vector given in (2.1), $\mathbf{A}$ in (2.4), and g be given in (2.6). Define $\mathbf{x}_0 = \mathbf{x}_{m+1} = \mathbf{0} \in \mathbb{R}^{\bar{d}}$. The following statements hold:*

(a) *Let $\widehat{\mathbf{x}} = \mathbf{A}^\top \mathbf{A}\mathbf{x} = (\widehat{\mathbf{x}}_1^\top, \ldots, \widehat{\mathbf{x}}_m^\top)^\top$ with each $\widehat{\mathbf{x}}_i \in \mathbb{R}^{\bar{d}}$. Then*

$$\operatorname{supp}(\widehat{\mathbf{x}}_i) \subset \operatorname{supp}(\mathbf{x}_{i-1}) \cup \operatorname{supp}(\mathbf{x}_i) \cup \operatorname{supp}(\mathbf{x}_{i+1}), \; \forall\, i \in [1, m]. \tag{2.34}$$

(b) *For any $\eta > 0$, let $\widetilde{\mathbf{x}} = \mathbf{prox}_{\eta g}(\mathbf{x}) = (\widetilde{\mathbf{x}}_1^\top, \ldots, \widetilde{\mathbf{x}}_m^\top)^\top$ with each $\widetilde{\mathbf{x}}_i \in \mathbb{R}^{\bar{d}}$. Then*

$$\operatorname{supp}(\widetilde{\mathbf{x}}_i) \subset \operatorname{supp}(\mathbf{x}_{i-1}) \cup \operatorname{supp}(\mathbf{x}_i) \cup \operatorname{supp}(\mathbf{x}_{i+1}), \; \forall\, i \in [1, m].$$

Proof. (a) The relation in (2.34) immediately follows from the observation

$$\mathbf{A}^\top \mathbf{A} = \left.\begin{bmatrix} \mathbf{B} & & & \\ & \mathbf{B} & & \\ & & \ddots & \\ & & & \mathbf{B} \end{bmatrix}\right\} 3m_2 \text{ blocks}, \;\; \text{with } \mathbf{B} = m^2 L_f^2 \left.\underbrace{\begin{bmatrix} \mathbf{I}_{\bar{d}} & -\mathbf{I}_{\bar{d}} & & & \\ -\mathbf{I}_{\bar{d}} & 2\mathbf{I}_{\bar{d}} & -\mathbf{I}_{\bar{d}} & & \\ & \ddots & \ddots & \ddots & \\ & & -\mathbf{I}_{\bar{d}} & 2\mathbf{I}_{\bar{d}} & -\mathbf{I}_{\bar{d}} \\ & & & -\mathbf{I}_{\bar{d}} & \mathbf{I}_{\bar{d}} \end{bmatrix}}_{m_1 \text{ blocks}}\right\} m_1 \text{ blocks}. \tag{2.35}$$

(b) Given any x_1 and x_2 in $\mathbb{R}$ and any $c > 0$, consider the following optimization problem in $\mathbb{R}^2$:

$$(\widetilde{x}_1, \widetilde{x}_2) = \mathop{\arg\min}_{z_1, z_2 \in \mathbb{R}} \frac{1}{2}(z_1 - x_1)^2 + \frac{1}{2}(z_2 - x_2)^2 + c|z_1 - z_2|.$$

The optimal solution of this problem is

$$(\widetilde{x}_1, \widetilde{x}_2) = \begin{cases} ((x_1 + x_2)/2, (x_1 + x_2)/2), & \text{if } |x_1 - x_2| \le 2c \\ (x_1 - c \cdot \operatorname{sign}(x_1 - x_2), x_2 + c \cdot \operatorname{sign}(x_1 - x_2)), & \text{if } |x_1 - x_2| > 2c. \end{cases} \tag{2.36}$$

Recall the definition of g in (2.6) and $\mathbf{prox}_{\eta g}$, we obtain that

$$\mathbf{prox}_{\eta g}(\mathbf{x}) = \mathop{\arg\min}_{\mathbf{y}} \eta\beta \sum_{i \in \mathcal{M}} \|\mathbf{x}_i - \mathbf{x}_{i+1}\|_1 + \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2.$$

It then holds that

$$\widetilde{\mathbf{x}}_i = \begin{cases} \arg\min_{\mathbf{y}_i} \frac{1}{2}\|\mathbf{x}_i - \mathbf{y}_i\|^2, & \text{if } i-1, i \notin \mathcal{M}, \\ \arg\min_{\mathbf{y}_i} \eta\beta\|\mathbf{x}_i - \mathbf{x}_{i-1}\|_1 + \frac{1}{2}\|\mathbf{x}_i - \mathbf{y}_i\|^2, & \text{if } i-1 \in \mathcal{M}, \\ \arg\min_{\mathbf{y}_i} \eta\beta\|\mathbf{x}_i - \mathbf{x}_{i+1}\|_1 + \frac{1}{2}\|\mathbf{x}_i - \mathbf{y}_i\|^2, & \text{if } i \in \mathcal{M}. \end{cases}$$

By (2.36) and the separability property of $\|\cdot\|_1$ and $\|\cdot\|^2$, we have that for any $j \in \{1, \ldots, \bar{d}\}$,

$$[\widetilde{\mathbf{x}}_i]_j = \begin{cases} [\mathbf{x}_i]_j, & \text{if } i-1, i \notin \mathcal{M}, \\ ([\mathbf{x}_{i-1}]_j + [\mathbf{x}_i]_j)/2, & \text{if } i-1 \in \mathcal{M} \text{ and } |[\mathbf{x}_{i-1}]_j - [\mathbf{x}_i]_j| \le \eta\beta, \\ ([\mathbf{x}_i]_j + [\mathbf{x}_{i+1}]_j)/2, & \text{if } i \in \mathcal{M} \text{ and } |[\mathbf{x}_i]_j - [\mathbf{x}_{i+1}]_j| \le \eta\beta, \\ [\mathbf{x}_i]_j + \eta\beta \cdot \operatorname{sign}([\mathbf{x}_{i-1}]_j - [\mathbf{x}_i]_j), & \text{if } i-1 \in \mathcal{M} \text{ and } |[\mathbf{x}_{i-1}]_j - [\mathbf{x}_i]_j| > \eta\beta, \\ [\mathbf{x}_i]_j - \eta\beta \cdot \operatorname{sign}([\mathbf{x}_{i-1}]_j - [\mathbf{x}_i]_j), & \text{if } i \in \mathcal{M} \text{ and } |[\mathbf{x}_{i-1}]_j - [\mathbf{x}_i]_j| > \eta\beta, \end{cases}$$

which implies

$$\operatorname{supp}(\widetilde{\mathbf{x}}_i) \subset \begin{cases} \operatorname{supp}(\mathbf{x}_i), & \text{if } i-1, i \notin \mathcal{M}, \\ \operatorname{supp}(\mathbf{x}_{i-1}) \cup \operatorname{supp}(\mathbf{x}_i), & \text{if } i-1 \in \mathcal{M}, \\ \operatorname{supp}(\mathbf{x}_i) \cup \operatorname{supp}(\mathbf{x}_{i+1}), & \text{if } i \in \mathcal{M}. \end{cases}$$

The proof is then completed. □

Now we are ready to show the following result on how fast $\operatorname{supp}(\bar{\mathbf{x}}^{(t)})$ can expand with t.

Proposition 2.3 *Under Assumption 2, suppose an algorithm is applied to instance $\mathcal{P}$ from $\mathbf{x}^{(0)} = \mathbf{0}$ and generates a sequence $\{\mathbf{x}^{(t)}\}_{t \ge 0}$. By notations in (2.32), it holds for any $\bar{j} \in \{2, 3, \ldots, \bar{d}\}$ that*

$$\operatorname{supp}(\mathbf{x}_i^{(t)}) \subset \{1, \ldots, \bar{j} - 1\} \textit{ for } i = 1, \ldots, m \textit{ and } t \le 1 + m(\bar{j} - 2)/6. \tag{2.37}$$

Proof. We prove the claim by induction on $\bar{j}$. Let $\boldsymbol{\xi}^{(t)} = (\boldsymbol{\xi}_1^{(t)\top}, \ldots, \boldsymbol{\xi}_m^{(t)\top})^\top$ with $\boldsymbol{\xi}_i^{(t)} \in \mathbb{R}^{\bar{d}}$ and $\boldsymbol{\zeta}^{(t)} = (\boldsymbol{\zeta}_1^{(t)\top}, \ldots, \boldsymbol{\zeta}_m^{(t)\top})^\top$ with $\boldsymbol{\zeta}_i^{(t)} \in \mathbb{R}^{\bar{d}}$ be the vectors defined in Assumption 2 for $t \ge 1$. Since $\mathbf{x}^{(0)} = \mathbf{0}$, we have $\operatorname{supp}(\nabla f_i(\mathbf{x}_i^{(0)})) \subset \{1\}, \forall\, i$ from Lemma 2.7. Notice $\mathbf{b} = \mathbf{0}$. Hence, $\operatorname{supp}(\boldsymbol{\xi}_i^{(1)}) \subset \{1\}, \forall\, i$, which further indicates $\operatorname{supp}(\boldsymbol{\zeta}_i^{(1)}) \subset \{1\}, \forall\, i$ by Lemma 2.8(b), and thus $\operatorname{supp}(\mathbf{x}_i^{(1)}) \subset \{1\}, \forall\, i$. This proves the claim in (2.37) for $\bar{j} = 2$. Now suppose that the claim (2.37) holds for some $\bar{j} \ge 2$. We go to prove it for $\bar{j} + 1$.

According to the hypothesis of the induction, we have

$$\mathrm{supp}(\mathbf{x}_i^{(r)}) \subset \{1,\ldots,\bar{j}-1\}, \forall\, i \in [1,m] \text{ and } \forall\, r \le \bar{t} := 1 + m(\bar{j}-2)/6. \tag{2.38}$$

Below we let $\widehat{\mathbf{x}}^{(r)} = \mathbf{A}^\top \mathbf{A}\mathbf{x}^{(r)}$ for any $r \ge 0$ and consider two cases: $\bar{j}$ is even and $\bar{j}$ is odd.

Case 1: Suppose $\bar{j}$ is even. We claim that, for $s = 0, 1, \ldots, \frac{m}{6}$,

$$\mathrm{supp}(\mathbf{x}_i^{(r)}) \subset \begin{cases} \{1,\ldots,\bar{j}\}, & \text{if } i \in \left[1, \frac{m}{3}+2s\right], \\ \{1,\ldots,\bar{j}-1\}, & \text{if } i \in \left[\frac{m}{3}+2s+1, m\right], \end{cases} \quad \forall\, r \le \bar{t}+s. \tag{2.39}$$

Notice (2.38) implies (2.39) for $s = 0$. Suppose (2.39) holds for some integer $s \in [0, \frac{m}{6}]$. Then by Lemma 2.7 and $\frac{m}{3} + 2s \le \frac{2m}{3}$, it holds

$$\mathrm{supp}(\nabla f_i(\mathbf{x}_i^{(r)})) \subset \begin{cases} \{1,\ldots,\bar{j}\}, & \text{if } i \in \left[1, \frac{m}{3}+2s\right], \\ \{1,\ldots,\bar{j}-1\}, & \text{if } i \in \left[\frac{m}{3}+2s+1, m\right], \end{cases} \quad \forall\, r \le \bar{t}+s.$$

In addition, by Lemma 2.8(a), we have from (2.39) that

$$\mathrm{supp}(\widehat{\mathbf{x}}_i^{(r)}) \subset \begin{cases} \{1,\ldots,\bar{j}\}, & \text{if } i \in \left[1, \frac{m}{3}+2s+1\right], \\ \{1,\ldots,\bar{j}-1\}, & \text{if } i \in \left[\frac{m}{3}+2s+2, m\right], \end{cases} \quad \forall\, r \le \bar{t}+s.$$

Hence, by Assumption 2, we have

$$\mathrm{supp}(\boldsymbol{\xi}_i^{(\bar{t}+s+1)}) \subset \begin{cases} \{1,\ldots,\bar{j}\}, & \text{if } i \in \left[1, \frac{m}{3}+2s+1\right], \\ \{1,\ldots,\bar{j}-1\}, & \text{if } i \in \left[\frac{m}{3}+2s+2, m\right], \end{cases}$$

and thus it follows from Lemma 2.8(b) that

$$\mathrm{supp}(\boldsymbol{\zeta}_i^{(\bar{t}+s+1)}) \subset \begin{cases} \{1,\ldots,\bar{j}\}, & \text{if } i \in \left[1, \frac{m}{3}+2s+2\right], \\ \{1,\ldots,\bar{j}-1\}, & \text{if } i \in \left[\frac{m}{3}+2s+3, m\right]. \end{cases}$$

Now since $\mathbf{x}_i^{(\bar{t}+s+1)} \in \mathbf{span}\left(\{\boldsymbol{\xi}_i^{(\bar{t}+s+1)}, \boldsymbol{\zeta}_i^{(\bar{t}+s+1)}\}\right)$ by Assumption 2, we have

$$\mathrm{supp}(\mathbf{x}_i^{(\bar{t}+s+1)}) \subset \begin{cases} \{1,\ldots,\bar{j}\}, & \text{if } i \in \left[1, \frac{m}{3}+2s+2\right], \\ \{1,\ldots,\bar{j}-1\}, & \text{if } i \in \left[\frac{m}{3}+2s+3, m\right], \end{cases}$$

which means (2.39) holds for $s+1$ as well. By induction, (2.39) holds for $s = 0, 1, \ldots, \frac{m}{6}$. Let $s = \frac{m}{6}$ in (2.39). We have $\mathrm{supp}(\mathbf{x}_i^{(r)}) \subset \{1,\ldots,\bar{j}\}$ for any i and $r \le \bar{t} + \frac{m}{6} = 1 + m(\bar{j}-2)/6 + \frac{m}{6} = 1 + m(\bar{j}-1)/6$.

Case 2: Suppose $\bar{j}$ is odd. We claim that, for $s = 0, 1, \ldots, \frac{m}{6}$,

$$\mathrm{supp}(\mathbf{x}_i^{(r)}) \subset \begin{cases} \{1,\ldots,\bar{j}-1\}, & \text{if } i \in \left[1, \frac{2m}{3}-2s\right], \\ \{1,\ldots,\bar{j}\}, & \text{if } i \in \left[\frac{2m}{3}-2s+1, m\right], \end{cases} \quad \forall\, r \le \bar{t}+s. \tag{2.40}$$

Again (2.38) implies (2.40) for $s = 0$. Suppose it holds for an integer $s \in [0, \frac{m}{6}]$. Then by Lemma 2.7,

$$\mathrm{supp}(\nabla f_i(\mathbf{x}_i^{(r)})) \subset \begin{cases} \{1,\ldots,\bar{j}-1\}, & \text{if } i \in \left[1, \frac{2m}{3}-2s\right], \\ \{1,\ldots,\bar{j}\}, & \text{if } i \in \left[\frac{2m}{3}-2s+1, m\right], \end{cases} \quad \forall\, r \le \bar{t}+s.$$

In addition, by Lemma 2.8(a) and (2.40), we have

$$\mathrm{supp}(\widehat{\mathbf{x}}_i^{(r)}) \subset \begin{cases} \{1,\ldots,\bar{j}-1\}, & \text{if } i \in \left[1, \frac{2m}{3}-2s-1\right], \\ \{1,\ldots,\bar{j}\}, & \text{if } i \in \left[\frac{2m}{3}-2s, m\right], \end{cases} \quad \forall r \le \bar{t}+s.$$

Hence, by Assumption 2, we have

$$\mathrm{supp}(\boldsymbol{\xi}_i^{(\bar{t}+s+1)}) \subset \begin{cases} \{1,\ldots,\bar{j}-1\}, & \text{if } i \in \left[1, \frac{2m}{3}-2s-1\right], \\ \{1,\ldots,\bar{j}\}, & \text{if } i \in \left[\frac{2m}{3}-2s, m\right], \end{cases}$$

and then it follows from Lemma 2.8(b) that

$$\mathrm{supp}(\boldsymbol{\zeta}_i^{(\bar{t}+s+1)}) \subset \begin{cases} \{1,\ldots,\bar{j}-1\}, & \text{if } i \in \left[1, \frac{2m}{3}-2s-2\right], \\ \{1,\ldots,\bar{j}\}, & \text{if } i \in \left[\frac{2m}{3}-2s-1, m\right]. \end{cases}$$

Again since $\mathbf{x}_i^{(\bar{t}+s+1)} \in \mathbf{span}\left(\{\boldsymbol{\xi}_i^{(\bar{t}+s+1)}, \boldsymbol{\zeta}_i^{(\bar{t}+s+1)}\}\right)$ by Assumption 2, we have

$$\mathrm{supp}(\mathbf{x}_i^{(\bar{t}+s+1)}) \subset \begin{cases} \{1,\ldots,\bar{j}-1\}, & \text{if } i \in \left[1, \frac{2m}{3}-2s-2\right], \\ \{1,\ldots,\bar{j}\}, & \text{if } i \in \left[\frac{2m}{3}-2s-1, m\right], \end{cases}$$

which means (2.40) holds for $s+1$ as well. By induction, (2.40) holds for $s = 0, 1, \ldots, \frac{m}{6}$. Let $s = \frac{m}{6}$ in (2.40). We have $\mathrm{supp}(\mathbf{x}_i^{(r)}) \subset \{1,\ldots,\bar{j}\}$ for any i and $r \le \bar{t} + \frac{m}{6} = 1 + m(\bar{j}-2)/6 + \frac{m}{6} = 1 + m(\bar{j}-1)/6$.

Therefore, we have proved that (2.37) holds for $\bar{j}+1$, when $\bar{j}$ is either even or odd. By induction, (2.37) holds for any integer $\bar{j} \in [2, \bar{d}]$, and we complete the proof. □

Finally, we are ready to give our main result about the lower bound of oracle complexity.

Theorem 2.1 *Let $\epsilon > 0$ and $L_f > 0$ be given. Suppose an algorithm is applied to problem (P) that satisfies Assumption 1 and generates a sequence $\{\mathbf{x}^{(t)}\}_{t\ge0}$ that satisfies Assumption 2. Then for any $\omega \in [0, \frac{150\pi\epsilon}{L_f})$, there exists an instance of problem (P), i.e., instance $\mathcal{P}$ in Definition 2, such that the algorithm requires at least $\left\lceil \frac{\kappa([\bar{\mathbf{A}};\mathbf{A}])L_f\Delta_{F_0}}{36000\pi^2}\epsilon^{-2} \right\rceil$ oracles to obtain a point that is ω-close to an ϵ-stationary point of instance $\mathcal{P}$, where $\Delta_{F_0} = F_0(\mathbf{x}^{(0)}) - \inf_{\mathbf{x}} F_0(\mathbf{x})$.*

Proof. As we discussed below (2.32), we assume $\mathbf{x}^{(0)} = \mathbf{0}$ without loss of generality. Thus by notation in (2.32), Proposition 2.3 indicates that $\mathrm{supp}(\mathbf{x}_i^{(t)}) \subset \{1,\ldots,\bar{d}-1\}$ for any $i \in [1,m]$ and any $t \le 1 + m(\bar{d}-2)/6$, which means $[\bar{\mathbf{x}}^{(t)}]_{\bar{d}} = 0$ if $t \le 1 + m(\bar{d}-2)/6$, where $\bar{\mathbf{x}}^{(t)} = \frac{1}{m}\sum_{i=1}^m \mathbf{x}_i^{(t)}$.

On the other hand, suppose $\mathbf{x}^*$ with the structure as in (2.32) is an ϵ-stationary point of instance $\mathcal{P}$. Then by Lemma 2.4, it must also be an ϵ-stationary point of (AP). Hence, by Lemmas 2.5 and 2.6, we have $|[\bar{\mathbf{x}}^*]_j| \ge \frac{150\pi\epsilon}{\sqrt{m}L_f}$ for all $j = 1, \ldots, \bar{d}$, where $\bar{\mathbf{x}}^* = \frac{1}{m}\sum_{i=1}^m \mathbf{x}_i^*$. Therefore, by the convexity of the square function, it follows that

$$\|\mathbf{x}^{(t)} - \mathbf{x}^*\|^2 \ge \sum_{i=1}^m \left([\mathbf{x}_i^{(t)}]_{\bar{d}} - [\mathbf{x}_i^*]_{\bar{d}}\right)^2 \ge m\left([\bar{\mathbf{x}}^{(t)}]_{\bar{d}} - [\bar{\mathbf{x}}^*]_{\bar{d}}\right)^2 \ge m\left(\frac{150\pi\epsilon}{\sqrt{m}L_f}\right)^2 > \omega^2,$$

and thus $\mathbf{x}^{(t)}$ is not ω-close to $\mathbf{x}^*$ if $t \le 1 + m(\bar{d}-2)/6$.

Moreover, by Lemma 2.2(a) and the fact that $g(\mathbf{x}^{(0)}) = 0$ and $g(\mathbf{x}) \geq 0, \forall \mathbf{x}$, it holds that

$$\bar{d} \geq \frac{L_f \left(F_0(\mathbf{x}^{(0)}) - \inf_{\mathbf{x}} F_0(\mathbf{x})\right)}{3000\pi^2}\epsilon^{-2} = \frac{L_f \Delta_{F_0}}{3000\pi^2}\epsilon^{-2}.$$

In other words, in order for $\mathbf{x}^{(t)}$ to be ω-close to an ϵ-stationary point of instant $\mathcal{P}$, the algorithm needs at least $t = 2 + m(\bar{d} - 2)/6$ oracles. We complete the proof by noticing

$$2 + m(\bar{d} - 2)/6 \geq m\bar{d}/12 \geq \frac{mL_f\Delta_{F_0}}{36000\pi^2}\epsilon^{-2} > \frac{\kappa([\bar{\mathbf{A}}; \mathbf{A}])L_f\Delta_{F_0}}{36000\pi^2\epsilon^2},$$

where the first inequality is because $\bar{d} \geq 5$, and the last one is by Lemma 2.3. □

3 Lower Bound of Oracle Complexity for Problems (P) and (SP) under Assumption 3

Under Assumption 2, an algorithm is allowed to call the operator $\mathbf{prox}_{\eta g}(\cdot)$ that may not be easy when g has the structure as that in Assumption 1. Calculating $\mathbf{prox}_{\eta g}(\cdot)$ to a high accuracy or exactly may require many (or even infinitely many) calls to $\bar{\mathbf{A}}$ and $\bar{\mathbf{b}}$. In contrast, $\bar{g}$ is simpler than g, making $\mathbf{prox}_{\eta\bar{g}}(\cdot)$ easier to compute such as when $\bar{g}(\cdot) = \lambda\|\cdot\|_1$ for some $\lambda > 0$ as in instance $\mathcal{P}$. These observations motivate us to reformulate (P) into (SP) and seek an ϵ-stationary point of (SP) under Assumption 3 which only requires the computation of $\mathbf{prox}_{\eta\bar{g}}(\cdot)$. Furthermore, the approach that is designed based on (SP) exhibits the potential to find a near ϵ-stationary point of problem (P), as shown in the next section.

