Accelerated first-order methods for convex optimization with locally Lipschitz continuous gradient

Zhaosong Lu * Sanyou Mei *

June 1, 2022 (Revised: February 15, 2023; Apri 8, 2023)

Abstract

In this paper we develop accelerated first-order methods for convex optimization with *locally Lipschitz* continuous gradient (LLCG), which is beyond the well-studied class of convex optimization with Lipschitz continuous gradient. In particular, we first consider unconstrained convex optimization with LLCG and propose accelerated proximal gradient (APG) methods for solving it. The proposed APG methods are equipped with a verifiable termination criterion and enjoy an operation complexity of $\mathcal{O}(\varepsilon^{-1/2}\log\varepsilon^{-1})$ and $\mathcal{O}(\log\varepsilon^{-1})$ for finding an ε-residual solution of an unconstrained convex and strongly convex optimization problem, respectively. We then consider constrained convex optimization with LLCG and propose an first-order proximal augmented Lagrangian method for solving it by applying one of our proposed APG methods to approximately solve a sequence of proximal augmented Lagrangian subproblems. The resulting method is equipped with a verifiable termination criterion and enjoys an operation complexity of $\mathcal{O}(\varepsilon^{-1}\log\varepsilon^{-1})$ and $\mathcal{O}(\varepsilon^{-1/2}\log\varepsilon^{-1})$ for finding an ε-KKT solution of a constrained convex and strongly convex optimization problem, respectively. All the proposed methods in this paper are *parameter-free* or *almost parameter-free* except that the knowledge on convexity parameter is required. In addition, preliminary numerical results are presented to demonstrate the performance of our proposed methods. To the best of our knowledge, no prior studies were conducted to investigate accelerated first-order methods with complexity guarantees for convex optimization with LLCG. All the complexity results obtained in this paper are entirely new.

Keywords: Convex optimization, locally Lipschitz continuous gradient, proximal gradient method, proximal augmented Lagrangian method, accelerated first-order methods, iteration complexity, operation complexity

Mathematics Subject Classification: 90C25, 90C30, 90C46, 49M37

1 Introduction

In this paper we first consider unconstrained convex optimization[1]

$$F^* = \min_x \{F(x) \coloneqq f(x) + P(x)\}, \tag{1}$$

where $F^* \in \mathbb{R}$, $f, P : \mathbb{R}^n \to (-\infty, \infty]$ are proper closed convex functions, f is differentiable on $\mathrm{cl}(\mathrm{dom}(P))$, and ∇f is *locally Lipschitz* continuous[2] on $\mathrm{cl}(\mathrm{dom}(P))$, where $\mathrm{dom}(P)$ denotes the domain of P and $\mathrm{cl}(\mathrm{dom}(P))$ denotes its closure. It shall be mentioned that $\mathrm{dom}(P)$ is possibly *unbounded*. Problem (1) is beyond the well-studied class of problems in the form of (1) yet with ∇f being (*globally*) *Lipschitz* continuous on $\mathrm{cl}(\mathrm{dom}(P))$ or $\mathbb{R}^n$. For example, the problem of minimizing a convex high-degree polynomial function over a closed unbounded convex set is a special case of (1), but it does not belong to the latter class in general. In addition, it is sometimes easier to verify local Lipschitz continuity than Lipschitz continuity of ∇f on $\mathrm{cl}(\mathrm{dom}(P))$. For example, when f is twice differentiable in an open set containing $\mathrm{cl}(\mathrm{dom}(P))$, it is straightforward to see that ∇f is locally Lipschitz continuous on $\mathrm{cl}(\mathrm{dom}(P))$; however, verifying Lipschitz continuity of ∇f may require exploring the expression of ∇f and can be a nontrivial task.

The well-known special case of problem (1) with ∇f being *Lipschitz* continuous on $\mathrm{cl}(\mathrm{dom}(P))$ or $\mathbb{R}^n$ has been extensively studied in the literature. In particular, accelerated proximal gradient (APG) methods [3, 16] and

*Department of Industrial and Systems Engineering, University of Minnesota, USA (email: zhaosong@umn.edu, mei00035@umn.edu).