Consequently, two intriguing questions arise: (i) whether finding an ϵ-stationary point of (SP) under Assumption 3 is easier or more challenging compared to finding an ϵ-stationary point of (P) under Assumption 2, and (ii) whether finding a near ϵ-stationary point of (P) under Assumption 3 is easier or harder compared to that under Assumption 2. We provide an answer to the first question, by showing that the same-order lower bound of complexity in Theorem 2.1 holds for finding an ϵ-stationary point of (SP) under Assumption 3. Moreover, we provide an answer to the second question, by showing that the lower bound of oracle complexity for finding a near ϵ-stationary point of (P) is $O(\kappa([\bar{\mathbf{A}}; \mathbf{A}])L_f\Delta_{F_0}\epsilon^{-2})$ under either Assumption 2 or 3; see Theorem 2.1 and Corollary 3.1.

Before we give our main results in this section, we introduce an instance of (SP) that is a reformulation of instant $\mathcal{P}$ in Definition 2. We show that an ϵ-stationary point of the reformulation (SP) of instant $\mathcal{P}$ is a 2ϵ-stationary point of the auxiliary problem (AP) of instant $\mathcal{P}$. This is formally stated in the lemma below.

Lemma 3.1 *Let $\epsilon > 0$ be given in Definition 2 for instance $\mathcal{P}$, and let $\widehat{\epsilon} \in [0, \epsilon]$. Suppose $(\mathbf{x}^*, \mathbf{y}^*)$ is an $\widehat{\epsilon}$-stationary point of the reformulation (SP) of instant $\mathcal{P}$. Then $\mathbf{x}^*$ is a $2\widehat{\epsilon}$-stationary point of the auxiliary problem (AP) of instant $\mathcal{P}$.*

Proof. By Definition 1, there exist $\mathbf{z}_1 \in \mathbb{R}^{\bar{n}}$ and $\mathbf{z}_2 \in \mathbb{R}^n$ such that the conditions in (1.8) hold. Hence, for some $\boldsymbol{\xi} \in \partial\bar{g}(\mathbf{y}^*)$, we have $\|\boldsymbol{\xi} - \mathbf{z}_1\| \leq \widehat{\epsilon}$, and

$$\|\nabla f_0(\mathbf{x}^*) + \bar{\mathbf{A}}^\top\boldsymbol{\xi} + \mathbf{A}^\top\mathbf{z}_2\| \leq \|\nabla f_0(\mathbf{x}^*) + \bar{\mathbf{A}}^\top\mathbf{z}_1 + \mathbf{A}^\top\mathbf{z}_2\| + \|\bar{\mathbf{A}}\|\widehat{\epsilon} \leq \widehat{\epsilon} + \|\bar{\mathbf{A}}\|\widehat{\epsilon}. \tag{3.1}$$

Moreover, (1.8) implies that, for any $\mathbf{v} \in \mathbf{Null}(\mathbf{A})$, we have

$$\begin{aligned} F'(\mathbf{x}^*, \mathbf{y}^*; \mathbf{v}, \bar{\mathbf{A}}\mathbf{v}) =& \mathbf{v}^\top\nabla f_0(\mathbf{x}^*) + \bar{g}'(\mathbf{y}^*; \bar{\mathbf{A}}\mathbf{v}) \\ \geq& \mathbf{v}^\top\nabla f_0(\mathbf{x}^*) + \mathbf{v}^\top\bar{\mathbf{A}}^\top\boldsymbol{\xi} = \mathbf{v}^\top\left(\nabla f_0(\mathbf{x}^*) + \bar{\mathbf{A}}^\top\boldsymbol{\xi} + \mathbf{A}^\top\mathbf{z}_2\right) \geq -\widehat{\epsilon}(1 + \|\bar{\mathbf{A}}\|)\|\mathbf{v}\|, \end{aligned} \tag{3.2}$$

where the first inequality follows from (2.22) with g replaced by $\bar{g}$, and the second one is by Cauchy-Schwarz inequality and (3.1).

Below we prove $\mathbf{y}^* = \mathbf{0}$. We write it into the block-structured form

$$\mathbf{y}^* = (\mathbf{y}_1^{*\top}, \ldots, \mathbf{y}_{3m_2-1}^{*\top})^\top \text{ with } \mathbf{y}_i^* \in \mathbb{R}^{\bar{d}}, \forall\, i = 1, 2, \ldots, 3m_2 - 1.$$

If $\mathbf{y}^* \neq \mathbf{0}$, then $\mathbf{y}_{\bar{i}}^* \neq \mathbf{0}$ for some $\bar{i} \in \{1, 2, \ldots, 3m_2 - 1\}$. Let $\mathbf{v}^* = (\mathbf{v}_1^{*\top}, \mathbf{v}_2^{*\top}, \ldots, \mathbf{v}_m^{*\top})^\top$ where $\mathbf{v}_i^* = \mathbf{y}_{\bar{i}}^*/(mL_f)$ for $i \leq \bar{i}m_1$, and $\mathbf{v}_i^* = \mathbf{0}$ otherwise. We then have $\mathbf{v}_i^* = \mathbf{v}_{i+1}^*$ for any $i \neq \bar{i}m_1$, so $\mathbf{A}\mathbf{v}^* = \mathbf{0}$ by Proposition 2.1(c). Moreover, let $\mathbf{u}^* = \bar{\mathbf{A}}\mathbf{v}^* = (\mathbf{u}_1^{*\top}, \ldots, \mathbf{u}_{3m_2-1}^{*\top})^\top$ with $\mathbf{u}_i^* \in \mathbb{R}^{\bar{d}}$ for $i = 1, 2, \ldots, 3m_2 - 1$. We must have $\mathbf{u}_i^* = -\mathbf{y}_{\bar{i}}^*$ for $i = \bar{i}$ and $\mathbf{u}_i^* = \mathbf{0}$ for $i \neq \bar{i}$. Therefore by (2.5), for any $s \in (0, 1)$,

$$\bar{g}\left(\mathbf{y}^* + s\mathbf{u}^*\right) = \frac{\beta}{mL_f} \sum_{i<\bar{i}} \|\mathbf{y}_i^*\|_1 + \frac{\beta}{mL_f} \|\mathbf{y}_{\bar{i}}^* - s\mathbf{y}_{\bar{i}}^*\|_1 + \frac{\beta}{mL_f} \sum_{i>\bar{i}} \|\mathbf{y}_i^*\|_1 = \bar{g}\left(\mathbf{y}^*\right) - \frac{s\beta}{mL_f} \|\mathbf{y}_{\bar{i}}^*\|_1. \tag{3.3}$$

Now by (2.25), the choice of $\mathbf{v}_i^*$, and the mean value theorem, we have for any $i = 1, \ldots, m$ that

$$f_i(\mathbf{x}_i^* + s\mathbf{v}_i^*) - f_i(\mathbf{x}_i^*) = s\nabla f_i(\mathbf{x}_i^* + s'\mathbf{v}_i^*)^\top \mathbf{v}_i^* \leq s\|\nabla f_i(\mathbf{x}_i^* + s'\mathbf{v}_i^*)\|_\infty \|\mathbf{v}_i^*\|_1 \leq \frac{50s\pi\epsilon}{\sqrt{m}} \cdot \frac{\|\mathbf{y}_{\bar{i}}^*\|_1}{mL_f},$$

where $s' \in (0, s)$. The inequality above, together with (3.3) and the definition of f_0 in (2.12), implies

$$\begin{aligned}
&\frac{1}{s}\Big(F(\mathbf{x}^* + s\mathbf{v}^*, \mathbf{y}^* + s\mathbf{u}^*) - F(\mathbf{x}^*, \mathbf{y}^*)\Big) = \frac{1}{s}\Big(f_0(\mathbf{x}^* + s\mathbf{v}^*) - f_0(\mathbf{x}^*) + \bar{g}(\mathbf{y}^* + s\mathbf{u}^*) - \bar{g}(\mathbf{y}^*)\Big) \\
=&\frac{1}{s}\left(\sum_{i=1}^m \left(f_i(\mathbf{x}_i^* + s\mathbf{v}_i^*) - f_i(\mathbf{x}_i^*)\right) + \bar{g}(\mathbf{y}^* + s\mathbf{u}^*) - \bar{g}(\mathbf{y}^*)\right) \\
\leq&\frac{1}{smL_f}\left(50\pi s\epsilon\sqrt{m}\|\mathbf{y}_{\bar{i}}^*\|_1 - \beta s\|\mathbf{y}_{\bar{i}}^*\|_1\right) = \frac{\left(50\pi\epsilon\sqrt{m} - \beta\right)\|\mathbf{y}_{\bar{i}}^*\|_1}{mL_f}.
\end{aligned}$$

Letting $s \downarrow 0$ in the inequality above gives $\frac{\left(50\pi\epsilon\sqrt{m} - \beta\right)\|\mathbf{y}_{\bar{i}}^*\|_1}{mL_f} \geq F'(\mathbf{x}^*, \mathbf{y}^*; \mathbf{v}^*, \mathbf{u}^*)$. Recall $\mathbf{A}\mathbf{v}^* = \mathbf{0}$ and $\mathbf{u}^* = \bar{\mathbf{A}}\mathbf{v}^*$. Hence, we have from (3.2) and the choice of $\mathbf{v}^*$ that

$$\frac{\left(50\pi\epsilon\sqrt{m} - \beta\right)\|\mathbf{y}_{\bar{i}}^*\|_1}{mL_f} \geq -\widehat{\epsilon}(1 + \|\bar{\mathbf{A}}\|)\|\mathbf{v}^*\| \geq -\widehat{\epsilon}(1 + \|\bar{\mathbf{A}}\|)\frac{\sqrt{\bar{i}}\|\mathbf{y}_{\bar{i}}^*\|}{mL_f} \geq -\widehat{\epsilon}(1 + \|\bar{\mathbf{A}}\|)\frac{\sqrt{m}\|\mathbf{y}_{\bar{i}}^*\|_1}{mL_f}.$$

Since $\beta > (50\pi + 1 + \|\mathbf{A}\|)\sqrt{m}\epsilon$, $\epsilon \geq \widehat{\epsilon}$ and $\|\mathbf{A}\| \geq \|\bar{\mathbf{A}}\|$, the inequalities above can hold only when $\mathbf{y}_{\bar{i}}^* = \mathbf{0}$. This contradicts to the hypothesis $\mathbf{y}_{\bar{i}}^* \neq \mathbf{0}$. Hence, $\mathbf{y}^* = \mathbf{0}$, which together with (1.8) gives $\|\bar{\mathbf{A}}\mathbf{x}^*\| \leq \widehat{\epsilon}$. Furthermore, $\|\mathbf{A}\mathbf{x}^*\| \leq \widehat{\epsilon}$ from (1.8). Thus $\|\mathbf{H}\mathbf{x}^*\| \leq \|\bar{\mathbf{A}}\mathbf{x}^*\| + \|\mathbf{A}\mathbf{x}^*\| \leq 2\widehat{\epsilon}$, which, together with $\|\nabla f_0(\mathbf{x}^*) + \bar{\mathbf{A}}^\top \mathbf{z}_1 + \mathbf{A}^\top \mathbf{z}_2\| \leq \widehat{\epsilon}$ from (1.8), indicates that $\mathbf{x}^*$ a $2\widehat{\epsilon}$-stationary point of (AP). □

With Lemma 3.1, we can establish the lower-bound complexity for solving problem (SP) in a similar way to show Theorem 2.1. In particular, Lemma 3.1 indicates that a solution $(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})$ cannot be an $\epsilon/2$-stationary point of the reformulation (SP) of instant $\mathcal{P}$, if $\mathbf{x}^{(t)}$ is not an ϵ-stationary point of (AP). By Lemmas 2.5 and 2.6, if there exists $\bar{j} \in \{1, 2, \ldots, \bar{d}\}$ such that $[\bar{\mathbf{x}}^{(t)}]_{\bar{j}} = 0$, where $\bar{\mathbf{x}}^{(t)} = \frac{1}{m}\sum_{i=1}^m \mathbf{x}_i^{(t)}$, then $\mathbf{x}^{(t)}$ cannot be an ϵ-stationary point of problem (AP) and thus cannot be an $\epsilon/2$-stationary point of the reformulation (SP) of instant $\mathcal{P}$. Finally, similar to Proposition 2.3, we can show that, for any algorithm that is applied to the reformulation (SP) of instance $\mathcal{P}$, if it starts from $(\mathbf{x}^{(0)}, \mathbf{y}^{(0)}) = (\mathbf{0}, \mathbf{0})$ and generates a sequence $\{(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\}_{t\geq 0}$ satisfying Assumption 3, then $\mathrm{supp}(\mathbf{x}^{(t)}) \subset \{1, \ldots, \bar{j}-1\}$ for some $\bar{j} \in \{1, 2, \ldots, \bar{d}\}$ if t is not large enough. Hence, $\mathbf{x}^{(t)}$ cannot be ω-close to any ϵ-stationary point of instance $\mathcal{P}$ for any $\omega \in [0, \frac{150\pi\epsilon}{L_f})$, according to the proof of Theorem 2.1. This way, we can obtain a lower bound of oracle complexity to produce an ϵ-stationary point of (SP) and also a lower bound to obtain a near ϵ-stationary point of (P). Since the aforementioned arguments are similar to those for proving Theorem 2.1, we simply present the lower complexity bounds in the theorem and the corollary below and put the proofs in Appendix A.

Theorem 3.1 *Let $\epsilon > 0$ and $L_f > 0$ be given. Suppose an algorithm is applied to problem (SP) that satisfies Assumption 1(a, c) and $\inf_{\mathbf{x},\mathbf{y}} F(\mathbf{x}, \mathbf{y}) > -\infty$. Suppose the generated sequence $\{(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\}_{t\geq 0}$ satisfies Assumption 3. Then there exists an instance of problem (SP), i.e., the reformulation (SP) of instance $\mathcal{P}$ given in Definition 2, such that the algorithm requires at least $\left\lceil \frac{\kappa([\bar{\mathbf{A}};\mathbf{A}])L_f \Delta_F}{72000\pi^2}\epsilon^{-2} \right\rceil$ oracles to obtain an ϵ-stationary point of that instance, where $\Delta_F = F(\mathbf{x}^{(0)}, \mathbf{y}^{(0)}) - \inf_{\mathbf{x},\mathbf{y}} F(\mathbf{x}, \mathbf{y})$.*

Corollary 3.1 *Let $\epsilon > 0$ and $L_f > 0$ be given. Suppose an algorithm is applied to problem (P) that satisfies Assumption 1 and generates a sequence $\{(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\}_{t\geq 0}$ that satisfies Assumption 3. Then for any $\omega \in [0, \frac{150\pi\epsilon}{L_f})$, there exists an instance of problem (P), i.e., instance $\mathcal{P}$ in Definition 2, such that the algorithm requires at least $\left\lceil \frac{\kappa([\bar{\mathbf{A}};\mathbf{A}])L_f \Delta_{F_0}}{18000\pi^2}\epsilon^{-2} \right\rceil$ oracles to obtain a point $\mathbf{x}^{(t)}$ that is ω-close to an ϵ-stationary point of instance $\mathcal{P}$, where $\Delta_{F_0} = F_0(\mathbf{x}^{(0)}) - \inf_{\mathbf{x}} F_0(\mathbf{x})$.*

4 Tightness of Lower Bounds of Oracle Complexity under Assumption 3

As we discussed in Section 1.2, various existing FOMs can be applied to (P) and (SP) under a non-smooth non-convex setting. However, the best-known upper bound of oracle complexity is $O\left(\kappa^2([\mathbf{A};\bar{\mathbf{A}}])L_f^2\Delta\epsilon^{-2}\right)$ (see [9]) with $\Delta = \Delta_{F_0}$ or Δ_F, which does not match our lower bounds. To close the gap between the upper and lower complexity bounds, we present a new *inexact proximal gradient* (IPG) method in this section that falls in the class of algorithms under Assumption 3. The oracle complexity of the IPG matches the lower bounds in Theorem 3.1 and Corollary 3.1, up to logarithmic factors, under a few more assumptions. More precisely, we will need Assumption 4 and either (but not necessary both) of Assumptions 5 and 6 below[5].

Assumption 4 $\inf_{\mathbf{x},\mathbf{y}} F(\mathbf{x}, \mathbf{y}) > -\infty$*; f_0 is l_f-Lipschitz continuous; $\bar{g}$ is l_g-Lipschitz continuous; $\mathbf{A}$ has a full-row rank. There exists a feasible point $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ of (SP) such that $\bar{\mathbf{y}}$ is in the relative interior of* $\mathrm{dom}(\bar{g})$.

Assumption 5 $[\bar{\mathbf{A}};\mathbf{A}]$ *has a full-row rank.*

Assumption 6 $\bar{g}(\mathbf{y}) = \max\{\mathbf{u}^\top\mathbf{y} : \mathbf{C}\mathbf{u} \leq \mathbf{d}, \mathbf{u} \in \mathbb{R}^{\bar{n}}\}$ *for some* $\mathbf{C}$ *and* $\mathbf{d}$.

We point out that instance $\mathcal{P}$ given in Definition 2 satisfies Assumptions 1, 4, 5 and 6. By Lemma 2.2, f_0 is $50\pi\epsilon\sqrt{m\bar{d}}$-Lipschitz continuous. Since $\bar{d} = \Theta(\epsilon^{-2})$ as mentioned in Theorem 2.1, the Lipschitz constant of f_0 is independent of ϵ. Also, the Lipschitz constant of $\bar{g}$ is $l_g = \sqrt{(3m_2-1)\bar{d}}\frac{\beta}{mL_f} = \Theta(\sqrt{3m_2\bar{d}}\epsilon/\sqrt{m})$, which is not dependent on ϵ either. In addition, $\mathrm{dom}(\bar{g})$ is the whole space. These observations, together with Lemmas 2.1 and 2.2, imply that Assumption 4 is satisfied by instance $\mathcal{P}$. Moreover, it can be easily checked

[5] To claim (near) tightness of our lower bounds, we also need $\mathbf{b} = \mathbf{0}$ and $\bar{\mathbf{b}} = \mathbf{0}$ under Assumptions 4 and 6.

that Assumptions 5 and 6 are also satisfied by instance $\mathcal{P}$. Hence, the lower bound $O(\kappa([\bar{\mathbf{A}};\mathbf{A}])L_f\varDelta\epsilon^{-2})$ with $\varDelta=\varDelta_F$ or $\varDelta_{F_0}$ remains valid for problems (P) and (SP) even with these additional assumptions, indicating that the oracle complexity of the IPG method (resp. the established lower bound) under these assumptions is optimal (resp. tight).

4.1 A New Inexact Proximal Gradient Method

The IPG method we propose generates a sequence $\{(\mathbf{x}^{(k)},\mathbf{y}^{(k)})\}$ by

$$\begin{aligned}(\mathbf{x}^{(k+1)},\mathbf{y}^{(k+1)})\approx \operatorname*{arg\,min}_{\mathbf{x},\mathbf{y}}\ &\underbrace{\left\langle \nabla f_0(\mathbf{x}^{(k)}),\mathbf{x}-\mathbf{x}^{(k)}\right\rangle+\frac{\tau}{2}\left\|\mathbf{x}-\mathbf{x}^{(k)}\right\|^2}_{:=\bar{f}(\mathbf{x})}+\bar{g}(\mathbf{y})\\ \text{s.t.}\ \ &\mathbf{y}=\bar{\mathbf{A}}\mathbf{x}+\bar{\mathbf{b}},\ \ \mathbf{A}\mathbf{x}+\mathbf{b}=\mathbf{0},\end{aligned}\tag{4.1}$$

for each $k\geq 0$, where $\tau\geq L_f$. Solving the minimization problem in (4.1) exactly can be difficult due to the coexistence of affine constraints and the regularization term. To obtain the inexact solution $(\mathbf{x}^{(k+1)},\mathbf{y}^{(k+1)})$ efficiently to a desired accuracy, we consider the following Lagrangian function of the problem in (4.1)

$$\mathcal{L}_k(\mathbf{x},\mathbf{y},\mathbf{z})=\bar{f}(\mathbf{x})+\bar{g}(\mathbf{y})-\mathbf{z}_1^\top\left(\mathbf{y}-\left(\bar{\mathbf{A}}\mathbf{x}+\bar{\mathbf{b}}\right)\right)+\mathbf{z}_2^\top\left(\mathbf{A}\mathbf{x}+\mathbf{b}\right),\tag{4.2}$$

where $\mathbf{z}=(\mathbf{z}_1^\top,\mathbf{z}_2^\top)^\top$, $\mathbf{z}_1\in\mathbb{R}^{\bar{n}}$ and $\mathbf{z}_2\in\mathbb{R}^n$ are dual variables. Let $\mathcal{D}_k$ be the negative Lagrangian dual function[6], i.e.,

$$\mathcal{D}_k(\mathbf{z}):=-\min_{\mathbf{x},\mathbf{y}}\mathcal{L}_k(\mathbf{x},\mathbf{y},\mathbf{z})=\frac{1}{2\tau}\|\bar{\mathbf{A}}^\top\mathbf{z}_1+\mathbf{A}^\top\mathbf{z}_2+\nabla f_0(\mathbf{x}^{(k)})-\tau\mathbf{x}^{(k)}\|^2+\bar{g}^\star(\mathbf{z}_1)-\mathbf{z}_1^\top\bar{\mathbf{b}}-\mathbf{z}_2^\top\mathbf{b},\forall\,\mathbf{z},$$

where $\bar{g}^\star$ is the convex conjugate function of $\bar{g}$, i.e., $\bar{g}^\star(\mathbf{z}_1)=\max_{\mathbf{y}}\{\mathbf{y}^\top\mathbf{z}_1-\bar{g}(\mathbf{y})\}$. We then define

$$\varOmega^{(k+1)}:=\operatorname*{Arg\,min}_{\mathbf{z}}\mathcal{D}_k(\mathbf{z})\quad\text{ and }\quad\mathcal{D}_k^*:=\min_{\mathbf{z}}\mathcal{D}_k(\mathbf{z}).\tag{4.3}$$

Note that $\mathbf{prox}_{\eta\bar{g}^\star}(\mathbf{z}_1)=\mathbf{z}_1-\mathbf{prox}_{\eta\bar{g}(\cdot/\eta)}(\mathbf{z}_1)=\mathbf{z}_1-\eta\mathbf{prox}_{\eta^{-1}\bar{g}(\cdot)}(\mathbf{z}_1/\eta)$ [33]. Thus the proximal operator $\mathbf{prox}_{\eta\bar{g}^\star}(\mathbf{z})$ can be calculated easily if $\mathbf{prox}_{\eta\bar{g}(\cdot/\eta)}(\mathbf{z}_1)$ can be. Moreover, when Assumption 5 holds, the objective function in (4.3) is strongly convex composite, so an *accelerated proximal gradient* (APG) method, e.g., the one in [29], can approach $\varOmega^{(k+1)}$ in (4.3) at a linear rate. When Assumption 6 holds, it is shown in [26, Theorem 10] that the objective function in (4.3) has a quadratic growth, so a restarted APG method can approach an optimal solution at a linear rate. This motivates us to apply a (restarted) APG method to find a nearly optimal solution of problem in (4.3), which is then used to obtain $(\mathbf{x}^{(k+1)},\mathbf{y}^{(k+1)})$.

More specifically, we find a near-optimal point $\mathbf{z}^{(k+1)}=((\mathbf{z}_1^{(k+1)})^\top,(\mathbf{z}_2^{(k+1)})^\top)^\top$ of $\min_{\mathbf{z}}\mathcal{D}_k(\mathbf{z})$ such that

$$\mathbf{z}_1^{(k+1)}\in\operatorname{dom}(\bar{g}^\star)\quad\text{ and }\quad\operatorname{dist}(\mathbf{z}^{(k+1)},\varOmega^{(k+1)})\leq\delta,\tag{4.4}$$

where δ is a small number (to be specified). Then, we obtain a primal solution from $\mathbf{z}^{(k+1)}$ by

$$\mathbf{x}^{(k+1)}=\mathbf{x}^{(k)}-\frac{1}{\tau}\left(\bar{\mathbf{A}}^\top\mathbf{z}_1^{(k+1)}+\mathbf{A}^\top\mathbf{z}_2^{(k+1)}+\nabla f_0(\mathbf{x}^{(k)})\right),\tag{4.5}$$

[6] For ease of discussion, we formulate the Lagrangian dual problem into a minimization one by negating the dual function.

$$\mathbf{y}^{(k+1)} = \mathbf{prox}_{\sigma^{-1}\bar{g}}\left(\sigma^{-1}\mathbf{z}_1^{(k+1)} + \bar{\mathbf{A}}\mathbf{x}^{(k+1)} + \bar{\mathbf{b}}\right). \tag{4.6}$$

This procedure is presented in Algorithm 1.

Algorithm 1 An inexact proximal gradient (IPG) method for problem (SP)

```
1: Input: a feasible initial point $(\mathbf{x}^{(0)}, \mathbf{y}^{(0)})$, $\sigma > 0$, $\epsilon > 0$, and $\tau > L_f$.
2: Choose $\delta > 0$ and let $k \leftarrow 0$.
3: while an $\epsilon$-stationary point of problem (SP) is not obtained do
4:     Calculate a near-optimal point $\mathbf{z}^{(k+1)}$ of problem $\min_{\mathbf{z}} \mathcal{D}_k(\mathbf{z})$ that satisfies the conditions in (4.4).
5:     Set $\mathbf{x}^{(k+1)}$ by (4.5) and $\mathbf{y}^{(k+1)}$ by (4.6).
6:     Let $k \leftarrow k + 1$.
7: end while
8: Output: $(\mathbf{x}^{(k)}, \mathbf{y}^{(k)})$.
```

4.2 Number of Outer Iterations for Finding an ϵ-stationary Point

To characterize the convergence property of Algorithm 1, we need the following two lemmas.

Lemma 4.1 *Suppose that Assumption 4 holds. For any $\sigma > 0$ and any $\bar{\mathbf{z}}^{(k+1)} \in \Omega^{(k+1)}$, let $(\bar{\mathbf{x}}^{(k+1)}, \bar{\mathbf{y}}^{(k+1)})$ be the optimal solution of the strongly convex problem*

$$\min_{\mathbf{x},\mathbf{y}} \left\{ \mathcal{L}_k(\mathbf{x}, \mathbf{y}, \bar{\mathbf{z}}^{(k+1)}) + \frac{\sigma}{2}\|\mathbf{y} - (\bar{\mathbf{A}}\mathbf{x} + \bar{\mathbf{b}})\|^2 + \frac{\sigma}{2}\|\mathbf{A}\mathbf{x} + \mathbf{b}\|^2 \right\}, \tag{4.7}$$

where $\mathcal{L}_k$ and $\Omega^{(k+1)}$ are defined in (4.2) and (4.3), respectively. Then it holds that for any $k \geq 0$,

$$\bar{\mathbf{y}}^{(k+1)} - (\bar{\mathbf{A}}\bar{\mathbf{x}}^{(k+1)} + \bar{\mathbf{b}}) = \mathbf{0}, \quad \mathbf{A}\bar{\mathbf{x}}^{(k+1)} + \mathbf{b} = \mathbf{0}, \tag{4.8}$$

$$\bar{\mathbf{x}}^{(k+1)} = \mathbf{x}^{(k)} - \frac{1}{\tau}\left(\bar{\mathbf{A}}^\top \bar{\mathbf{z}}_1^{(k+1)} + \mathbf{A}^\top \bar{\mathbf{z}}_2^{(k+1)} + \nabla f_0(\mathbf{x}^{(k)})\right), \tag{4.9}$$

and

$$\bar{\mathbf{y}}^{(k+1)} = \mathbf{prox}_{\sigma^{-1}\bar{g}}\left(\sigma^{-1}\bar{\mathbf{z}}_1^{(k+1)} + \bar{\mathbf{A}}\bar{\mathbf{x}}^{(k+1)} + \bar{\mathbf{b}}\right). \tag{4.10}$$

Proof. Let $(\widehat{\mathbf{x}}^{(k+1)}, \widehat{\mathbf{y}}^{(k+1)})$ be an optimal solution of the problem in (4.1). Under Assumption 4, the strong duality holds [3, Section 5.2.3]. Then the optimal objective value of the minimization problem in (4.1) is $\max_{\mathbf{z}} \mathcal{L}_k(\widehat{\mathbf{x}}^{(k+1)}, \widehat{\mathbf{y}}^{(k+1)}, \mathbf{z})$. By the definition of $\bar{\mathbf{z}}^{(k+1)}$ and the strong duality, it holds that

$$\mathcal{L}_k(\widehat{\mathbf{x}}^{(k+1)}, \widehat{\mathbf{y}}^{(k+1)}, \bar{\mathbf{z}}^{(k+1)}) \geq \min_{\mathbf{x},\mathbf{y}} \mathcal{L}_k(\mathbf{x}, \mathbf{y}, \bar{\mathbf{z}}^{(k+1)}) = \max_{\mathbf{z}} \mathcal{L}_k(\widehat{\mathbf{x}}^{(k+1)}, \widehat{\mathbf{y}}^{(k+1)}, \mathbf{z}) \geq \mathcal{L}_k(\widehat{\mathbf{x}}^{(k+1)}, \widehat{\mathbf{y}}^{(k+1)}, \bar{\mathbf{z}}^{(k+1)}).$$

Hence, the inequalities above must hold with equalities. Thus $(\widehat{\mathbf{x}}^{(k+1)}, \widehat{\mathbf{y}}^{(k+1)}) \in \operatorname{Arg\,min}_{\mathbf{x},\mathbf{y}} \mathcal{L}_k(\mathbf{x}, \mathbf{y}, \bar{\mathbf{z}}^{(k+1)})$. By this fact and also that $(\widehat{\mathbf{x}}^{(k+1)}, \widehat{\mathbf{y}}^{(k+1)})$ solves the problem in (4.1), we obtain

$$\begin{aligned}
\mathcal{L}_k(\widehat{\mathbf{x}}^{(k+1)}, \widehat{\mathbf{y}}^{(k+1)}, \bar{\mathbf{z}}^{(k+1)}) &= \min_{\mathbf{x},\mathbf{y}} \mathcal{L}_k(\mathbf{x}, \mathbf{y}, \bar{\mathbf{z}}^{(k+1)}) \leq \mathcal{L}_k(\bar{\mathbf{x}}^{(k+1)}, \bar{\mathbf{y}}^{(k+1)}, \bar{\mathbf{z}}^{(k+1)}) \\
&\leq \mathcal{L}_k(\bar{\mathbf{x}}^{(k+1)}, \bar{\mathbf{y}}^{(k+1)}, \bar{\mathbf{z}}^{(k+1)}) + \frac{\sigma}{2}\|\bar{\mathbf{y}}^{(k+1)} - (\bar{\mathbf{A}}\bar{\mathbf{x}}^{(k+1)} + \bar{\mathbf{b}})\|^2 + \frac{\sigma}{2}\|\mathbf{A}\bar{\mathbf{x}}^{(k+1)} + \mathbf{b}\|^2 \\
&= \min_{\mathbf{x},\mathbf{y}} \left\{ \mathcal{L}_k(\mathbf{x}, \mathbf{y}, \bar{\mathbf{z}}^{(k+1)}) + \frac{\sigma}{2}\|\mathbf{y} - (\bar{\mathbf{A}}\mathbf{x} + \bar{\mathbf{b}})\|^2 + \frac{\sigma}{2}\|\mathbf{A}\mathbf{x} + \mathbf{b}\|^2 \right\}
\end{aligned}$$

$$\begin{aligned}&\leq \mathcal{L}_k(\widehat{\mathbf{x}}^{(k+1)}, \widehat{\mathbf{y}}^{(k+1)}, \bar{\mathbf{z}}^{(k+1)}) + \frac{\sigma}{2}\|\widehat{\mathbf{y}}^{(k+1)} - (\bar{\mathbf{A}}\widehat{\mathbf{x}}^{(k+1)} + \bar{\mathbf{b}})\|^2 + \frac{\sigma}{2}\|\mathbf{A}\widehat{\mathbf{x}}^{(k+1)} + \mathbf{b}\|^2\\ &= \mathcal{L}_k(\widehat{\mathbf{x}}^{(k+1)}, \widehat{\mathbf{y}}^{(k+1)}, \bar{\mathbf{z}}^{(k+1)}),\end{aligned}$$

where the second equality is by the definition of $(\bar{\mathbf{x}}^{(k+1)}, \bar{\mathbf{y}}^{(k+1)})$. Hence all the inequalities above must hold with equalities. Thus (4.8) follows and $(\bar{\mathbf{x}}^{(k+1)}, \bar{\mathbf{y}}^{(k+1)}) \in \operatorname{Arg\,min}_{\mathbf{x},\mathbf{y}} \mathcal{L}_k(\mathbf{x}, \mathbf{y}, \bar{\mathbf{z}}^{(k+1)})$, the optimality condition of which gives (4.9) and (4.10). This completes the proof. □

Remark 4.1 The proof of Lemma 4.1 also implies that $(\bar{\mathbf{x}}^{(k+1)}, \bar{\mathbf{y}}^{(k+1)})$ is an optimal solution of the problem in (4.1). The lemma below bounds the inexactness of $(\mathbf{x}^{(k+1)}, \mathbf{y}^{(k+1)})$ generated in Algorithm 1.

Lemma 4.2 *Suppose that Assumption 4 holds. Let $\{(\mathbf{x}^{(k+1)}, \mathbf{y}^{(k+1)})\}_{k\geq 0}$ be generated from Algorithm 1. Denote the vector used to produce $(\mathbf{x}^{(k+1)}, \mathbf{y}^{(k+1)})$ in (4.4)-(4.6) by $\mathbf{z}^{(k+1)} = ((\mathbf{z}_1^{(k+1)})^\top, (\mathbf{z}_2^{(k+1)})^\top)^\top$. Let $(\bar{\mathbf{x}}^{(k+1)}, \bar{\mathbf{y}}^{(k+1)})$ be defined as in Lemma 4.1 with $\bar{\mathbf{z}}^{(k+1)} = \mathbf{proj}_{\Omega^{(k+1)}}(\mathbf{z}^{(k+1)})$. Then the following inequalities hold for all $k \geq 0$:*

$$\|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k+1)}\| \leq \frac{1}{\tau}\left\|[\bar{\mathbf{A}}; \mathbf{A}]\right\|\delta, \quad \|\bar{\mathbf{y}}^{(k+1)} - \mathbf{y}^{(k+1)}\| \leq \frac{1}{\tau}\left\|\bar{\mathbf{A}}\right\|\left\|[\bar{\mathbf{A}}; \mathbf{A}]\right\|\delta + \sigma^{-1}\delta, \tag{4.11}$$

$$\|\mathbf{y}^{(k+1)} - (\bar{\mathbf{A}}\mathbf{x}^{(k+1)} + \bar{\mathbf{b}})\| + \|\mathbf{A}\mathbf{x}^{(k+1)} + \mathbf{b}\| \leq B_1\delta, \tag{4.12}$$

$$\|\mathbf{z}_1^{(k+1)}\| \leq l_g, \quad \|\mathbf{z}_2^{(k+1)}\| \leq B_2 + B_3\delta, \tag{4.13}$$

$$\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| \leq B_4 + \frac{1}{\tau}\left\|[\bar{\mathbf{A}}; \mathbf{A}]\right\| B_3\delta, \tag{4.14}$$

where B_1, B_2, B_3, B_4 are some constants defined by

$$\begin{aligned}&B_1 := \tfrac{1}{\tau}\|\bar{\mathbf{A}}\|\left\|[\bar{\mathbf{A}}; \mathbf{A}]\right\| + \sigma^{-1} + \tfrac{1}{\tau}(\|\bar{\mathbf{A}}\| + \|\mathbf{A}\|)\left\|[\bar{\mathbf{A}}; \mathbf{A}]\right\|, \ B_2 := \|(\mathbf{A}\mathbf{A}^\top)^{-1}\mathbf{A}\|(l_f + \|\bar{\mathbf{A}}\|l_g),\\ &B_3 := (1 + \|(\mathbf{A}\mathbf{A}^\top)^{-1}\mathbf{A}\|\left\|[\bar{\mathbf{A}}; \mathbf{A}]\right\|), \ B_4 := \tfrac{1}{\tau}\left(l_f + \left\|[\bar{\mathbf{A}}; \mathbf{A}]\right\|(l_g + B_2)\right).\end{aligned}$$

Proof. By (4.4), we have $\|\mathbf{z}^{(k+1)} - \bar{\mathbf{z}}^{(k+1)}\| \leq \delta$. Then we obtain from (4.5) and (4.9) that

$$\|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k+1)}\| \leq \frac{1}{\tau}\left\|[\bar{\mathbf{A}}; \mathbf{A}]\right\|\|\mathbf{z}^{(k+1)} - \bar{\mathbf{z}}^{(k+1)}\| \leq \frac{1}{\tau}\left\|[\bar{\mathbf{A}}; \mathbf{A}]\right\|\delta,$$

and from (4.4), (4.6) and (4.10) that

$$\|\bar{\mathbf{y}}^{(k+1)} - \mathbf{y}^{(k+1)}\| \leq \|\bar{\mathbf{A}}\|\|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k+1)}\| + \sigma^{-1}\|\mathbf{z}^{(k+1)} - \bar{\mathbf{z}}^{(k+1)}\| \leq \frac{1}{\tau}\left\|\bar{\mathbf{A}}\right\|\left\|[\bar{\mathbf{A}}; \mathbf{A}]\right\|\delta + \sigma^{-1}\delta.$$

Hence, the two inequalities in (4.11) hold.

In addition, by (4.8), we have

$$\begin{aligned}&\|\mathbf{y}^{(k+1)} - (\bar{\mathbf{A}}\mathbf{x}^{(k+1)} + \bar{\mathbf{b}})\| + \|\mathbf{A}\mathbf{x}^{(k+1)} + \mathbf{b}\|\\ =& \left\|\mathbf{y}^{(k+1)} - (\bar{\mathbf{A}}\mathbf{x}^{(k+1)} + \bar{\mathbf{b}}) - \bar{\mathbf{y}}^{(k+1)} + (\bar{\mathbf{A}}\bar{\mathbf{x}}^{(k+1)} + \bar{\mathbf{b}})\right\| + \left\|(\mathbf{A}\bar{\mathbf{x}}^{(k+1)} + \mathbf{b}) - (\mathbf{A}\mathbf{x}^{(k+1)} + \mathbf{b})\right\|\\ \leq& \left\|\mathbf{y}^{(k+1)} - \bar{\mathbf{y}}^{(k+1)}\right\| + \left\|\bar{\mathbf{A}}\mathbf{x}^{(k+1)} - \bar{\mathbf{A}}\bar{\mathbf{x}}^{(k+1)}\right\| + \left\|\mathbf{A}\mathbf{x}^{(k+1)} - \mathbf{A}\bar{\mathbf{x}}^{(k+1)}\right\|\\ \leq& \left\|\mathbf{y}^{(k+1)} - \bar{\mathbf{y}}^{(k+1)}\right\| + \left(\|\bar{\mathbf{A}}\| + \|\mathbf{A}\|\right)\left\|\mathbf{x}^{(k+1)} - \bar{\mathbf{x}}^{(k+1)}\right\| \leq \delta B_1,\end{aligned}$$

where the last inequality is from (4.11) and the definition of B_1. Hence, the claim in (4.12) holds.

Moreover, recall $\mathbf{z}_1^{(k+1)} \in \operatorname{dom}(\bar{g}^\star)$ in (4.4). Since $\bar{g}$ is l_g-Lipschitz continuous by Assumption 4, we must have $\|\mathbf{z}_1^{(k+1)}\| \leq l_g$ [33]. Write $\bar{\mathbf{z}}^{(k+1)} = ((\bar{\mathbf{z}}_1^{(k+1)})^\top, (\bar{\mathbf{z}}_2^{(k+1)})^\top)^\top$. Then $\bar{\mathbf{z}}_1^{(k+1)} \in \operatorname{dom}(\bar{g}^\star)$. Thus for the same reason, we have $\|\bar{\mathbf{z}}_1^{(k+1)}\| \leq l_g$. By Assumption 4, $\mathbf{A}\mathbf{A}^\top$ is non-singular, so the optimality condition of (4.3) implies

$$\bar{\mathbf{z}}_2^{(k+1)} = -\left(\mathbf{A}\mathbf{A}^\top\right)^{-1}\left(\mathbf{A}\bar{\mathbf{A}}^\top\bar{\mathbf{z}}_1^{(k+1)} + \mathbf{A}\nabla f_0(\mathbf{x}^{(k)}) - \tau\mathbf{A}\mathbf{x}^{(k)} - \tau\mathbf{b}\right). \tag{4.15}$$

For $k \geq 1$, we obtain from the second inequality in (4.8) and the first inequality in (4.11) that

$$\begin{aligned}\|\left(\mathbf{A}\mathbf{A}^\top\right)^{-1}(\mathbf{A}\mathbf{x}^{(k)} + \mathbf{b})\| =& \|\left(\mathbf{A}\mathbf{A}^\top\right)^{-1}(\mathbf{A}\mathbf{x}^{(k)} + \mathbf{b} - \mathbf{A}\bar{\mathbf{x}}^{(k)} - \mathbf{b})\| \\ &\leq \|(\mathbf{A}\mathbf{A}^\top)^{-1}\mathbf{A}\|\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)}\| \leq \frac{1}{\tau}\|(\mathbf{A}\mathbf{A}^\top)^{-1}\mathbf{A}\|\left\|[\bar{\mathbf{A}};\mathbf{A}]\right\|\delta.\end{aligned}$$

Since $\mathbf{A}\mathbf{x}^{(0)} + \mathbf{b} = \mathbf{0}$, the inequality above also holds for $k = 0$. Hence,

$$\begin{aligned}\|\mathbf{z}_2^{(k+1)}\| \leq& \|\mathbf{z}_2^{(k+1)} - \bar{\mathbf{z}}_2^{(k+1)}\| + \|\bar{\mathbf{z}}_2^{(k+1)}\| \leq \delta + \left\|\left(\mathbf{A}\mathbf{A}^\top\right)^{-1}\left(\mathbf{A}\bar{\mathbf{A}}^\top\bar{\mathbf{z}}_1^{(k+1)} + \mathbf{A}\nabla f_0(\mathbf{x}^{(k)}) - \tau\mathbf{A}\mathbf{x}^{(k)} - \tau\mathbf{b}\right)\right\| \\ \leq& \delta + \|(\mathbf{A}\mathbf{A}^\top)^{-1}\mathbf{A}\|(l_f + \|\bar{\mathbf{A}}\|l_g) + \|(\mathbf{A}\mathbf{A}^\top)^{-1}\mathbf{A}\|\left\|[\bar{\mathbf{A}};\mathbf{A}]\right\|\delta \leq B_2 + B_3\delta.\end{aligned}$$

Thus both inequalities in (4.13) hold.

Finally, it follows from the updating rule in (4.5) and the inequalities in (4.13) that

$$\begin{aligned}\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| =& \frac{1}{\tau}\left\|[\bar{\mathbf{A}};\mathbf{A}]^\top\mathbf{z}^{(k+1)} + \nabla f_0(\mathbf{x}^{(k)})\right\| \\ \leq& \frac{1}{\tau}\left(\left\|[\bar{\mathbf{A}};\mathbf{A}]\right\|\left\|\mathbf{z}_1^{(k+1)}\right\| + \left\|[\bar{\mathbf{A}};\mathbf{A}]\right\|\left\|\mathbf{z}_2^{(k+1)}\right\| + \left\|\nabla f_0(\mathbf{x}^{(k)})\right\|\right) \\ \leq& \frac{1}{\tau}l_f + \frac{1}{\tau}\left\|[\bar{\mathbf{A}};\mathbf{A}]\right\|l_g + \frac{1}{\tau}\left\|[\bar{\mathbf{A}};\mathbf{A}]\right\|(B_2 + B_3\delta).\end{aligned}$$

Hence by the definition of B_4, (4.14) is obtained, and we complete the proof. □

With Lemmas 4.1 and 4.2, we can characterize the number of outer iterations that Algorithm 1 needs to find an ϵ-stationary point of problem (SP).

Theorem 4.1 *Suppose that Assumptions 1 and 4 hold. Given $\epsilon > 0$, in Algorithm 1, let $\tau = 2L_f$ and $\delta = \delta_\epsilon$ with*

$$\delta_\epsilon := \min\left\{\frac{\epsilon}{B_1\sigma}, \frac{\epsilon}{B_1}, \frac{\epsilon^2}{48L_fB_1\left(B_2 + \sigma\|\bar{\mathbf{A}}\|B_4 + l_g\right)}, \sqrt{\frac{\epsilon^2}{48L_fB_1B_3\left(1 + \frac{1}{\tau}\sigma\|\bar{\mathbf{A}}\|\left\|[\bar{\mathbf{A}};\mathbf{A}]\right\|\right)}}\right\}, \tag{4.16}$$

where B_1, B_2, B_3 and B_4 are given in Lemma 4.2. Let

$$K_\epsilon := \left\lceil 12L_f\Delta_F\epsilon^{-2}\right\rceil. \tag{4.17}$$

For each $k \geq 0$, let $(\mathbf{x}^{(k)}, \mathbf{y}^{(k)})$ be generated from Algorithm 1. Let $k' = \arg\min_{k=0,\ldots,K_\epsilon-1}\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\|$. Then $(\mathbf{x}^{(k'+1)}, \mathbf{y}^{(k'+1)})$ is an ϵ-stationary point of problem (SP).

Proof. Since $\delta \leq \frac{\epsilon}{B_1}$ from (4.16), we obtain from inequality (4.12) that, for any $k \geq 0$,

$$\|\mathbf{y}^{(k+1)} - (\bar{\mathbf{A}}\mathbf{x}^{(k+1)} + \bar{\mathbf{b}})\| + \|\mathbf{A}\mathbf{x}^{(k+1)} + \mathbf{b}\| \leq \delta B_1 \leq \epsilon. \tag{4.18}$$

In addition, by (4.6), it holds for any $k \geq 0$ that

$$\mathbf{0} \in \partial \bar{g}(\mathbf{y}^{(k+1)}) - \mathbf{z}_1^{(k+1)} + \sigma\left(\mathbf{y}^{(k+1)} - (\bar{\mathbf{A}}\mathbf{x}^{(k+1)} + \bar{\mathbf{b}})\right), \tag{4.19}$$

which implies that there exists $\boldsymbol{\xi}^{(k+1)} \in \partial \bar{g}(\mathbf{y}^{(k+1)})$ such that

$$\|\boldsymbol{\xi}^{(k+1)} - \mathbf{z}_1^{(k+1)}\| = \sigma\|\mathbf{y}^{(k+1)} - (\bar{\mathbf{A}}\mathbf{x}^{(k+1)} + \bar{\mathbf{b}})\| \overset{(4.12)}{\leq} \delta B_1 \sigma \overset{(4.16)}{\leq} \epsilon. \tag{4.20}$$

Moreover, it holds by (4.5) that for any $k \geq 0$,

$$\mathbf{0} = \nabla f_0(\mathbf{x}^{(k)}) + [\bar{\mathbf{A}}; \mathbf{A}]^\top \mathbf{z}^{(k+1)} + \tau(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}). \tag{4.21}$$

Hence, by the L_f-Lipschitz continuity of ∇f_0 and the fact that $\tau = 2L_f$, we have for any $k \geq 0$,

$$\begin{aligned}\left\|\nabla f_0(\mathbf{x}^{(k+1)}) + [\bar{\mathbf{A}}; \mathbf{A}]^\top \mathbf{z}^{(k+1)}\right\| \leq & \left\|\nabla f_0(\mathbf{x}^{(k)}) + [\bar{\mathbf{A}}; \mathbf{A}]^\top \mathbf{z}^{(k+1)}\right\| + L_f\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| \\ &= (\tau + L_f)\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| = 3L_f\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|.\end{aligned} \tag{4.22}$$

Below we bound the average of the square of the right-hand side of (4.22) over K terms. First, by (4.12) and the feasibility of $(\mathbf{x}^{(0)}, \mathbf{y}^{(0)})$, i.e, $\mathbf{A}\mathbf{x}^{(0)} + \mathbf{b} = \mathbf{0}$ and $\mathbf{y}^{(0)} = \bar{\mathbf{A}}\mathbf{x}^{(0)} + \bar{\mathbf{b}}$, we have for any $k \geq 0$,

$$\left\|\mathbf{y}^{(k+1)} - \mathbf{y}^{(k)} - \bar{\mathbf{A}}(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)})\right\| = \left\|\mathbf{y}^{(k+1)} - (\bar{\mathbf{A}}\mathbf{x}^{(k+1)} + \bar{\mathbf{b}}) - \mathbf{y}^{(k)} + (\bar{\mathbf{A}}\mathbf{x}^{(k)} + \bar{\mathbf{b}})\right\| \leq 2\delta B_1, \tag{4.23}$$

$$\left\|\mathbf{A}(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)})\right\| = \left\|(\mathbf{A}\mathbf{x}^{(k)} + \mathbf{b}) - (\mathbf{A}\mathbf{x}^{(k+1)} + \mathbf{b})\right\| \leq 2\delta B_1. \tag{4.24}$$

Second, it follows from the L_f-Lipschitz continuity of ∇f_0 that

$$\begin{aligned} & f_0(\mathbf{x}^{(k+1)}) - f_0(\mathbf{x}^{(k)}) \leq \left\langle \nabla f_0(\mathbf{x}^{(k)}), \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} \right\rangle + \frac{L_f}{2}\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\|^2 \\ \overset{(4.21)}{=} & \left\langle -\tau\left(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right) - \bar{\mathbf{A}}^\top \mathbf{z}_1^{(k+1)} - \mathbf{A}^\top \mathbf{z}_2^{(k+1)}, \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} \right\rangle + \frac{L_f}{2}\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\|^2 \\ = & \frac{L_f - 2\tau}{2}\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\|^2 - \left\langle \bar{\mathbf{A}}^\top \boldsymbol{\xi}^{(k+1)}, \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} \right\rangle - \left\langle \mathbf{z}_2^{(k+1)}, \mathbf{A}(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}) \right\rangle \\ & - \left\langle \bar{\mathbf{A}}^\top (\mathbf{z}_1^{(k+1)} - \boldsymbol{\xi}^{(k+1)}), \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} \right\rangle \\ \leq & -\frac{3L_f}{2}\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\|^2 - \left\langle \bar{\mathbf{A}}^\top \boldsymbol{\xi}^{(k+1)}, \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} \right\rangle + 2\delta B_1\|\mathbf{z}_2^{(k+1)}\| + 2\delta B_1\sigma\|\bar{\mathbf{A}}\|\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| \\ \leq & -\frac{3L_f}{2}\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\|^2 - \left\langle \bar{\mathbf{A}}^\top \boldsymbol{\xi}^{(k+1)}, \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} \right\rangle + 2\delta B_1\left(B_2 + B_3\delta\right) \\ & + 2\delta B_1\sigma\|\bar{\mathbf{A}}\|\left(B_4 + \frac{1}{\tau}\left\|[\bar{\mathbf{A}}; \mathbf{A}]\right\| B_3\delta\right), \end{aligned} \tag{4.25}$$

where the second inequality holds from $\tau = 2L_f$, (4.20) and (4.24), and the last inequality follows from (4.13) and (4.14). Third, by the convexity of $\bar{g}$ and the fact that $\boldsymbol{\xi}^{(k+1)} \in \partial \bar{g}(\mathbf{y}^{(k+1)})$, we have

$$\begin{aligned} &\bar{g}(\mathbf{y}^{(k+1)}) - \bar{g}(\mathbf{y}^{(k)}) \leq \left\langle \boldsymbol{\xi}^{(k+1)}, \mathbf{y}^{(k+1)} - \mathbf{y}^{(k)} \right\rangle \\ =& \left\langle \boldsymbol{\xi}^{(k+1)}, \bar{\mathbf{A}}(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}) \right\rangle + \left\langle \boldsymbol{\xi}^{(k+1)}, \mathbf{y}^{(k+1)} - \mathbf{y}^{(k)} - \bar{\mathbf{A}}(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}) \right\rangle \\ \leq& \left\langle \boldsymbol{\xi}^{(k+1)}, \bar{\mathbf{A}}(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}) \right\rangle + 2\delta B_1 \|\boldsymbol{\xi}^{(k+1)}\| \\ \leq& \left\langle \boldsymbol{\xi}^{(k+1)}, \bar{\mathbf{A}}(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}) \right\rangle + 2\delta B_1 l_g, \end{aligned} \tag{4.26}$$

where the second inequality is from (4.23), and the last one holds by $\|\boldsymbol{\xi}^{(k+1)}\| \leq l_g$ from the l_g-Lipschitz continuity of $\bar{g}$. Adding (4.25) and (4.26) and combining terms give

$$\begin{aligned} & f_0(\mathbf{x}^{(k+1)}) + \bar{g}(\mathbf{y}^{(k+1)}) - f_0(\mathbf{x}^{(k)}) - \bar{g}(\mathbf{y}^{(k)}) \\ \leq & -\frac{3L_f}{2} \left\| \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} \right\|^2 + 2\delta B_1 \left(B_2 + \sigma \|\bar{\mathbf{A}}\| B_4 + l_g \right) + 2\delta^2 B_1 B_3 \left(1 + \frac{1}{\tau} \sigma \|\bar{\mathbf{A}}\| \left\| [\bar{\mathbf{A}}; \mathbf{A}] \right\| \right). \end{aligned}$$

Multiplying $3L_f$ to the inequality above and summing it over $k = 0, 1, \ldots, K-1$, we obtain

$$\begin{aligned} \frac{(3L_f)^2}{K} \sum_{k=0}^{K-1} \left\| \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} \right\|^2 \leq & \frac{6L_f \left(F(\mathbf{x}^{(0)}, \mathbf{y}^{(0)}) - \inf_{\mathbf{x},\mathbf{y}} F(\mathbf{x}, \mathbf{y}) \right)}{K} \\ & + 12 L_f \delta B_1 \left(B_2 + \sigma \|\bar{\mathbf{A}}\| B_4 + l_g \right) + 12 L_f \delta^2 B_1 B_3 \left(1 + \frac{1}{\tau} \sigma \|\bar{\mathbf{A}}\| \left\| [\bar{\mathbf{A}}; \mathbf{A}] \right\| \right), \forall K \geq 1. \end{aligned} \tag{4.27}$$

Let $K = K_\epsilon$ in (4.27). Since $K_\epsilon \geq 12 L_f \left(F(\mathbf{x}^{(0)}, \mathbf{y}^{(0)}) - \inf_{\mathbf{x},\mathbf{y}} F(\mathbf{x}, \mathbf{y}) \right) \epsilon^{-2}$, it holds that

$$\frac{6L_f \left(F(\mathbf{x}^{(0)}, \mathbf{y}^{(0)}) - \inf_{\mathbf{x},\mathbf{y}} F(\mathbf{x}, \mathbf{y}) \right)}{K_\epsilon} \leq \frac{1}{2} \epsilon^2. \tag{4.28}$$

Moreover, the choice of $\delta = \delta_\epsilon$ in (4.16) ensures

$$12 L_f \delta B_1 \left(B_2 + \sigma \|\bar{\mathbf{A}}\| B_4 + l_g \right) \leq \frac{1}{4} \epsilon^2 \quad \text{and} \quad 12 L_f \delta^2 B_1 B_3 \left(1 + \frac{1}{\tau} \sigma \|\bar{\mathbf{A}}\| \left\| [\bar{\mathbf{A}}; \mathbf{A}] \right\| \right) \leq \frac{1}{4} \epsilon^2. \tag{4.29}$$

Applying the bounds in (4.28) and (4.29) to (4.27) with $K = K_\epsilon$ gives

$$\frac{(3L_f)^2}{K_\epsilon} \sum_{k=0}^{K_\epsilon - 1} \left\| \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} \right\|^2 \leq \epsilon^2. \tag{4.30}$$

Hence, it follows that $3L_f \min_{k=0,\ldots,K_\epsilon-1} \left\| \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} \right\| \leq \epsilon$. Then it results from the definition of k' and (4.22) that

$$\left\| \nabla f_0(\mathbf{x}^{(k'+1)}) + [\bar{\mathbf{A}}; \mathbf{A}]^\top \mathbf{z}^{(k'+1)} \right\| \leq 3L_f \left\| \mathbf{x}^{(k'+1)} - \mathbf{x}^{(k')} \right\| = 3L_f \min_{k=0,\ldots,K_\epsilon-1} \left\| \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} \right\| \leq \epsilon, \tag{4.31}$$

which together with (4.18) and (4.20) shows that $(\mathbf{x}^{(k'+1)}, \mathbf{y}^{(k'+1)})$ is an ϵ-stationary point of problem (SP) by Definition 1. This completes the proof. □

Theorem 4.1 above only shows that Algorithm 1 can produce an ϵ-stationary point of problem (SP). In fact, with a δ that is slightly smaller than δ_ϵ in Algorithm 1, we can also show that Algorithm 1 can produce a near ϵ-stationary point of (P) with an outer-iteration number similar to that in Theorem 4.1. This result is presented in the following theorem.

Theorem 4.2 *Suppose that Assumptions 1 and 4 hold. Given $\epsilon > 0$, in Algorithm 1, let $\tau = 2L_f$ and $\delta = \bar{\delta}_\epsilon := \min\left\{\frac{\epsilon}{6\left\|[\bar{\mathbf{A}};\mathbf{A}]\right\|}, \frac{\Delta_{F_0}}{B_1 l_g}, \delta_\epsilon\right\}$, where B_1 is given in Lemma 4.2 and δ_ϵ is given in (4.16). Let*

$$\bar{K}_\epsilon := \lceil 192 L_f \Delta_{F_0} \epsilon^{-2} \rceil.$$

For each $k \geq 0$, let $(\mathbf{x}^{(k)}, \mathbf{y}^{(k)})$ and $\mathbf{z}^{(k)}$ be generated from Algorithm 1, $(\bar{\mathbf{x}}^{(k+1)}, \bar{\mathbf{y}}^{(k+1)})$ be defined as in Lemma 4.1 with $\bar{\mathbf{z}}^{(k+1)} = \mathbf{proj}_{\Omega^{(k+1)}}(\mathbf{z}^{(k+1)})$. Let $k' = \underset{k=0,\ldots,\bar{K}_\epsilon-1}{\arg\min} \left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\|$. Then $(\mathbf{x}^{(k'+1)}, \mathbf{y}^{(k'+1)})$ is an ϵ-stationary point of problem (SP), $\bar{\mathbf{x}}^{(k'+1)}$ is an ϵ-stationary point of problem (P), and $\mathbf{x}^{(k'+1)}$ is a near ϵ-stationary point of problem (P). More specifically, $\|\mathbf{x}^{(k'+1)} - \bar{\mathbf{x}}^{(k'+1)}\| \leq \frac{\epsilon}{12L_f} \in [0, \frac{150\pi\epsilon}{L_f})$.

Proof. Notice $\mathbf{0} = \nabla f_0(\mathbf{x}^{(k)}) + [\bar{\mathbf{A}}; \mathbf{A}]^\top \bar{\mathbf{z}}^{(k+1)} + \tau(\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k)})$ from (4.9). Hence through the same arguments to obtain (4.22), we arrive at

$$\left\|\nabla f_0(\bar{\mathbf{x}}^{(k+1)}) + [\bar{\mathbf{A}}; \mathbf{A}]^\top \bar{\mathbf{z}}^{(k+1)}\right\| \leq 3L_f \left\|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k)}\right\|, \forall\, k \geq 0. \tag{4.32}$$

By the triangle inequality, it holds $\left\|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k)}\right\| \leq \left\|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k+1)}\right\| + \left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\|$. Thus from (4.11), we have $\left\|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k)}\right\| \leq \left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\| + \frac{1}{\tau}\left\|[\bar{\mathbf{A}}; \mathbf{A}]\right\| \delta$, which together with (4.32) and $\tau = 2L_f$ gives

$$\left\|\nabla f_0(\bar{\mathbf{x}}^{(k+1)}) + [\bar{\mathbf{A}}; \mathbf{A}]^\top \bar{\mathbf{z}}^{(k+1)}\right\| \leq 3L_f \left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\| + \tfrac{3}{2}\left\|[\bar{\mathbf{A}}; \mathbf{A}]\right\| \delta, \forall\ k \geq 0. \tag{4.33}$$

By using the same method to obtain argument (4.27) and the definition of $\bar{\delta}_\epsilon$, we have

$$\frac{(3L_f)^2}{K} \sum_{k=0}^{K-1} \left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\|^2 \leq \frac{6L_f\left(F(\mathbf{x}^{(0)}, \mathbf{y}^{(0)}) - [f_0(\mathbf{x}^{(K+1)}) + \bar{g}(\mathbf{y}^{(K+1)})]\right)}{K} + \frac{\epsilon^2}{2}, \forall\, K \geq 1. \tag{4.34}$$

Recall that $(\mathbf{x}^{(0)}, \mathbf{y}^{(0)})$ is a feasible point of problem (SP). We have

$$\begin{aligned}
&F(\mathbf{x}^{(0)}, \mathbf{y}^{(0)}) - \left[f_0(\mathbf{x}^{(K+1)}) + \bar{g}(\mathbf{y}^{(K+1)})\right] \\
=&\left(f_0(\mathbf{x}^{(0)}) + \bar{g}(\bar{\mathbf{A}}\mathbf{x}^{(0)} + \bar{\mathbf{b}}) - \left[f_0(\mathbf{x}^{(K+1)}) + \bar{g}(\bar{\mathbf{A}}\mathbf{x}^{(K+1)} + \bar{\mathbf{b}})\right]\right) + \left[\bar{g}(\bar{\mathbf{A}}\mathbf{x}^{(K+1)} + \bar{\mathbf{b}}) - \bar{g}(\mathbf{y}^{(K+1)})\right] \\
\leq&\Delta_{F_0} + l_g\|\bar{\mathbf{A}}\mathbf{x}^{(K+1)} + \bar{\mathbf{b}} - \mathbf{y}^{(K+1)}\| \\
\leq&\Delta_{F_0} + \Delta_{F_0} = 2\Delta_{F_0}, \quad \forall\, K \geq 1,
\end{aligned}$$

where the first inequality is because $\bar{g}$ is l_g-Lipschitz continuous by Assumption 4 and the second inequality is because $\|\mathbf{y}^{(K+1)} - (\bar{\mathbf{A}}\mathbf{x}^{(K+1)} + \bar{\mathbf{b}})\| \leq \bar{\delta}_\epsilon B_1 \leq \frac{\Delta_{F_0}}{l_g}$ according to (4.18) and the definition of $\bar{\delta}_\epsilon$. Now substituting the above inequality into (4.34) with $K = \bar{K}_\epsilon = \lceil 192 L_f \Delta_{F_0} \epsilon^{-2} \rceil$, we obtain

$$\frac{(3L_f)^2}{K} \sum_{k=0}^{K-1} \left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\|^2 \leq \frac{12 L_f \Delta_{F_0}}{192 L_f \Delta_{F_0} \epsilon^{-2}} + \frac{\epsilon^2}{2} = \frac{\epsilon^2}{16} + \frac{\epsilon^2}{2} = \frac{9\epsilon^2}{16}.$$

Since $k' = \arg\min_{k=0,\dots,\bar{K}_\epsilon-1} \left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\|$, we have $3L_f \left\|\mathbf{x}^{(k'+1)} - \mathbf{x}^{(k')}\right\| \le \frac{3\epsilon}{4}$. Then, it follows from the same arguments in the proof of Theorem 4.1 that $(\mathbf{x}^{(k'+1)}, \mathbf{y}^{(k'+1)})$ is an ϵ-stationary point of problem (SP).

Also, we obtain from (4.33) and the choice of δ that

$$\left\|\nabla f_0(\bar{\mathbf{x}}^{(k'+1)}) + [\bar{\mathbf{A}}; \mathbf{A}]^\top \bar{\mathbf{z}}^{(k'+1)}\right\| \le \epsilon. \tag{4.35}$$

On the other hand, we have from (4.8) and (4.10) that

$$\bar{\mathbf{y}}^{(k'+1)} = (\bar{\mathbf{A}}\bar{\mathbf{x}}^{(k'+1)} + \bar{\mathbf{b}}), \quad \mathbf{A}\bar{\mathbf{x}}^{(k'+1)} + \mathbf{b} = \mathbf{0}, \quad \bar{\mathbf{z}}_1^{(k'+1)} \in \partial\bar{g}(\bar{\mathbf{y}}^{(k'+1)}). \tag{4.36}$$

Recall $g(\mathbf{x}) = \bar{g}(\bar{\mathbf{A}}\mathbf{x} + \bar{\mathbf{b}})$. By [8], it holds that $\partial g(\bar{\mathbf{x}}^{(k'+1)}) = \overline{\mathbf{co}}\left(\left\{\bar{\mathbf{A}}^\top \boldsymbol{\xi} : \boldsymbol{\xi} \in \partial\bar{g}\left(\bar{\mathbf{A}}\bar{\mathbf{x}}^{(k'+1)} + \bar{\mathbf{b}}\right)\right\}\right)$, and thus from (4.36), it follows $\bar{\mathbf{A}}^\top \bar{\mathbf{z}}_1^{(k'+1)} \in \partial g(\bar{\mathbf{x}}^{(k'+1)})$. Then by (4.35), we have

$$\operatorname{dist}\left(\mathbf{0}, \nabla f_0\left(\bar{\mathbf{x}}^{(k'+1)}\right) + \partial g\left(\bar{\mathbf{x}}^{(k'+1)}\right) + \mathbf{A}^\top \bar{\mathbf{z}}_2^{(k'+1)}\right) \le \left\|\nabla f_0\left(\bar{\mathbf{x}}^{(k'+1)}\right) + \bar{\mathbf{A}}^\top \bar{\mathbf{z}}_1^{(k'+1)} + \mathbf{A}^\top \bar{\mathbf{z}}_2^{(k'+1)}\right\| \le \epsilon,$$

which, together with $\mathbf{A}\bar{\mathbf{x}}^{(k'+1)} + \mathbf{b} = \mathbf{0}$ from (4.36), indicates that $\bar{\mathbf{x}}^{(k'+1)}$ is an ϵ-stationary point of (P).

By (4.11), $\|\bar{\mathbf{x}}^{(k'+1)} - \mathbf{x}^{(k'+1)}\| \le \frac{1}{\tau}\left\|[\bar{\mathbf{A}}; \mathbf{A}]\right\| \delta_\epsilon \le \frac{\epsilon}{12L_f}$, where the second inequality is from the definition of $\bar{\delta}_\epsilon$ which ensures $\bar{\delta}_\epsilon \le \frac{\epsilon}{6\|[\bar{\mathbf{A}};\mathbf{A}]\|}$. □

4.3 Number of Inner Iterations for Finding $\mathbf{z}^{(k+1)}$ Satisfying (4.4)

To obtain the total number of inner iterations of Algorithm 1 to find an ϵ-stationary point of (SP), we still need to evaluate the number of iterations for computing $\mathbf{z}^{(k+1)}$ that satisfies the criterion in (4.4) for each $k \ge 0$. The problem in (4.3) has the convex composite structure. Additionally, the gradient of the smooth function $\mathcal{D}_k(\mathbf{z}) - \bar{g}^\star(\mathbf{z}_1)$ in the objective function $\mathcal{D}_k$ in (4.3) is $L_\mathcal{D}$-Lipschitz continuous with

$$L_\mathcal{D} := \lambda_{\max}([\bar{\mathbf{A}}; \mathbf{A}][\bar{\mathbf{A}}; \mathbf{A}]^\top)/\tau.$$

Hence, we apply the APG method in [1] with a standard restarting technique to (4.3) to find $\mathbf{z}^{(k+1)}$. The restarted APG algorithm instantiated on (4.3) is presented in Algorithm 2. The algorithm has double loops and, in addition to the main iterate $\mathbf{z}^{(j,k)}$, it also generates an auxiliary iterate $\widehat{\mathbf{z}}^{(j,k)}$. In the algorithm, we use $\widehat{\mathcal{G}}_1^j$ and $\widehat{\mathcal{G}}_2^j$ to represent the gradients of $\mathcal{D}_k(\mathbf{z}) - \bar{g}^\star(\mathbf{z}_1)$ with respect to $\mathbf{z}_1$ and $\mathbf{z}_2$, respectively, at $\mathbf{z} = \widehat{\mathbf{z}}^{(j,k)}$. The algorithm is restarted after every j_k APG steps. We call each APG step, i.e., Line 5-9 in Algorithm 2, as one inner iteration of Algorithm 1.

Algorithm 2 Restarted accelerated proximal gradient method for (4.3)

```
1: Input: an initial point $\mathbf{z}^{\text{ini}} = ((\mathbf{y}^{(0)})^\top, (\mathbf{A}\mathbf{x}^{(0)})^\top)^\top$ where $(\mathbf{x}^{(0)}, \mathbf{y}^{(0)})$ is the same as in Algorithm 1.
2: $\mathbf{z}^{(0,k)} \leftarrow \mathbf{z}^{\text{ini}}$, $\widehat{\mathbf{z}}^{(0,k)} \leftarrow \mathbf{z}^{(0,k)}$, $\alpha_0 \leftarrow 1$.
3: for $i = 0, \dots, i_k - 1$ do
4:     for $j = 0, \dots, j_k - 1$ do
5:         $\widehat{\mathcal{G}}_1^j \leftarrow \frac{1}{\tau}\bar{\mathbf{A}}\left(\bar{\mathbf{A}}^\top \widehat{\mathbf{z}}_1^{(j,k)} + \mathbf{A}^\top \widehat{\mathbf{z}}_2^{(j,k)} + \nabla f_0(\mathbf{x}^{(k)}) - \tau \mathbf{x}^{(k)}\right) - \bar{\mathbf{b}}$.
6:         $\widehat{\mathcal{G}}_2^j \leftarrow \frac{1}{\tau}\mathbf{A}\left(\bar{\mathbf{A}}^\top \widehat{\mathbf{z}}_1^{(j,k)} + \mathbf{A}^\top \widehat{\mathbf{z}}_2^{(j,k)} + \nabla f_0(\mathbf{x}^{(k)}) - \tau \mathbf{x}^{(k)}\right) - \mathbf{b}$.
7:         $\mathbf{z}_1^{(j+1,k)} \leftarrow \mathbf{prox}_{L_{\mathcal{D}}^{-1}\bar{g}^*}\left(\widehat{\mathbf{z}}_1^{(j,k)} - \frac{1}{L_{\mathcal{D}}}\widehat{\mathcal{G}}_1^j\right)$, $\mathbf{z}_2^{(j+1,k)} \leftarrow \widehat{\mathbf{z}}_2^{(j,k)} - \frac{1}{L_{\mathcal{D}}}\widehat{\mathcal{G}}_2^j$.
8:         $\alpha_{j+1} \leftarrow \frac{1+\sqrt{1+4\alpha_j^2}}{2}$.
9:         $\widehat{\mathbf{z}}^{(j+1,k)} \leftarrow \mathbf{z}^{(j+1,k)} + \left(\frac{\alpha_j - 1}{\alpha_{j+1}}\right)\left(\mathbf{z}^{(j+1,k)} - \mathbf{z}^{(j,k)}\right)$.
10:    end for
11:    $\mathbf{z}^{(0,k)} \leftarrow \mathbf{z}^{(j_k,k)}$, $\widehat{\mathbf{z}}^{(0,k)} \leftarrow \mathbf{z}^{(0,k)}$, $\alpha_0 \leftarrow 1$.
12: end for
13: Output: $\mathbf{z}^{(k+1)} = \mathbf{z}^{(0,k)}$.
```

The number of APG steps performed in Algorithm 2 to find a point satisfying (4.4) is known when Assumption 4 and either one of Assumptions 5 and 6 hold. We first present the result when Assumptions 4 and 5 hold.

Theorem 4.3 *Suppose Assumptions 1, 4 and 5 hold. Given any $\delta > 0$, if $j_k = \lceil 2\sqrt{2}\kappa([\bar{\mathbf{A}}; \mathbf{A}])\rceil$ and $i_k = \left\lceil \log_2(\frac{2(\mathcal{D}_k(\mathbf{z}^{ini}) - \mathcal{D}_k^*)}{\mu_{\mathcal{D}}\delta^2})\right\rceil$ in Algorithm 2, the output $\mathbf{z}^{(k+1)}$ satisfies the criteria in (4.4). Here, $\mu_{\mathcal{D}} = \lambda_{\min}([\bar{\mathbf{A}}; \mathbf{A}][\bar{\mathbf{A}}; \mathbf{A}]^\top)/\tau$.*

Proof. With Assumption 5, the problem in (4.3) is $\mu_{\mathcal{D}}$-strongly convex. Consider the end of the first inner loop (i.e., $i = 0$) of Algorithm 2. According to [1, Theorem 4.4.], after j_k APG steps, the APG method generates an output $\mathbf{z}^{(j_k,k)}$ satisfying

$$\mathcal{D}_k(\mathbf{z}^{(j_k,k)}) - \mathcal{D}_k^* \leq \frac{2L_{\mathcal{D}}\text{dist}^2(\mathbf{z}^{\text{ini}}, \Omega^{(k+1)})}{j_k^2} \leq \frac{4L_{\mathcal{D}}[\mathcal{D}_k(\mathbf{z}^{\text{ini}}) - \mathcal{D}_k^*]}{\mu_{\mathcal{D}} j_k^2} = \frac{4\kappa^2([\bar{\mathbf{A}}; \mathbf{A}])[\mathcal{D}_k(\mathbf{z}^{\text{ini}}) - \mathcal{D}_k^*]}{j_k^2},$$

where the second inequality is by the $\mu_{\mathcal{D}}$-strong convexity of $\mathcal{D}_k$. Since $j_k = \lceil 2\sqrt{2}\kappa([\bar{\mathbf{A}}; \mathbf{A}])\rceil$, we have $\mathcal{D}_k(\mathbf{z}^{(j_k,k)}) - \mathcal{D}_k^* \leq \frac{1}{2}\left(\mathcal{D}_k(\mathbf{z}^{\text{ini}}) - \mathcal{D}_k^*\right)$, meaning that we can reduce the objective gap by a half with each inner loop of Algorithm 2. Since the next inner loop is started at $\mathbf{z}^{(j_k,k)}$ according to Line 11 of Algorithm 2, we repeat this argument $i_k = \left\lceil \log_2(\frac{2(\mathcal{D}_k(\mathbf{z}^{\text{ini}}) - \mathcal{D}_k^*)}{\mu_{\mathcal{D}}\delta^2})\right\rceil$ times and show that the final output $\mathbf{z}^{(k+1)}$ satisfies $\mathcal{D}_k(\mathbf{z}^{(k+1)}) - \mathcal{D}_k^* \leq \frac{\mu_{\mathcal{D}}\delta^2}{2}$ and thus

$$\text{dist}(\mathbf{z}^{(k+1)}, \Omega^{(k+1)}) \leq \sqrt{\frac{2(\mathcal{D}_k(\mathbf{z}^{(k+1)}) - \mathcal{D}_k^*)}{\mu_{\mathcal{D}}}} \leq \delta,$$

which means $\mathbf{z}^{(k+1)}$ satisfies (4.4). □

Below we derive the number of APG steps performed in Algorithm 2 to find a point satisfying (4.4) under Assumptions 4 and 6. In this case, $\bar{g}^\star(\mathbf{z}_1) = \iota_{\mathbf{C}\mathbf{z}_1 \leq \mathbf{d}}(\mathbf{z}_1)$ because the conjugate of $\iota_{\mathbf{C}\mathbf{z}_1 \leq \mathbf{d}}(\mathbf{z}_1)$ is $\max\{\mathbf{u}^\top \mathbf{z}_1 : \mathbf{C}\mathbf{u} \leq \mathbf{d}\} = \bar{g}(\mathbf{z}_1)$ and the conjugate of the conjugate of a closed convex function is just the function itself. In this case, the minimization problem in (4.3) may not be strongly convex, but it is a quadratic program for which a restarted APG method can still have a linear convergence rate as shown in [26]. Recall that $\Omega^{(k+1)}$ is the solution set in (4.3). Since $\|\cdot\|^2$ is strongly convex, by the same proof of Eqn. (40) in [26], there exists $\boldsymbol{\nu}^{(k)} \in \mathbb{R}^d$ such that

$$[\bar{\mathbf{A}}; \mathbf{A}]^\top \bar{\mathbf{z}}^{(k+1)} + \nabla f_0(\mathbf{x}^{(k)}) - \tau \mathbf{x}^{(k)} = \boldsymbol{\nu}^{(k)},\ \forall\, \bar{\mathbf{z}}^{(k+1)} \in \Omega^{(k+1)}. \tag{4.37}$$

As a result, we have

$$-(\bar{\mathbf{z}}_1^{(k+1)})^\top \bar{\mathbf{b}} - (\bar{\mathbf{z}}_2^{(k+1)})^\top \mathbf{b} = \mathcal{D}_k^* - \frac{1}{2\tau}\|\boldsymbol{\nu}^{(k)}\|^2,\ \forall\, \bar{\mathbf{z}}^{(k+1)} \in \Omega^{(k+1)}. \tag{4.38}$$

Therefore, $\Omega^{(k+1)}$ can be characterized as the solution set of the linear system as follows

$$\Omega^{(k+1)} = \left\{ \mathbf{z} = (\mathbf{z}_1^\top, \mathbf{z}_2^\top)^\top \,\middle|\, \begin{array}{c} [\bar{\mathbf{A}}; \mathbf{A}]^\top \mathbf{z} = \boldsymbol{\nu}^{(k)} - \nabla f_0(\mathbf{x}^{(k)}) + \tau \mathbf{x}^{(k)}, \\ -\mathbf{z}_1^\top \bar{\mathbf{b}} - \mathbf{z}_2^\top \mathbf{b} = \mathcal{D}_k^* - \frac{1}{2\tau}\|\boldsymbol{\nu}^{(k)}\|^2, \\ \mathbf{C}\mathbf{z}_1 \le \mathbf{d} \end{array} \right\}. \tag{4.39}$$

The Hoffman constant is a constant $\theta(\bar{\mathbf{A}}, \mathbf{A}, \bar{\mathbf{b}}, \mathbf{b}, \mathbf{C})$ such that

$$\operatorname{dist}(\mathbf{z}, \Omega^{(k+1)}) \le \theta(\bar{\mathbf{A}}, \mathbf{A}, \bar{\mathbf{b}}, \mathbf{b}, \mathbf{C}) \left\| \begin{bmatrix} [\bar{\mathbf{A}}; \mathbf{A}]^\top \mathbf{z} + \nabla f_0(\mathbf{x}^{(k)}) - \tau \mathbf{x}^{(k)} - \boldsymbol{\nu}^{(k)} \\ -\mathbf{z}_1^\top \bar{\mathbf{b}} - \mathbf{z}_2^\top \mathbf{b} - \mathcal{D}_k^* + \frac{1}{2\tau}\|\boldsymbol{\nu}^{(k)}\|^2 \\ (\mathbf{C}\mathbf{z}_1 - \mathbf{d})_+ \end{bmatrix} \right\|, \forall \mathbf{z}. \tag{4.40}$$

Note that $\theta(\bar{\mathbf{A}}, \mathbf{A}, \bar{\mathbf{b}}, \mathbf{b}, \mathbf{C})$ does not depend on the right-hand sides of the linear system in (4.39). As shown in [26], the linear convergence rate of the APG depends on $\theta(\bar{\mathbf{A}}, \mathbf{A}, \bar{\mathbf{b}}, \mathbf{b}, \mathbf{C})$. In a special case where $\bar{\mathbf{b}} = \mathbf{0}$ and $\mathbf{b} = \mathbf{0}$, we drop the equality $-\mathbf{z}_1^\top \bar{\mathbf{b}} - \mathbf{z}_2^\top \mathbf{b} = \mathcal{D}_k^* - \frac{1}{2\tau}\|\boldsymbol{\nu}^{(k)}\|^2$ from (4.39) and denote the corresponding Hoffman constant by $\theta(\bar{\mathbf{A}}, \mathbf{A}, \mathbf{C})$ which satisfies

$$\operatorname{dist}(\mathbf{z}, \Omega^{(k+1)}) \le \theta(\bar{\mathbf{A}}, \mathbf{A}, \mathbf{C}) \left\| \begin{bmatrix} [\bar{\mathbf{A}}; \mathbf{A}]^\top \mathbf{z} + \nabla f_0(\mathbf{x}^{(k)}) - \tau \mathbf{x}^{(k)} - \boldsymbol{\nu}^{(k)} \\ (\mathbf{C}\mathbf{z}_1 - \mathbf{d})_+ \end{bmatrix} \right\|, \forall \mathbf{z}. \tag{4.41}$$

With these notations, the number of APG steps to obtain $\mathbf{z}^{k+1}$ satisfying (4.4) is characterized as follows.

Theorem 4.4 *Suppose Assumptions 1, 4 and 6 hold. Given any $\delta > 0$, if $j_k = \left\lceil 2\sqrt{2L_{\mathcal{D}}/\rho_k} \right\rceil$ and $i_k = \left\lceil \log_2\!\left(\frac{2(\mathcal{D}_k(\mathbf{z}^{ini}) - \mathcal{D}_k^*)}{\rho_k \delta^2}\right) \right\rceil$ in Algorithm 2, the output $\mathbf{z}^{(k+1)}$ satisfies the criteria in (4.4). Here,*

$$\rho_k = \begin{cases} \left[\theta^2(\bar{\mathbf{A}}, \mathbf{A}, \bar{\mathbf{b}}, \mathbf{b}, \mathbf{C})(\tau + \mathcal{D}_k(\mathbf{z}^{ini}) - \mathcal{D}_k^* + 2\tau\|\boldsymbol{\nu}^{(k)}\|^2)\right]^{-1} & \text{if } \bar{\mathbf{b}} \ne \mathbf{0} \text{ or } \mathbf{b} \ne \mathbf{0}, \\ \left[\theta^2(\bar{\mathbf{A}}, \mathbf{A}, \mathbf{C})\tau\right]^{-1} & \text{if } \bar{\mathbf{b}} = \mathbf{0} \text{ and } \mathbf{b} = \mathbf{0}, \end{cases} \tag{4.42}$$

with $\boldsymbol{\nu}^{(k)}$ defined in (4.37) for $k \ge 0$.

Proof. According to [26, Theorem 10], $\mathcal{D}_k(\mathbf{z})$ has a quadratic growth on the level set

$$\mathcal{Z}_k := \left\{ \mathbf{z} \middle| \mathbf{C}\mathbf{z}_1 \le \mathbf{d},\ \mathcal{D}_k(\mathbf{z}) \le \mathcal{D}_k(\mathbf{z}^{\text{ini}}) \right\}.$$

More specifically, it holds that

$$\frac{\rho_k}{2}\operatorname{dist}^2(\mathbf{z}, \Omega^{(k+1)}) \le \mathcal{D}_k(\mathbf{z}) - \mathcal{D}_k^*, \quad \forall\, \mathbf{z} \in \mathcal{Z}_k, \tag{4.43}$$

where ρ_k is defined in (4.42).

Consider the end of the first inner loop (i.e., $i = 0$) of Algorithm 2. According to [1, Theorem 4.4], after j_k APG steps, the APG method generates an output $\mathbf{z}^{(j_k,k)}$ satisfying

$$\mathcal{D}_k(\mathbf{z}^{(j_k,k)}) - \mathcal{D}_k^* \leq \frac{2L_{\mathcal{D}}\text{dist}^2(\mathbf{z}^{\text{ini}}, \Omega^{(k+1)})}{j_k^2} \overset{(4.43)}{\leq} \frac{4L_{\mathcal{D}}[\mathcal{D}_k(\mathbf{z}^{\text{ini}}) - \mathcal{D}_k^*]}{\rho_k j_k^2}.$$

Since $j_k = \left\lceil 2\sqrt{2L_{\mathcal{D}}/\rho_k}\right\rceil$, we have $\mathcal{D}_k(\mathbf{z}^{(j_k,k)}) - \mathcal{D}_k^* \leq \frac{1}{2}\left(\mathcal{D}_k(\mathbf{z}^{\text{ini}}) - \mathcal{D}_k^*\right)$, meaning that we can reduce the objective gap by a half with each inner loop of Algorithm 2. Since the next inner loop is started at $\mathbf{z}^{(j_k,k)}$ according to Line 11 of Algorithm 2, we repeat this argument $i_k = \left\lceil \log_2(\frac{2(\mathcal{D}_k(\mathbf{z}^{\text{ini}})-\mathcal{D}_k^*)}{\rho_k\delta^2})\right\rceil$ times and show that the final output $\mathbf{z}^{(k+1)}$ satisfies $\mathcal{D}_k(\mathbf{z}^{(k+1)}) - \mathcal{D}_k^* \leq \frac{\rho_k\delta^2}{2}$ and thus

$$\text{dist}(\mathbf{z}^{(k+1)}, \Omega^{(k+1)}) \leq \sqrt{\frac{2(\mathcal{D}_k(\mathbf{z}^{(k+1)}) - \mathcal{D}_k^*)}{\rho_k}} \leq \delta,$$

which means $\mathbf{z}^{(k+1)}$ satisfies (4.4). The proof is then completed. □

4.4 Total Number of Inner Iterations for Finding an ϵ-stationary Point

Combining Theorems 4.1, 4.3 and 4.4, we derive the total number of inner iterations for computing an ϵ-stationary point of (SP) as follows.

Corollary 4.1 *Suppose Assumptions 1 and 4 hold and Algorithm 1 uses Algorithm 2 at Step 4. Let $\tau = 2L_f$ and $\delta = \delta_\epsilon$ in Algorithm 1 with δ_ϵ defined in (4.16). Then the following statements hold.*

(a) *If Assumption 5 holds, Algorithm 1 finds an ϵ-stationary point of problem (SP) with total number*

$$O\left(\kappa([\bar{\mathbf{A}};\mathbf{A}])\log\left(\tfrac{\Delta_F}{\epsilon}\right)\frac{L_f\Delta_F}{\epsilon^2}\right) \tag{4.44}$$

of inner iterations (i.e, APG steps), where $\Delta_F = F(\mathbf{x}^{(0)}, \mathbf{y}^{(0)}) - \inf_{\mathbf{x},\mathbf{y}} F(\mathbf{x}, \mathbf{y})$.

(b) *If Assumption 6 holds and, in addition, $\bar{\mathbf{b}} = \mathbf{0}, \mathbf{b} = \mathbf{0}$, the same conclusion in statement (a) holds except that the total number of inner iterations becomes*

$$O\left(\sqrt{\lambda_{\max}([\bar{\mathbf{A}};\mathbf{A}][\bar{\mathbf{A}};\mathbf{A}]^\top)\theta^2(\bar{\mathbf{A}},\mathbf{A},\mathbf{C})}\log\left(\tfrac{\Delta_F}{\epsilon}\right)\frac{L_f\Delta_F}{\epsilon^2}\right), \tag{4.45}$$

where $\theta(\bar{\mathbf{A}},\mathbf{A},\mathbf{C})$ is the Hoffman constant given in (4.41).

(c) *If Assumption 6 holds and, in addition, the sequence $\{\mathbf{x}^{(k)}\}_{k\geq 0}$ is bounded, the same conclusion in statement (a) holds except that the total number of inner iterations becomes*

$$O\left(\sqrt{\lambda_{\max}([\bar{\mathbf{A}};\mathbf{A}][\bar{\mathbf{A}};\mathbf{A}]^\top)\theta^2(\bar{\mathbf{A}},\mathbf{A},\bar{\mathbf{b}},\mathbf{b},\mathbf{C})(1+B_6/\tau+2B_5)}\log\left(\tfrac{B_6}{\epsilon}\right)\frac{L_f\Delta_F}{\epsilon^2}\right), \tag{4.46}$$

where $\theta(\bar{\mathbf{A}}, \mathbf{A}, \bar{\mathbf{b}}, \mathbf{b}, \mathbf{C})$ *is the Hoffman constant given in* (4.40), B_5 *and* B_6 *are constants*[7] *independent of* ϵ *such that*

$$\|\boldsymbol{\nu}^{(k)}\|^2 \leq B_5, \quad \mathcal{D}_k(\mathbf{z}^{ini}) - \mathcal{D}_k^* \leq B_6, \quad \forall\, k \geq 0, \tag{4.47}$$

with $\boldsymbol{\nu}^{(k)}$ *defined in* (4.37).

Proof. According to Theorem 4.1, Algorithm 1 needs K_ϵ outer iterations to find an ϵ-stationary point of problem (SP), where K_ϵ is given in (4.17). Moreover, in each outer iteration of Algorithm 1, Algorithm 2 needs $i_k j_k$ APG steps to find a point $\mathbf{z}^{(k+1)}$ satisfying (4.4) with i_k and j_k specified in Theorems 4.3 and 4.4 for different assumptions. Hence, the total number of inner iterations by Algorithm 1 is $\sum_{k=0}^{K_\epsilon-1} i_k j_k$.

Below we prove statement (c) first. In this case, i_k and j_k are given in Theorem 4.4 and they depend on ρ_k in (4.42) which further depends on $\|\boldsymbol{\nu}^{(k)}\|$ and $\mathcal{D}_k(\mathbf{z}^{\text{ini}}) - \mathcal{D}_k^*$. Using (4.47) and the fact that $\log\frac{1}{\delta} = O(\log\frac{1}{\epsilon})$. We can bound i_k and j_k from above and obtain (4.46).

To prove statements (a) and (b), we only need to derive an upper bound for $\mathcal{D}_k(\mathbf{z}^{\text{ini}}) - \mathcal{D}_k^*$ that appears in the formula of i_k in Theorems 4.3 and 4.4. First, we observe that, for any $k \leq K_\epsilon$, it holds that

$$\begin{aligned}\|\mathbf{x}^{(k)}\| \leq & \|\mathbf{x}^{(0)}\| + \sum_{s=0}^{k-1} \|\mathbf{x}^{(s)} - \mathbf{x}^{(s+1)}\| \leq \|\mathbf{x}^{(0)}\| + \sum_{s=0}^{K_\epsilon-1} \|\mathbf{x}^{(s)} - \mathbf{x}^{(s+1)}\| \\ \leq & \|\mathbf{x}^{(0)}\| + \sqrt{K_\epsilon} \cdot \sqrt{\sum_{s=0}^{K_\epsilon-1} \|\mathbf{x}^{(s)} - \mathbf{x}^{(s+1)}\|^2} \leq \|\mathbf{x}^{(0)}\| + \sqrt{K_\epsilon} \cdot \sqrt{\frac{K_\epsilon \cdot \epsilon^2}{9L_f^2}} \leq \|\mathbf{x}^{(0)}\| + \frac{4\Delta_F}{\epsilon},\end{aligned} \tag{4.48}$$

where the third inequality is by the Cauchy–Schwarz inequality, the fourth one is by (4.30), and the last one follows from (4.17). Recall that $\|\bar{\mathbf{z}}_1^{(k+1)}\| \leq l_g$ and thus from (4.15) and (4.48), we have

$$\|\bar{\mathbf{z}}_2^{(k+1)}\| \leq l_g + \left(l_f + \tau\|\mathbf{x}^{(0)}\| + \frac{4\tau\Delta_F}{\epsilon}\right) \left\|\left(\mathbf{A}\mathbf{A}^\top\right)^{-1}\mathbf{A}\right\| + \tau\left\|\left(\mathbf{A}\mathbf{A}^\top\right)^{-1}\mathbf{b}\right\|, \tag{4.49}$$

where we have used the l_f-Lipschitz continuity of f_0 from Assumption 4.

Moreover, for any $\mathbf{z}$, we have

$$\begin{aligned}\|\nabla(\mathcal{D}_k - \bar{g}^\star)(\mathbf{z})\| = & \left\|\frac{1}{\tau}\left([\bar{\mathbf{A}}; \mathbf{A}]\big([\bar{\mathbf{A}}; \mathbf{A}]^\top \mathbf{z} + \nabla f_0(\mathbf{x}^{(k)}) - \tau\mathbf{x}^{(k)}\big)\right) - [\bar{\mathbf{b}}; \mathbf{b}]\right\| \\ \leq & \frac{1}{\tau}\|[\bar{\mathbf{A}}; \mathbf{A}]\|^2 \cdot \|\mathbf{z}\| + \frac{1}{\tau}\|[\bar{\mathbf{A}}; \mathbf{A}]\| \left(l_f + \tau\|\mathbf{x}^{(0)}\| + \frac{4\tau\Delta_F}{\epsilon}\right) + \|[\bar{\mathbf{b}}; \mathbf{b}]\| =: \widehat{\mathcal{D}}(\mathbf{z}).\end{aligned} \tag{4.50}$$

Hence, it follows from the convexity of $\mathcal{D}_k$ that for any $\boldsymbol{\xi} \in \partial\bar{g}^\star(\mathbf{z}^{\text{ini}})$,

$$\mathcal{D}_k(\mathbf{z}^{\text{ini}}) - \mathcal{D}_k^* \leq \left\langle \nabla(\mathcal{D}_k - \bar{g}^\star)(\mathbf{z}^{\text{ini}}) + \boldsymbol{\xi}, \mathbf{z}^{\text{ini}} - \bar{\mathbf{z}}^{(k+1)} \right\rangle \leq \left(\widehat{\mathcal{D}}(\mathbf{z}^{\text{ini}}) + \|\boldsymbol{\xi}\|\right) \left(\|\mathbf{z}^{\text{ini}}\| + l_g + \|\bar{\mathbf{z}}_2^{(k+1)}\|\right).$$

Keeping only the key quantities ϵ and Δ_F, we then have $\mathcal{D}_k(\mathbf{z}^{\text{ini}}) - \mathcal{D}_k^* = O\left(\frac{\Delta_F}{\epsilon}\right)$. Applying this bound of $\mathcal{D}_k(\mathbf{z}^{\text{ini}}) - \mathcal{D}_k^*$ to the i_k's in Theorems 4.3 and 4.4, we obtain (4.44) and (4.45). □

In the following corollary, we also derive from Theorems 4.2, 4.3 and 4.4 the required total number of inner iterations for obtaining a near ϵ-stationary point of (P). We skip its proof because it is essentially the same as that of Corollary 4.1.

[7] Such B_5 and B_6 exist because of the boundedness of $\{\mathbf{x}^{(k)}\}_{k\geq 0}$.

Corollary 4.2 *Suppose Assumptions 1 and 4 hold and Algorithm 1 uses Algorithm 2 at Step 4. Let $\tau = 2L_f$ and $\delta = \bar{\delta}_\epsilon$ in Algorithm 1 with $\bar{\delta}_\epsilon$ defined in Theorem 4.2. Then the following statements hold.*

(a) *If Assumption 5 holds, Algorithm 1 finds a point that is $\frac{\epsilon}{12L_f}$-close to an ϵ-stationary point of problem* (P) *with total number*

$$O\left(\kappa([\bar{\mathbf{A}};\mathbf{A}])\log\left(\tfrac{\Delta_{F_0}}{\epsilon}\right)\frac{L_f\Delta_{F_0}}{\epsilon^2}\right) \tag{4.51}$$

of inner iterations (i.e, APG steps), where $\Delta_{F_0} = F_0(\mathbf{x}^{(0)}) - \inf_{\mathbf{x}} F_0(\mathbf{x})$.

(b) *If Assumption 6 holds and, in addition, $\bar{\mathbf{b}} = \mathbf{0}, \mathbf{b} = \mathbf{0}$, the same conclusion in statement* (a) *holds except that the total number of inner iterations becomes*

$$O\left(\sqrt{\lambda_{\max}([\bar{\mathbf{A}};\mathbf{A}][\bar{\mathbf{A}};\mathbf{A}]^\top)\theta^2(\bar{\mathbf{A}},\mathbf{A},\mathbf{C})}\log\left(\tfrac{\Delta_{F_0}}{\epsilon}\right)\frac{L_f\Delta_{F_0}}{\epsilon^2}\right), \tag{4.52}$$

where $\theta(\bar{\mathbf{A}},\mathbf{A},\mathbf{C})$ is the Hoffman constant given in (4.41).

(c) *If Assumption 6 holds and, in addition, the sequence $\{\mathbf{x}^{(k)}\}_{k\geq 0}$ is bounded, the same conclusion in statement* (a) *holds except that the total number of inner iterations becomes*

$$O\left(\sqrt{\lambda_{\max}([\bar{\mathbf{A}};\mathbf{A}][\bar{\mathbf{A}};\mathbf{A}]^\top)\theta^2(\bar{\mathbf{A}},\mathbf{A},\bar{\mathbf{b}},\mathbf{b},\mathbf{C})(1+B_6/\tau+2B_5)}\log\left(\tfrac{B_6}{\epsilon}\right)\frac{L_f\Delta_{F_0}}{\epsilon^2}\right), \tag{4.53}$$

where $\theta(\bar{\mathbf{A}},\mathbf{A},\bar{\mathbf{b}},\mathbf{b},\mathbf{C})$ is the Hoffman constant given in (4.40)*, B_5 and B_6 are given in Corollary 4.1.*

4.5 Oracle Complexity Matching the Lower Bounds

Suppose Algorithm 1 uses Algorithm 2 at Step 4 to find a point $\mathbf{z}^{(k+1)}$ satisfying (4.4). We can show that Algorithm 1 satisfies Assumption 3, i.e., the iterate points of Algorithm 1 can be generated by only querying the oracles and performing the operations given in Assumption 3. Then, we show that the oracle complexity of Algorithm 1 matches the lower complexity bound presented in Section 3. To do so, we first present the following observations.

Lemma 4.3 *Let $\mathbf{z}^{(j,k)}$ be the solution generated in the j-th inner iteration of the k-th outer iteration of Algorithm 2. Let $\mathbf{x}^{(j,k)} = \mathbf{A}^\top\mathbf{z}_2^{(j,k)}$. It holds that, for all $j \geq 1$ and $k \geq 0$,*

$$\mathbf{x}^{(j,k)} \in \mathbf{A}^\top\mathbf{A}\cdot\mathbf{span}\left(\left\{\mathbf{x}^{(k)},\nabla f_0(\mathbf{x}^{(k)}),\mathbf{x}^{(0)},\bar{\mathbf{A}}^\top\bar{\mathbf{b}}\right\}\bigcup\cup_{i=0}^{j-1}\left\{\mathbf{x}^{(i,k)},\bar{\mathbf{A}}^\top\mathbf{z}_1^{(i,k)}\right\}\right) \tag{4.54a}$$

$$\mathbf{z}_1^{(j,k)} \in \mathbf{span}\left(\left\{\boldsymbol{\zeta}^{(j,k)},\boldsymbol{\xi}^{(j,k)}\right\}\right), \tag{4.54b}$$

where $\boldsymbol{\zeta}^{(j,k)} \in \left\{\mathbf{prox}_{\eta\bar{g}}(\boldsymbol{\xi}^{(j,k)}) \mid \eta > 0\right\}$ and

$$\boldsymbol{\xi}^{(j,k)} \in \mathbf{span}\left(\left\{\bar{\mathbf{A}}\mathbf{x}^{(k)},\bar{\mathbf{A}}\nabla f_0(\mathbf{x}^{(k)}),\bar{\mathbf{b}}\right\}\bigcup\cup_{i=0}^{j-1}\left\{\mathbf{z}_1^{(i,k)},\bar{\mathbf{A}}\bar{\mathbf{A}}^\top\mathbf{z}_1^{(i,k)},\bar{\mathbf{A}}\mathbf{x}^{(i,k)}\right\}\right). \tag{4.55}$$

Proof. Consider any $j \geq 0$ and $k \geq 0$. By Steps 2 and 9 in Algorithm 2, we have $\widehat{\mathbf{z}}^{(0,k)} = \mathbf{z}^{(0,k)}$ and $\widehat{\mathbf{z}}^{(j+1,k)} = \mathbf{z}^{(j+1,k)} + \left(\frac{\alpha_j - 1}{\alpha_{j+1}}\right)\left(\mathbf{z}^{(j+1,k)} - \mathbf{z}^{(j,k)}\right)$ for $j \geq 0$. This further implies $\mathbf{A}^\top \widehat{\mathbf{z}}_2^{(0,k)} = \mathbf{x}^{(0,k)}$ and $\mathbf{A}^\top \widehat{\mathbf{z}}_2^{(j+1,k)} = \mathbf{x}^{(j+1,k)} + \left(\frac{\alpha_j - 1}{\alpha_{j+1}}\right)\left(\mathbf{x}^{(j+1,k)} - \mathbf{x}^{(j,k)}\right)$ for $j \geq 0$.

Hence, by Steps 5 and 7 in Algorithm 2,

$$\widehat{\mathbf{z}}_1^{(j-1,k)} - \frac{1}{L_{\mathcal{D}}}\widehat{\mathcal{G}}_1^{j-1} \in \mathbf{span}\left(\left\{\bar{\mathbf{A}}\mathbf{x}^{(k)}, \bar{\mathbf{A}}\nabla f_0(\mathbf{x}^{(k)}), \bar{\mathbf{b}}\right\} \bigcup \cup_{i=0}^{j-1}\left\{\mathbf{z}_1^{(i,k)}, \bar{\mathbf{A}}\bar{\mathbf{A}}^\top \mathbf{z}_1^{(i,k)}, \bar{\mathbf{A}}\mathbf{x}^{(i,k)}\right\}\right)$$

and

$$\mathbf{z}_1^{(j,k)} = \mathbf{prox}_{L_{\mathcal{D}}^{-1}\bar{g}^*}\left(\widehat{\mathbf{z}}_1^{(j-1,k)} - L_{\mathcal{D}}^{-1}\widehat{\mathcal{G}}_1^{j-1}\right) = \widehat{\mathbf{z}}_1^{(j-1,k)} - L_{\mathcal{D}}^{-1}\widehat{\mathcal{G}}_1^{j-1} - L_{\mathcal{D}}^{-1}\mathbf{prox}_{L_{\mathcal{D}}\bar{g}}\left(L_{\mathcal{D}}\widehat{\mathbf{z}}_1^{(j-1,k)} - \widehat{\mathcal{G}}_1^{j-1}\right)$$

which means (4.54b) holds. Recall that $\mathbf{b} = -\mathbf{A}\mathbf{x}^{(0)}$. By Steps 6 and 7 in Algorithm 2,

$$\mathbf{A}^\top \widehat{\mathcal{G}}_2^{j-1} \in \mathbf{A}^\top \mathbf{A} \cdot \mathbf{span}\left(\left\{\mathbf{x}^{(k)}, \nabla f_0(\mathbf{x}^{(k)}), \mathbf{x}^{(0)}, \bar{\mathbf{A}}^\top \bar{\mathbf{b}}\right\} \bigcup \cup_{i=0}^{j-1}\left\{\mathbf{x}^{(i,k)}, \bar{\mathbf{A}}^\top \mathbf{z}_1^{(i,k)}\right\}\right)$$

and

$$\begin{aligned}\mathbf{x}_2^{(j,k)} =& \mathbf{A}^\top \mathbf{z}_2^{(j,k)} = \mathbf{A}^\top \widehat{\mathbf{z}}_2^{(j-1,k)} - L_{\mathcal{D}}^{-1}\mathbf{A}^\top \widehat{\mathcal{G}}_2^{j-1} \\ \in& \mathbf{A}^\top \mathbf{A} \cdot \mathbf{span}\left(\left\{\mathbf{x}^{(k)}, \nabla f_0(\mathbf{x}^{(k)}), \mathbf{x}^{(0)}, \bar{\mathbf{A}}^\top \bar{\mathbf{b}}\right\} \bigcup \cup_{i=0}^{j-1}\left\{\mathbf{x}^{(i,k)}, \bar{\mathbf{A}}^\top \mathbf{z}_1^{(i,k)}\right\}\right)\end{aligned}$$

which means (4.54a) holds. □

With these observations, we can formally show that Algorithm 1 follows Assumption 3.

Proposition 4.1 *The updating schemes in Algorithm 1 using Algorithm 2 at Step 4 satisfy Assumption 3.*

Proof. We first claim that Algorithms 1 and 2 can be both implemented by computing and updating $\mathbf{z}_1^{(j,k)}$ and $\mathbf{x}^{(j,k)} = \mathbf{A}^\top \mathbf{z}_2^{(j,k)}$ without explicitly computing $\mathbf{z}_2^{(j,k)}$. For Algorithm 2, this claim can be verified by its updating steps. For Algorithms 1, updating steps (4.5) and (4.6) can be also implemented directly based on $\mathbf{z}_1^{(j,k)}$ and $\mathbf{x}^{(j_k,k)}$ without knowing $\mathbf{z}_2^{(k+1)}$ because $\mathbf{A}^\top \mathbf{z}_2^{(k+1)} = \mathbf{A}^\top \mathbf{z}_2^{(j_k,k)} = \mathbf{x}^{(j_k,k)}$.

Comparing (4.54) and (4.55) with the updating schemes allowed by Assumption 3, we can prove that Algorithm 1 satisfies Assumption 3 by proving that $\mathbf{x}^{(0,k)} = \mathbf{A}^\top \mathbf{z}_2^{(0,k)}$ and $\mathbf{x}^{(k)}$ for any k are generated only by the updating schemes allowed by Assumption 3. We will prove this statement by induction on k.

Suppose $k = 0$. Consider $\mathbf{x}^{(0,k)} = \mathbf{A}^\top \mathbf{z}_2^{(0,k)}$ in the first inner loop (i.e., $i = 0$) of Algorithm 2. Since $\mathbf{x}^{(0,k)} = \mathbf{A}^\top \mathbf{z}_2^{(0,k)} = \mathbf{A}^\top \mathbf{A}\mathbf{x}^{(0)}$, $\mathbf{x}^{(0,k)}$ is generated by the schemes allowed by Assumption 3. According to (4.54) and (4.55) and the fact that $k = 0$, we conclude that all iterates during the first inner loop of Algorithm 2 are generated only by the updating schemes allowed by Assumption 3. Consider the second inner loop (i.e., $i = 1$) of Algorithm 2. Recall that the initial solution of this inner loop is created as $\mathbf{x}^{(0,k)} = \mathbf{A}^\top \mathbf{z}_2^{(0,k)} = \mathbf{A}^\top \mathbf{z}_2^{(j_k,k)} = \mathbf{x}^{(j_k,k)}$, where $\mathbf{x}^{(j_k,k)}$ is the outputs of the first inner loop of Algorithm 2. We then also conclude that all iterates during the second inner loop of Algorithm 2 are generated only by the schemes allowed by Assumption 3. Specifically, all iterates during the second inner loop of Algorithm 2 can be generated by accessing two oracles given in Assumption 3. Repeating this argument for all inner loops of Algorithm 2, we prove the statement for $k = 0$.

Suppose the statement is true for $k \geq 0$. Then $\mathbf{z}_1^{(k+1)}$ and $\mathbf{A}^\top \mathbf{z}_2^{(k+1)}$ are generated by the updating schemes allowed by Assumption 3, so are $\mathbf{x}^{(k+1)}$ and $\mathbf{y}^{(k+1)}$ according to (4.5) and (4.6). By the same argument as when $k = 0$, we can prove that the statement is true for $k + 1$. The proof is completed by induction. □

By Proposition 4.1 and its proof, we have that the oracle complexity of Algorithm 1 equals $O(1)$ times of the total number of inner iterations (i.e., APG steps) of Algorithm 1. We then finish the main body of this paper with several remarks to point out that the lower bound of oracle complexity matches, up to a difference of logarithmic factors, the oracle complexity in Corollaries 4.1 and 4.2 under two cases: (i) Assumptions 4 and 5 hold; (ii) Assumptions 4 and 6 hold and in addition, $\bar{\mathbf{b}} = \mathbf{0}$ and $\mathbf{b} = \mathbf{0}$.

Remark 4.2 A few remarks follow on the oracle complexity of FOMs for solving problem (SP).

(i) By Proposition 4.1, the oracle complexity of Algorithm 1 using Algorithm 2 at Step 4 is subject to the lower bound established in Theorem 3.1 for finding an ϵ-stationary point of (SP). We then assert that this lower bound and the oracle complexity[8] in (4.44) are both not improvable (up to a logarithmic factor) under Assumptions 4 and 5 by comparing their dependency on ϵ, L_f and Δ_F and $\kappa([\bar{\mathbf{A}}; \mathbf{A}])$.

(ii) Suppose Algorithm 1 is applied to the reformulation (SP) of instance $\mathcal{P}$ given in Definition 2. By Lemma 2.2, instance $\mathcal{P}$ satisfies Assumptions 4 and 6. In particular, for instance $\mathcal{P}$, it holds that $\bar{g}(\mathbf{y}) = \frac{\beta}{mL_f}\|\mathbf{y}\|_1 = \max_{\|\mathbf{z}_1\|_\infty \leq \frac{\beta}{mL_f}} \mathbf{z}_1^\top \mathbf{y}$ and $\bar{g}^\star(\mathbf{z}_1) = \iota_{\|\mathbf{z}_1\|_\infty \leq \frac{\beta}{mL_f}}$. Moreover, since $\bar{\mathbf{b}} = \mathbf{0}$ and $\mathbf{b} = \mathbf{0}$ in instance $\mathcal{P}$, Algorithm 1 finds an ϵ-stationary point of (SP) with oracle complexity given in (4.45). In this case, $\theta(\bar{\mathbf{A}}, \mathbf{A}, \mathbf{C})$ becomes the Hoffman constant of the linear system

$$\left\{ \mathbf{z} = (\mathbf{z}_1^\top, \mathbf{z}_2^\top)^\top \;\middle|\; \begin{array}{c} [\bar{\mathbf{A}}; \mathbf{A}]^\top \mathbf{z} = \boldsymbol{\nu}^{(k)} - \nabla f_0(\mathbf{x}^{(k)}) + \tau \mathbf{x}^{(k)}, \\ -\frac{\beta}{mL_f} \leq [\mathbf{z}_1]_i \leq \frac{\beta}{mL_f}, \forall i \end{array} \right\},$$

for some $\boldsymbol{\nu}^{(k)}$. According to Proposition 6 and the discussion on page 12 of [32], we can define a reference polyhedron

$$\mathcal{R} = \left\{ \mathbf{z} = [\mathbf{z}_1; \mathbf{z}_2] : -\frac{\beta}{mL_f} \leq [\mathbf{z}_1]_i \leq \frac{\beta}{mL_f}, \forall i \right\}$$

and characterize $\theta(\bar{\mathbf{A}}, \mathbf{A}, \mathbf{C})$ as

$$\frac{1}{\theta(\bar{\mathbf{A}}, \mathbf{A}, \mathbf{C})} = \min_{\mathcal{K} \in \mathcal{S}} \min_{\mathbf{u}, \mathbf{v}} \left\{ \|\mathbf{u}\| \;\middle|\; \|\mathbf{H}^\top \mathbf{v}\| = 1, \;\mathbf{v} \in \mathcal{K}, \;\mathbf{H}\mathbf{H}^\top \mathbf{v} - \mathbf{u} \in \mathcal{K}^* \right\},$$

where $\mathbf{H}$ is defined in (2.2), $\mathcal{K}^*$ is the dual cone of $\mathcal{K}$, and

$$\mathcal{S} = \{\mathcal{K} : \mathcal{K} \text{ is a tangent cone of } \mathcal{R} \text{ at some point of } \mathcal{R}, \text{ and } \mathbf{H}^\top \mathcal{K} \text{ is a linear space}\}.$$

Next we show $\mathcal{S} = \{\mathbb{R}^d\}$. By the definition of $\mathcal{R}$, any tangent cone of $\mathcal{R}$ must look like

$$\mathcal{K} = \left\{ \mathbf{z} = [\mathbf{z}_1; \mathbf{z}_2] \;\middle|\; [\mathbf{z}_1]_i \geq 0, \forall i \in \mathcal{J}_1, [\mathbf{z}_1]_j \leq 0, \forall j \in \mathcal{J}_2 \right\}$$

for some disjoint index sets $\mathcal{J}_1$ and $\mathcal{J}_2$. Suppose $\mathcal{K} \in \mathcal{S}$. Then, for any $\mathbf{v} \in \mathcal{K}$, we must have $-\mathbf{H}^\top \mathbf{v} \in \mathbf{H}^\top \mathcal{K}$, namely, there exists $\mathbf{v}' \in \mathcal{K}$ such that $-\mathbf{H}^\top \mathbf{v} = \mathbf{H}^\top \mathbf{v}'$. Since $\mathbf{H}^\top$ has a full-column rank, $\mathbf{v} + \mathbf{v}' = \mathbf{0}$. This means for any $\mathbf{v} \in \mathcal{K}$, we must have $-\mathbf{v} \in \mathcal{K}$. Since $\mathcal{J}_1 \cap \mathcal{J}_2 = \emptyset$, this happens only if $\mathcal{J}_1 = \mathcal{J}_2 = \emptyset$. Thus $\mathcal{K} = \mathbb{R}^d$ and $\mathcal{K}^* = \{\mathbf{0}\}$.

[8] The dependency of oracle complexity in (4.44) on l_f and l_g is only logarithmic and has been suppressed in $O(\cdot)$.

With this fact, we have

$$\theta(\bar{\mathbf{A}}, \mathbf{A}, \mathbf{C}) = \frac{1}{\min\limits_{\mathbf{v}:\|\mathbf{H}^\top \mathbf{v}\|=1} \|\mathbf{H}\mathbf{H}^\top \mathbf{v}\|} = \frac{1}{\sqrt{\lambda_{\min}(\mathbf{H}\mathbf{H}^\top)}}.$$

Hence by Corollary 4.1(b), the oracle complexity of Algorithm 1 becomes

$$O\left(\sqrt{\frac{\lambda_{\max}(\mathbf{H}\mathbf{H}^\top)}{\lambda_{\min}(\mathbf{H}\mathbf{H}^\top)}} \log(\tfrac{1}{\epsilon}) \frac{L_f \Delta_F}{\epsilon^2}\right) = O\left(\kappa([\bar{\mathbf{A}}; \mathbf{A}]) \log(\tfrac{1}{\epsilon}) \frac{L_f \Delta_F}{\epsilon^2}\right).$$

Again, we conclude that the lower bound given in Theorem 3.1 and the oracle complexity above are neither improvable (up to a logarithmic factor) under Assumptions 4 and 6 when $\bar{\mathbf{b}} = \mathbf{0}$ and $\mathbf{b} = \mathbf{0}$.

Remark 4.3 A few remarks follow on the oracle complexity of FOMs for solving problem (P).

(i) By Proposition 4.1 again, the oracle complexity of Algorithm 1 using Algorithm 2 at Step 4 is subject to the lower bound established in Corollary 3.1 for finding a point $\mathbf{x}^{(k)}$ that is ω-close to an ϵ-stationary point of problem (P) for some $\omega \in [0, \frac{150\pi\epsilon}{L_f})$. We then assert that this lower bound and the oracle complexity in (4.51) (with $\omega = \frac{\epsilon}{12L_f}$) are both not improvable, up to a logarithmic factor, under Assumptions 4 and 5 by comparing their dependency on ϵ, L_f and Δ_{F_0} and $\kappa([\bar{\mathbf{A}}; \mathbf{A}])$.

(ii) Suppose Algorithm 1 is applied to instance $\mathcal{P}$ given in Definition 2. We obtain from Remark 4.2(ii) that if $\bar{\mathbf{b}} = \mathbf{0}$ and $\mathbf{b} = \mathbf{0}$, then Algorithm 1 finds a near ϵ-stationary point of (P) with oracle complexity (4.52), which equals $O\left(\kappa([\bar{\mathbf{A}}; \mathbf{A}]) \log(\frac{1}{\epsilon}) \frac{L_f \Delta_{F_0}}{\epsilon^2}\right)$ under Assumptions 4 and 6. Hence, the lower bound given in Corollary 3.1 and the oracle complexity above are neither improvable (up to a logarithmic factor) under Assumptions 4 and 6 when $\bar{\mathbf{b}} = \mathbf{0}$ and $\mathbf{b} = \mathbf{0}$.

5 Conclusion and Open Questions

We present a lower bound for the oracle complexity of first-order methods for finding a (near) ϵ-stationary point of a non-convex composite non-smooth optimization problem with affine equality constraints. We also show that the same-order lower bound holds for first-order methods applied to a reformulation of the original problem. In addition, we design an inexact proximal gradient method for the reformulation. To find an ϵ-stationary point (that is also a near ϵ-stationary point of the original problem), our designed method has an oracle complexity that matches the established lower bounds under two scenarios, up to a difference of a logarithmic factor. This shows that both the lower bound and the oracle complexity of the inexact proximal gradient method are nearly not improvable.

Though we make the first attempt on establishing lower complexity bounds of first-order methods for solving affine-constrained non-convex composite non-smooth problems, there are still a few open questions that are worth further exploration. First, under Assumption 2, can the lower bound $O(\kappa([\bar{\mathbf{A}}; \mathbf{A}]) L_f \Delta_{F_0} \epsilon^{-2})$ for problem (P) be achieved? The second question is whether the lower bound $O(\kappa([\bar{\mathbf{A}}; \mathbf{A}]) L_f \Delta_F \epsilon^{-2})$ for problem (SP) can be achieved by an algorithm satisfying Assumption 3 without Assumptions 4, 5 or 6. Third, what will the lower bound look like if there are convex nonlinear inequality constraints?

Acknowledgements

This work is partly supported by NSF awards DMS-2053493 and IIS-2147253 and the ONR award N00014-22-1-2573.

References

1. 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. 29, 30, 32
2. W. Bian and X. Chen. Linearly constrained non-Lipschitz optimization for image restoration. *SIAM Journal on Imaging Sciences*, 8(4):2294–2322, 2015. 3
3. S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004. 23
4. Y. Carmon, J. C. Duchi, O Hinder, and A. Sidford. Lower bounds for finding stationary points I. *Mathematical Programming*, 184(1-2):71–120, 2020. 3, 4, 5, 7
5. Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. Lower bounds for finding stationary points II: first-order methods. *Mathematical Programming*, 185(1-2):315–355, 2021. 3, 5
6. C. Cartis, N. Gould, and P. Toint. On the complexity of finding first-order critical points in constrained nonlinear optimization. *Mathematical Programming*, 144(1-2):93–106, 2014. 3
7. C. Cartis and N. Gouldand P. Toint. Corrigendum: on the complexity of finding first-order critical points in constrained nonlinear optimization. *Mathematical Programming*, 161:611–626, 2017. 3
8. F. H. Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990. 11, 29
9. 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. *Preprint, arXiv:1702.01850*, 2017. 5, 21
10. G. N. Grapiglia and Y. Yuan. On the complexity of an augmented Lagrangian method for nonconvex optimization. *IMA Journal of Numerical Analysis*, 41(2):1546–1568, 2021. 3
11. R. M. Gray. Toeplitz and circulant matrices: a review. *Foundations and Trends® in Communications and Information Theory*, 2(3):155–239, 2006. 10
12. G. Haeser, H. Liu, and Y. Ye. Optimality condition and complexity analysis for linearly-constrained optimization without differentiability on the boundary. *Mathematical Programming*, 178:263–299, 2019. 3
13. D. Hajinezhad and M. Hong. Perturbed proximal primal–dual algorithm for nonconvex nonsmooth optimization. *Mathematical Programming*, 176(1-2):207–245, 2019. 5
14. M. Hong. Decomposing linearly constrained nonconvex problems by a proximal primal dual approach: algorithms, convergence, and applications. *Preprint, arXiv:1604.00543*, 2016. 5
15. M. Hong, Z. Luo, and M. Razaviyayn. Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. *SIAM Journal on Optimization*, 26(1):337–364, 2016. 5
16. B. Jiang, T. Lin, S. Ma, and S. Zhang. Structured nonconvex and nonsmooth optimization: algorithms and iteration complexity analysis. *Computational Optimization and Applications*, 72(1):115–157, 2019. 3, 5
17. 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 Journal on Optimization*, 29(4):2566–2593, 2019. 3, 5
18. 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. *Mathematics of Operations Research*, 2022. 3
19. H. Li, Y. Tian, J. Zhang, and A. Jadbabaie. Complexity lower bounds for nonconvex-strongly-concave min-max optimization. *Advances in Neural Information Processing Systems*, 34:1792–1804, 2021. 3
20. Q. Lin, R. Ma, and Y. Xu. Complexity of an inexact proximal-point penalty method for constrained smooth non-convex optimization. *Computational Optimization and Applications*, 82(1):175–224, 2022. 3, 5
21. W. Liu, X. Liu, and X. Chen. An inexact augmented Lagrangian algorithm for training leaky ReLU neural network with group sparsity. *Preprint, arXiv:2205.05428*, 2022. 10
22. W. Liu, X. Liu, and X. Chen. Linearly constrained nonsmooth optimization for training autoencoders. *SIAM Journal on Optimization*, 32(3):1931–1957, 2022. 10
23. J. G. Melo and R. D. C. Monteiro. Iteration-complexity of a Jacobi-type non-Euclidean ADMM for multi-block linearly constrained nonconvex programs. *Preprint, arXiv:1705.07229*, 2017. 5

24. 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. *Preprint, arXiv:2006.08048*, 2020. 3, 5
25. K. G. Murty and S. N. Kabadi. Some NP-complete problems in quadratic and nonlinear programming. Technical report, 1985. 2
26. I. Necoara, Y. Nesterov, and F. Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175:69–107, 2019. 22, 31
27. A. S. Nemirovsky and D. B. Yudin. Problem complexity and method efficiency in optimization. 1983. 3
28. Y. Nesterov. How to make the gradients small. *Optima. Mathematical Optimization Society Newsletter*, (88):10–11, 2012. 3, 5
29. Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical programming*, 140(1):125–161, 2013. 22
30. Y. Ouyang and Y. Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, (184):1–35, 2021. 4
31. M. O'Neill and S. J. Wright. A log-barrier Newton-CG method for bound constrained optimization with complexity guarantees. *IMA Journal of Numerical Analysis*, 41(1):84–121, 2021. 3
32. J. Pena, J. C. Vera, and L. F. Zuluaga. New characterizations of Hoffman constants for systems of linear constraints. *Mathematical Programming*, 187:79–109, 2021. 36
33. R Tyrrell Rockafellar. *Convex analysis*. Number 28. Princeton university press, 1970. 22, 25
34. M. F. Sahin, A. Eftekhari, A. Alacaoglu, F. Latorre, and V. Cevher. An inexact augmented Lagrangian framework for nonconvex optimization with nonlinear constraints. In *Advances in Neural Information Processing Systems*, pages 13965–13977, 2019. 3
35. H. Sun and M. Hong. Distributed non-convex first-order optimization and information processing: lower complexity bounds and rate optimal algorithms. *IEEE Transactions on Signal processing*, 67(22):5912–5928, 2019. 3, 4, 5, 7, 9, 12
36. N. Xiao, X. Liu, and K.-C. Toh. Dissolving constraints for Riemannian optimization. *Mathematics of Operations Research*, 2023. 3
37. Y. Xie and S. J. Wright. Complexity of proximal augmented Lagrangian for nonconvex optimization with nonlinear equality constraints. *Journal of Scientific Computing*, 86:1–30, 2021. 3
38. M. Yashtini. Convergence and rate analysis of a proximal linearized ADMM for nonconvex nonsmooth optimization. *Journal of Global Optimization*, 84(4):913–939, 2022. 5
39. J. Zhang, H. Lin, S. Jegelka, S. Sra, and S. Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In *International Conference on Machine Learning*, pages 11173–11182. PMLR, 2020. 3
40. J. Zhang and Z. Luo. A proximal alternating direction method of multiplier for linearly constrained nonconvex minimization. *SIAM Journal on Optimization*, 30(3):2272–2302, 2020. 3, 5, 6
41. J. Zhang and Z. Luo. A global dual error bound and its application to the analysis of linearly constrained nonconvex optimization. *SIAM Journal on Optimization*, 32(3):2319–2346, 2022. 3, 5, 6
42. J. Zhang, W. Pu, and Z. Luo. On the iteration complexity of smoothed proximal ALM for nonconvex optimization problem with convex constraints. *Preprint, arXiv:2207.06304*, 2022. 6
43. S. Zhang, J. Yang, C. Guzmán, N. Kiyavash, and N. He. The complexity of nonconvex-strongly-concave minimax optimization. In *Uncertainty in Artificial Intelligence*, pages 482–492. PMLR, 2021. 3

A Proofs of Theorem 3.1 and Corollary 3.1

In this section, we give a complete proof of Theorem 3.1 and Corollary 3.1. We first show a lemma and a proposition. According to the structure of $\mathbf{A}$ and $\bar{\mathbf{A}}$ given in (2.4), $\mathrm{supp}((\mathbf{A}^\top\mathbf{A}\mathbf{x})_i)$ and $\mathrm{supp}((\bar{\mathbf{A}}^\top\bar{\mathbf{A}}\mathbf{x})_i)$ are determined by $\mathrm{supp}(\mathbf{x}_{i-1})$, $\mathrm{supp}(\mathbf{x}_i)$ and $\mathrm{supp}(\mathbf{x}_{i+1})$. Also, $\mathrm{supp}(\mathbf{prox}_{\eta\bar{g}}(\mathbf{y}))$, $\bar{\mathbf{A}}^\top\mathbf{y}$ and $\mathbf{A}\mathbf{x}$ have a similar property. They are stated in the following lemma.

Lemma A.1 *Let $\mathbf{x}$ be the structured vector given in (2.1), $\mathbf{A}$ in (2.4), and $\bar{g}$ be given in (2.5). Define $\mathbf{x}_0 = \mathbf{x}_{m+1} = \mathbf{0} \in \mathbb{R}^{\bar{d}}$. The following statements hold:*

(a) *Let $\widehat{\mathbf{x}} = (\widehat{\mathbf{x}}_1^\top, \ldots, \widehat{\mathbf{x}}_m^\top)^\top \in \mathbf{span}\{\mathbf{A}^\top\mathbf{A}\mathbf{x}, \bar{\mathbf{A}}^\top\bar{\mathbf{A}}\mathbf{x}\}$ with $\widehat{\mathbf{x}}_i \in \mathbb{R}^{\bar{d}}$. Then*

$$\mathrm{supp}(\widehat{\mathbf{x}}_i) \subset \mathrm{supp}(\mathbf{x}_{i-1}) \cup \mathrm{supp}(\mathbf{x}_i) \cup \mathrm{supp}(\mathbf{x}_{i+1}),\ \forall\, i \in [1, m]. \tag{A.1}$$

(b) *Let $\mathbf{y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_{3m_2-1}^\top)^\top$ with $\mathbf{y}_j \in \mathbb{R}^{\bar{d}}$ and $\widetilde{\mathbf{x}} = \bar{\mathbf{A}}^\top\mathbf{y}$. Then*

$$\mathrm{supp}(\widetilde{\mathbf{x}}_i) \subset \begin{cases} \emptyset & \textit{if } i-1, i \notin \mathcal{M}, \\ \mathrm{supp}(\mathbf{y}_j) & \textit{if } i-1 = jm_1 \in \mathcal{M}, \\ \mathrm{supp}(\mathbf{y}_j) & \textit{if } i = jm_1 \in \mathcal{M}, \end{cases} \quad \forall i \in [1, m]. \tag{A.2}$$

(c) *Let* $\widehat{\mathbf{y}} = \bar{\mathbf{A}}\mathbf{x}$ *and* $\widehat{\mathbf{y}} = (\widehat{\mathbf{y}}_1^\top, \ldots, \widehat{\mathbf{y}}_{3m_2-1}^\top)^\top$ *with* $\widehat{\mathbf{y}}_j \in \mathbb{R}^{\bar{d}}$*. Then*

$$\operatorname{supp}(\widehat{\mathbf{y}}_j) \subset \operatorname{supp}(\mathbf{x}_{jm_1}) \cup \operatorname{supp}(\mathbf{x}_{jm_1+1}), \ \forall j \in [1, 3m_2 - 1]. \tag{A.3}$$

(d) *It holds that*

$$\operatorname{supp}(\mathbf{y}) = \operatorname{supp}(\bar{\mathbf{A}}\bar{\mathbf{A}}^\top \mathbf{y}), \ \forall \mathbf{y} \in \mathbb{R}^{\bar{n}}. \tag{A.4}$$

(e) *For any given* $\eta > 0$*, let* $\widetilde{\mathbf{y}} = \mathbf{prox}_{\eta\bar{g}}(\mathbf{y}) = (\widetilde{\mathbf{y}}_1^\top, \ldots, \widetilde{\mathbf{y}}_{3m_2-1}^\top)^\top$ *with* $\widetilde{\mathbf{y}}_j \in \mathbb{R}^{\bar{d}}$*. Then*

$$\operatorname{supp}(\widetilde{\mathbf{y}}_j) \subset \operatorname{supp}(\mathbf{y}_j), \ \forall j \in [1, 3m_2 - 1]. \tag{A.5}$$

Proof. (a) Recall that $\mathbf{H}$ are split into $\bar{\mathbf{A}}$ and $\mathbf{A}$ in rows. The relation in (A.1) immediately follows from (2.35) the observation

$$\mathbf{H}^\top\mathbf{H} = m^2 L_f^2 \underbrace{\left.\begin{bmatrix} \mathbf{I}_{\bar{d}} & -\mathbf{I}_{\bar{d}} & & & \\ -\mathbf{I}_{\bar{d}} & 2\mathbf{I}_{\bar{d}} & -\mathbf{I}_{\bar{d}} & & \\ & \ddots & \ddots & \ddots & \\ & & -\mathbf{I}_{\bar{d}} & 2\mathbf{I}_{\bar{d}} & -\mathbf{I}_{\bar{d}} \\ & & & -\mathbf{I}_{\bar{d}} & \mathbf{I}_{\bar{d}} \end{bmatrix}\right\}}_{m-1 \text{ blocks}} m-1 \text{ blocks} \quad \text{and} \quad \bar{\mathbf{A}}^\top\bar{\mathbf{A}} = \mathbf{H}^\top\mathbf{H} - \mathbf{A}^\top\mathbf{A}. \tag{A.6}$$

(b) The relation in (A.2) immediately follows from the definitions of $\bar{\mathbf{A}}$ and $\mathcal{M}$ in (2.3).
(c) The relation in (A.3) immediately follows from the definition of $\bar{\mathbf{A}}$ and $\mathcal{M}$ in (2.3).
(d) The relation in (A.4) immediately follows from the fact that $\bar{\mathbf{A}}\bar{\mathbf{A}}^\top = 2m^2L_f^2\mathbf{I}_{\bar{n}}$ by the definition of $\bar{\mathbf{A}}$ and $\mathcal{M}$ in (2.3).
(e) Given any $y \in \mathbb{R}$ and any $c > 0$, consider the following optimization problem in $\mathbb{R}$:

$$\widetilde{y} = \arg\min_{z \in \mathbb{R}} \frac{1}{2}(z-y)^2 + c|z| = \operatorname{sign}(y) \cdot (|y| - c)_+. \tag{A.7}$$

Recall the definition of $\bar{g}$ in (2.5), we obtain that

$$\widetilde{\mathbf{y}} = \mathbf{prox}_{\eta\bar{g}}(\mathbf{y}) = \arg\min_{\mathbf{y}'} \frac{\beta}{mL_f}\|\mathbf{y}'\|_1 + \frac{1}{2}\|\mathbf{y}' - \mathbf{y}\|^2.$$

Applying (A.7) to each coordinate of $\widetilde{\mathbf{y}}$ above, we have (A.5) for $j = 1, \ldots, 3m_2 - 1$. The proof is then completed. □

Now we are ready to show the following result on how fast $\operatorname{supp}(\bar{\mathbf{x}}^{(t)})$ and $\operatorname{supp}(\bar{\mathbf{y}}^{(t)})$ can expand with t.

Proposition A.1 *Suppose an algorithm is applied to the reformulation* (SP) *of instance* $\mathcal{P}$ *started from an initial solution* $\mathbf{x}^{(0)} = \mathbf{0}$ *and* $\mathbf{y}^{(0)} = \mathbf{0}$*, and generates a sequence* $\{(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\}_{t\geq 0}$ *that satisfies Assumption 3. By notations in* (2.32) *and* $\mathbf{y}^{(t)} = (\mathbf{y}_1^{(t)\top}, \ldots, \mathbf{y}_{3m_2-1}^{(t)\top})^\top$ *with* $\mathbf{y}_j^{(t)} \in \mathbb{R}^{\bar{d}}$*. It holds, for any* $\bar{j} \in \{2, 3, \ldots, \bar{d}\}$*, that*

$$\operatorname{supp}(\mathbf{x}_i^{(t)}) \subset \{1, \ldots, \bar{j} - 1\} \text{ and } \operatorname{supp}(\mathbf{y}_j^{(t)}) \subset \{1, \ldots, \bar{j} - 1\} \tag{A.8}$$

for $i = 1, \ldots, m, \ j = 1, \ldots, 3m_2 - 1$ *and* $t \leq 1 + m(\bar{j} - 2)/3$.

Proof. We prove this claim by induction on $\bar{j}$. Let $\boldsymbol{\xi}^{(t)} = (\boldsymbol{\xi}_1^{(t)\top}, \ldots, \boldsymbol{\xi}_{3m_2-1}^{(t)\top})^\top$ with $\boldsymbol{\xi}_j^{(t)} \in \mathbb{R}^{\bar{d}}$ and $\boldsymbol{\zeta}^{(t)} = (\boldsymbol{\zeta}_1^{(t)\top}, \ldots, \boldsymbol{\zeta}_{3m_2-1}^{(t)\top})^\top$ with $\boldsymbol{\zeta}_j^{(t)} \in \mathbb{R}^{\bar{d}}$ defined as in Assumption 3 for $t \geq 1$. Since the algorithm is initialized with $\mathbf{x}^{(0)} = \mathbf{0}$ and $\mathbf{y}^{(0)} = \mathbf{0}$, we have $\operatorname{supp}(\nabla f_i(\mathbf{x}_i^{(0)})) \subset \{1\}$ for any i according to Lemma 2.7. Notice $\mathbf{b} = \mathbf{0}$ and $\bar{\mathbf{b}} = \mathbf{0}$. By Assumption 3 and (A.5), we have $\operatorname{supp}(\mathbf{x}_i^{(1)}) \subset \{1\}$ for any i. Meanwhile, we have $\boldsymbol{\xi}_j^{(1)} = \mathbf{0}$ and $\boldsymbol{\zeta}_j^{(1)} = \mathbf{0}$ for any j. This implies $\mathbf{y}_j^{(1)} = \mathbf{0}$. Thus the claim in (A.8) holds for $\bar{j} = 2$. Suppose that we have proved the claim in (A.8) for all $\bar{j} \geq 2$. We next prove it for $\bar{j} + 1$. According to the hypothesis of the induction, we have

$$\begin{aligned} &\operatorname{supp}(\mathbf{x}_i^{(t)}) \subset \{1, \ldots, \bar{j} - 1\} \text{ and } \operatorname{supp}(\mathbf{y}_j^{(t)}) \subset \{1, \ldots, \bar{j} - 1\}, \\ &\forall\, i \in [1, m], \ \forall j \in [1, 3m_2 - 1] \text{ and } r \leq \bar{t} := 1 + m(\bar{j} - 2)/3. \end{aligned} \tag{A.9}$$

Below we let $\widehat{\mathbf{x}}^{(s)}$ be any vector in $\mathbf{span}\left\{\mathbf{A}^\top\mathbf{A}\mathbf{x}^{(s)}, \bar{\mathbf{A}}^\top\bar{\mathbf{A}}\mathbf{x}^{(s)}\right\}$, $\widetilde{\mathbf{x}}^{(s)} = \bar{\mathbf{A}}^\top\mathbf{y}^{(s)}$ and $\widehat{\mathbf{y}}^{(s)} = \bar{\mathbf{A}}\mathbf{x}^{(s)}$ for any $s \geq 0$, and we consider two cases: $\bar{j}$ is even and $\bar{j}$ is odd.

Case 1: Suppose $\bar{j}$ is even. We claim that, for $s = 0, 1, \ldots, \frac{m}{3}$,

$$\mathrm{supp}(\mathbf{x}_i^{(r)}) \subset \left\{\begin{array}{ll} \{1,\ldots,\bar{j}\}, & \text{if } i \in \left[1, \frac{m}{3}+s\right], \\ \{1,\ldots,\bar{j}-1\}, & \text{if } i \in \left[\frac{m}{3}+s+1, m\right], \end{array}\right. \quad \forall r \in [\bar{t}+s] \text{ and} \tag{A.10}$$

$$\mathrm{supp}(\mathbf{y}_j^{(r)}) \subset \left\{\begin{array}{ll} \{1,\ldots,\bar{j}\}, & \text{if } j \in \left[1, m_2 + \lfloor\frac{s}{m_1}\rfloor\right], \\ \{1,\ldots,\bar{j}-1\}, & \text{if } j \in \left[m_2 + \lfloor\frac{s}{m_1}\rfloor + 1, 3m_2 - 1\right], \end{array}\right. \quad \forall r \in [\bar{t}+s]. \tag{A.11}$$

Notice (A.9) implies (A.10) and (A.11) for $s = 0$. Suppose (A.10) and (A.11) hold for an integer s satisfying $0 \leq s \leq \frac{m}{3}$. According to Lemma 2.7 and $\frac{m}{3} + s \leq \frac{2m}{3}$,

$$\mathrm{supp}(\nabla f_i(\mathbf{x}_i^{(r)})) \subset \left\{\begin{array}{ll} \{1,\ldots,\bar{j}\}, & \text{if } i \in \left[1, \frac{m}{3}+s\right], \\ \{1,\ldots,\bar{j}-1\}, & \text{if } i \in \left[\frac{m}{3}+s+1, m\right], \end{array}\right. \quad \forall r \in [\bar{t}+s].$$

In addition, by Lemma A.1(a), we have from (A.2) that

$$\mathrm{supp}(\widehat{\mathbf{x}}_i^{(r)}),\ \mathrm{supp}(\widetilde{\mathbf{x}}_i^{(r)}) \subset \left\{\begin{array}{ll} \{1,\ldots,\bar{j}\}, & \text{if } i \in \left[1, \frac{m}{3}+s+1\right], \\ \{1,\ldots,\bar{j}-1\}, & \text{if } i \in \left[\frac{m}{3}+s+2, m\right], \end{array}\right. \quad \forall r \in [\bar{t}+s].$$

Hence, by Assumption 3, we have

$$\mathrm{supp}(\mathbf{x}_i^{(\bar{t}+s+1)}) \subset \left\{\begin{array}{ll} \{1,\ldots,\bar{j}\}, & \text{if } i \in \left[1, \frac{m}{3}+s+1\right], \\ \{1,\ldots,\bar{j}-1\}, & \text{if } i \in \left[\frac{m}{3}+s+2, m\right]. \end{array}\right.$$

This means the claim in (A.10) holds for $s+1$ as well.

In addition, by the relation in (A.3), we have

$$\mathrm{supp}(\widehat{\mathbf{y}}_j^{(r)}) \subset \left\{\begin{array}{ll} \{1,\ldots,\bar{j}\}, & \text{if } j \in \left[1, m_2 + \lfloor\frac{s}{m_1}\rfloor\right], \\ \{1,\ldots,\bar{j}-1\}, & \text{if } j \in \left[m_2 + \lfloor\frac{s}{m_1}\rfloor + 1, 3m_2 - 1\right], \end{array}\right. \quad \forall r \in [\bar{t}+s].$$

Together with (A.11) and (A.4), the inclusion above implies that

$$\mathrm{supp}(\boldsymbol{\xi}_j^{(\bar{t}+s+1)}) \subset \left\{\begin{array}{ll} \{1,\ldots,\bar{j}\}, & \text{if } j \in \left[1, m_2 + \lfloor\frac{s}{m_1}\rfloor\right], \\ \{1,\ldots,\bar{j}-1\}, & \text{if } j \in \left[m_2 + \lfloor\frac{s}{m_1}\rfloor + 1, 3m_2 - 1\right]. \end{array}\right.$$

It then follows from (A.5) that

$$\mathrm{supp}(\boldsymbol{\zeta}_j^{(\bar{t}+s+1)}) \subset \left\{\begin{array}{ll} \{1,\ldots,\bar{j}\}, & \text{if } j \in \left[1, m_2 + \lfloor\frac{s}{m_1}\rfloor\right], \\ \{1,\ldots,\bar{j}-1\}, & \text{if } j \in \left[m_2 + \lfloor\frac{s}{m_1}\rfloor + 1, 3m_2 - 1\right]. \end{array}\right.$$

By Assumption 3, we have

$$\mathrm{supp}(\mathbf{y}_j^{(\bar{t}+s+1)}) \subset \left\{\begin{array}{ll} \{1,\ldots,\bar{j}\}, & \text{if } j \in \left[1, m_2 + \lfloor\frac{s}{m_1}\rfloor\right], \\ \{1,\ldots,\bar{j}-1\}, & \text{if } j \in \left[m_2 + \lfloor\frac{s}{m_1}\rfloor + 1, 3m_2 - 1\right]. \end{array}\right.$$

This means the claim in (A.11) holds for $s+1$ as well. By induction, (A.10) and (A.11) hold for $s = 0, 1, \ldots, \frac{m}{3}$. Let $s = \frac{m}{3}$ in them. We have $\mathrm{supp}(\mathbf{x}_i^{(r)}) \subset \{1,\ldots,\bar{j}\}$ and $\mathrm{supp}(\mathbf{y}_j^{(r)}) \subset \{1,\ldots,\bar{j}\}$ for any i, j and $r \leq \bar{t} + \frac{m}{3} = 1 + m(\bar{j}-2)/3 + \frac{m}{3} = 1 + m(\bar{j}-1)/3$.

Case 2: Suppose $\bar{j}$ is odd. We claim that, for $s = 0, 1, \ldots, \frac{m}{3}$,

$$\mathrm{supp}(\mathbf{x}_i^{(r)}) \subset \left\{\begin{array}{ll} \{1,\ldots,\bar{j}-1\}, & \text{if } i \in \left[1, \frac{2m}{3}-s\right], \\ \{1,\ldots,\bar{j}\}, & \text{if } i \in \left[\frac{2m}{3}-s+1, m\right], \end{array}\right. \quad \forall r \in [\bar{t}+s], \text{ and} \tag{A.12}$$

$$\mathrm{supp}(\mathbf{y}_j^{(r)}) \subset \begin{cases} \{1,\dots,\bar{j}-1\}, & \text{if } j \in \left[1, 2m_2 - \lceil \frac{s}{m_1} \rceil\right], \\ \{1,\dots,\bar{j}\}, & \text{if } j \in \left[2m_2 - \lceil \frac{s}{m_1} \rceil + 1, 3m_2 - 1\right], \end{cases} \quad \forall r \in [\bar{t}+s]. \tag{A.13}$$

Again (A.9) implies (A.12) and (A.13) for $s = 0$. Suppose (A.12) and (A.13) hold for an integer s satisfying $0 \leq s \leq \frac{m}{3}$. According to Lemma 2.7 and $\frac{2m}{3} - s \geq \frac{m}{3}$,

$$\mathrm{supp}(\nabla f_i(\mathbf{x}_i^{(r)})) \subset \begin{cases} \{1,\dots,\bar{j}-1\}, & \text{if } i \in \left[1, \frac{2m}{3} - s\right], \\ \{1,\dots,\bar{j}\}, & \text{if } i \in \left[\frac{2m}{3} - s + 1, m\right], \end{cases} \quad \forall r \in [\bar{t}+s].$$

In addition, by Lemma A.1(a) and (A.2), we have

$$\mathrm{supp}(\widehat{\mathbf{x}}_i^{(r)}),\ \mathrm{supp}(\widetilde{\mathbf{x}}_i^{(r)}) \subset \begin{cases} \{1,\dots,\bar{j}-1\}, & \text{if } i \in \left[1, \frac{2m}{3} - s - 1\right], \\ \{1,\dots,\bar{j}\}, & \text{if } i \in \left[\frac{2m}{3} - s, m\right], \end{cases} \quad \forall r \in [\bar{t}+s].$$

Hence, by Assumption 3, we have

$$\mathrm{supp}(\mathbf{x}_i^{(\bar{t}+s+1)}) \subset \begin{cases} \{1,\dots,\bar{j}-1\}, & \text{if } i \in \left[1, \frac{2m}{3} - s - 1\right] \\ \{1,\dots,\bar{j}\}, & \text{if } i \in \left[\frac{2m}{3} - s, m\right]. \end{cases}$$

This means claim (A.12) holds for $s+1$ as well.

In addition, by (A.3), we have

$$\mathrm{supp}(\widehat{\mathbf{y}}_j^{(\bar{t}+s+1)}) \subset \begin{cases} \{1,\dots,\bar{j}-1\}, & \text{if } j \in \left[1, 2m_2 - \lceil \frac{s+1}{m_1} \rceil\right], \\ \{1,\dots,\bar{j}\}, & \text{if } j \in \left[2m_2 - \lceil \frac{s+1}{m_1} \rceil + 1, 3m_2 - 1\right]. \end{cases}$$

Together with (A.13), the inclusion above implies that

$$\mathrm{supp}(\boldsymbol{\xi}_j^{(\bar{t}+s+1)}) \subset \begin{cases} \{1,\dots,\bar{j}-1\}, & \text{if } j \in \left[1, 2m_2 - \lceil \frac{s+1}{m_1} \rceil\right], \\ \{1,\dots,\bar{j}\}, & \text{if } j \in \left[2m_2 - \lceil \frac{s+1}{m_1} \rceil + 1, 3m_2 - 1\right]. \end{cases}$$

It then follows from (A.5) that

$$\mathrm{supp}(\boldsymbol{\zeta}_j^{(\bar{t}+s+1)}) \subset \begin{cases} \{1,\dots,\bar{j}-1\}, & \text{if } j \in \left[1, 2m_2 - \lceil \frac{s+1}{m_1} \rceil\right], \\ \{1,\dots,\bar{j}\}, & \text{if } j \in \left[2m_2 - \lceil \frac{s+1}{m_1} \rceil + 1, 3m_2 - 1\right]. \end{cases}$$

By Assumption 3, we have

$$\mathrm{supp}(\mathbf{y}_j^{(\bar{t}+s+1)}) \subset \begin{cases} \{1,\dots,\bar{j}-1\}, & \text{if } j \in \left[1, 2m_2 - \lceil \frac{s+1}{m_1} \rceil\right], \\ \{1,\dots,\bar{j}\}, & \text{if } j \in \left[2m_2 - \lceil \frac{s+1}{m_1} \rceil + 1, 3m_2 - 1\right], \end{cases}$$

which means (A.13) holds for $s+1$ as well. By induction, (A.12) and (A.13) holds for $s = 0, 1, \dots, \frac{m}{3}$. Let $s = \frac{m}{3}$ in (A.13). We have $\mathrm{supp}(\mathbf{x}_i^{(r)}) \subset \{1,\dots,\bar{j}\}$ and $\mathrm{supp}(\mathbf{y}_j^{(r)}) \subset \{1,\dots,\bar{j}\}$ for any i, j and $r \leq \bar{t} + \frac{m}{3} = 1 + m(\bar{j}-2)/3 + \frac{m}{3} = 1 + m(\bar{j}-1)/3$.

Therefore, we have proved that (A.9) holds for $\bar{j}+1$, when $\bar{j}$ is either even or odd. By induction, (A.9) holds for any integer $\bar{j} \in [2, \bar{d}]$, and we complete the proof. □

Now, we are ready to prove Theorem 3.1.

Proof.[of Theorem 3.1] As we discussed below (2.32), we assume $\mathbf{x}^{(0)} = \mathbf{0}$ and $\mathbf{y}^{(0)} = \mathbf{0}$ without loss of generality. Thus by notation in (2.32), Proposition A.1 indicates that $\mathrm{supp}(\mathbf{x}_i^{(t)}) \subset \{1,\dots,\bar{d}-1\}$ and $\mathrm{supp}(\mathbf{y}_j^{(t)}) \subset \{1,\dots,\bar{d}-1\}$ for $i = 1,\dots,m$ and $j = 1, 2, \dots, 3m_2 - 1$ for all $t \leq 1 + m(\bar{d}-2)/3$, which means $[\bar{\mathbf{x}}^{(t)}]_{\bar{d}} = 0$ if $t \leq 1 + m(\bar{d}-2)/3$. Hence, by Lemmas 2.5 and 2.6, we have

$$\max\left\{\left\|\mathbf{H}\mathbf{x}^{(t)}\right\|, \min_{\boldsymbol{\gamma}} \left\|\nabla f_0(\mathbf{x}^{(t)}) + \mathbf{H}^\top \boldsymbol{\gamma}\right\|\right\} \geq \frac{\sqrt{m}}{2} \left\| \frac{1}{m} \sum_{i=1}^{m} \nabla f_i(\bar{\mathbf{x}}^{(t)}) \right\| > \epsilon,\ \forall t \leq 1 + m(\bar{d}-2)/3.$$

Hence, $\mathbf{x}^{(t)}$ is not an ϵ-stationary point of problem (AP) if $t \leq m(\bar{d}-1)/3$. Thus, $(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})$ cannot be an $\epsilon/2$-stationary point of the reformulation (SP) of instance $\mathcal{P}$ according to Corollary 3.1. Moreover, by Lemma 2.2(a) and the facts that $\inf_{\mathbf{y}} \bar{g}(\mathbf{y}) = \bar{g}(\mathbf{y}^{(0)}) = 0$, $\inf_{\mathbf{x},\mathbf{y}}[f_0(\mathbf{x}) + \bar{g}(\mathbf{y})] \geq \inf_{\mathbf{x}} f_0(\mathbf{x})$, it holds that

$$\bar{d} \geq \frac{L_f\left(f_0(\mathbf{x}^{(0)}) - \inf_{\mathbf{x}} f_0(\mathbf{x})\right)}{3000\pi^2}\epsilon^{-2} \geq \frac{L_f\left(f_0(\mathbf{x}^{(0)}) + \bar{g}(\mathbf{y}^{(0)}) - \inf_{\mathbf{x},\mathbf{y}}\left[f_0(\mathbf{x}) + \bar{g}(\mathbf{y})\right]\right)}{3000\pi^2}\epsilon^{-2} = \frac{L_f \Delta_F}{3000\pi^2}\epsilon^{-2}.$$

In other words, in order for $(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})$ to be an $\epsilon/2$-stationary point of (SP), the algorithm needs at least $t = 2 + m(\bar{d}-2)/3$ oracles. Notice

$$2 + m(\bar{d}-2)/3 \geq m\bar{d}/6 \geq \frac{mL_f\Delta_F}{18000\pi^2}\epsilon^{-2} > \frac{\kappa([\bar{\mathbf{A}};\mathbf{A}])L_f\Delta_F}{18000\pi^2}\epsilon^{-2}, \tag{A.14}$$

where the second inequality is because $\bar{d} \geq 5$ and the last inequality is by Lemma 2.3. The conclusion is then proved by replacing ϵ in (A.14) to 2ϵ. □

Finally, we give the proof to Corollary 3.1.

Proof.[of Corollary 3.1] As we discussed in Section 3, we assume $\mathbf{x}^{(0)} = \mathbf{0}$ and $\mathbf{y}^{(0)} = \mathbf{0}$ without loss of generality. Thus by notation in (2.32), Proposition A.1 indicates that $\mathrm{supp}(\mathbf{x}_i^{(t)}) \subset \{1, \ldots, \bar{d}-1\}$ for any $i \in [1, m]$ and any $t \leq 1 + m(\bar{d}-2)/3$, which means $[\bar{\mathbf{x}}^{(t)}]_{\bar{d}} = 0$ if $t \leq 1 + m(\bar{d}-2)/3$, where $\bar{\mathbf{x}}^{(t)} = \frac{1}{m}\sum_{i=1}^m \mathbf{x}_i^{(t)}$.

On the other hand, suppose $\mathbf{x}^*$ with the structure as in (2.32) is an ϵ-stationary point of instance $\mathcal{P}$. Then by Lemma 2.4, it must also be an ϵ-stationary point of (AP). Hence, by Lemmas 2.5 and 2.6, we have $|[\bar{\mathbf{x}}^*]_j| \geq \frac{150\pi\epsilon}{\sqrt{m}L_f}$ for all $j = 1, \ldots, \bar{d}$, where $\bar{\mathbf{x}}^* = \frac{1}{m}\sum_{i=1}^m \mathbf{x}_i^*$. Therefore, by the convexity of the square function, it follows that

$$\|\mathbf{x}^{(t)} - \mathbf{x}^*\|^2 \geq \sum_{i=1}^m \left([\mathbf{x}_i^{(t)}]_{\bar{d}} - [\mathbf{x}_i^*]_{\bar{d}}\right)^2 \geq m\left([\bar{\mathbf{x}}^{(t)}]_{\bar{d}} - [\bar{\mathbf{x}}^*]_{\bar{d}}\right)^2 \geq m\left(\frac{150\pi\epsilon}{\sqrt{m}L_f}\right)^2 > \omega^2,$$

and thus $\mathbf{x}^{(t)}$ is not ω-close to $\mathbf{x}^*$ if $t \leq 1 + m(\bar{d}-2)/3$.

Moreover, by Lemma 2.2(a) and the fact that $g(\mathbf{x}^{(0)}) = 0$ and $g(\mathbf{x}) \geq 0, \forall\, \mathbf{x}$, it holds that

$$\bar{d} \geq \frac{L_f\left(F_0(\mathbf{x}^{(0)}) - \inf_{\mathbf{x}} F_0(\mathbf{x})\right)}{3000\pi^2}\epsilon^{-2} = \frac{L_f\Delta_{F_0}}{3000\pi^2}\epsilon^{-2}.$$

In other words, in order for $\mathbf{x}^{(t)}$ to be ω-close to an ϵ-stationary point of instant $\mathcal{P}$, the algorithm needs at least $t = 2 + m(\bar{d}-2)/3$ oracles. The proof is then completed, by observing

$$2 + m(\bar{d}-2)/3 \geq m\bar{d}/6 \geq \frac{mL_f\Delta_{F_0}}{18000\pi^2}\epsilon^{-2} > \frac{\kappa([\bar{\mathbf{A}};\mathbf{A}])L_f\Delta_{F_0}}{18000\pi^2},$$

where the first inequality is because $\bar{d} \geq 5$, and the last one is by Lemma 2.3. □