[1] We refer to problem (1) as an unconstrained optimization problem just for convenience. Strictly speaking, it can be a constrained optimization problem. For example, when P is the indicator function of a closed convex set, it reduces to the problem of minimizing f over this set.

[2] See Subsection 1.1 for the definition of locally Lipschitz continuity.

their variants [4, 9, 20] were proposed for solving it. From theoretical perspective, these methods enjoy an optimal iteration complexity of $\mathcal{O}(\varepsilon^{-1/2})$ for finding an ε-gap solution of (1), namely, a point x satisfying $F(x) - F^* \le \varepsilon$. However, since F^* is typically unknown, there is a lack of a verifiable termination criterion for them to find an ε-gap solution of (1) in general. To overcome this issue, a nearly optimal proximal gradient method was recently proposed in [6] for solving such a special case of (1). This method is equipped with a verifiable termination criterion based on the norm of a gradient mapping of (1) and enjoys an iteration complexity of $\mathcal{O}(\varepsilon^{-1/2}\log\varepsilon^{-1})$ for finding an ε-norm solution of (1), namely, a point at which the norm of a gradient mapping of (1) is no more than ε. It shall be mentioned that these methods [3, 4, 6, 9, 16, 20] and their analysis rely on the *Lipschitz* continuity of ∇f on $\mathrm{cl}(\mathrm{dom}(P))$ or $\mathbb{R}^n$. Indeed, they require either an explicitly known global Lipschitz constant of ∇f [4, 9, 20] or an estimated one obtained by a backtracking line search scheme [3, 6, 16]. When ∇f is merely locally Lipschitz continuous, a global Lipschitz constant of ∇f clearly does not exist and also the sequence of estimated Lipschitz constants in [3, 6, 16] can blow up because the solution sequence is possibly unbounded. If the latter case occurs, the methods may not converge and the complexity analysis of the methods in [3, 6, 16] will no longer hold. As a result, these methods are not applicable to (1) or lack complexity guarantees in general when ∇f is merely *locally Lipschitz* continuous on $\mathrm{cl}(\mathrm{dom}(P))$.

To handle the challenge of the local Lipschitz continuity of ∇f, we modify [9, Algorithm 1 with a single block] by incorporating a backtracking line search scheme and an adaptive update strategy on the algorithm parameters to propose an APG method (see Algorithm 1) for solving problem (1). Interestingly, the solution sequence and the sequence of estimated (local) Lipschitz constants obtained by the proposed APG method can be proved to be bounded, which overcome the aforementioned issues of the methods in [3, 6, 16]. Moreover, this method is shown to enjoy a nice iteration complexity of $\mathcal{O}(\varepsilon^{-1/2})$ and $\mathcal{O}(\log\varepsilon^{-1})$ for finding an ε-gap solution of (1) when f is convex and strongly convex, respectively. Yet, since F^* is typically unknown, it is difficult to come up with a verifiable termination criterion for this method to find an ε-gap solution of (1). To circumvent this issue, we further propose an APG method with a *verifiable* termination criterion (see Algorithm 2) for (1) with a *strongly convex* f, and show that it enjoys an iteration and operation complexity[3] of $\mathcal{O}(\log\varepsilon^{-1})$ for finding an ε-residual solution of (1), namely, a point x satisfying $\mathrm{dist}(0, \partial F(x)) \le \varepsilon$.[4] We also propose an APG method with a *verifiable* termination criterion (see Algorithm 4) for (1) with a *convex but non-strongly convex* f by applying Algorithm 2 to a sequence of strongly convex optimization problems arising from a perturbation of (1), and show that it enjoys an operation complexity of $\mathcal{O}(\varepsilon^{-1/2}\log\varepsilon^{-1})$ for finding an ε-residual solution of (1). All the proposed APG methods are *parameter-free* or *almost parameter-free* except that the knowledge on convexity parameter of f is required.

Secondly, we consider constrained convex optimization in the form of

$$\begin{array}{rl} \bar{F}^* = \min & \{F(x) := f(x) + P(x)\} \\ \text{s.t.} & -g(x) \in \mathcal{K}, \end{array} \tag{2}$$

where $\mathcal{K} \subseteq \mathbb{R}^m$ is a closed convex cone, $f, P : \mathbb{R}^n \to (-\infty, \infty]$ are proper closed convex functions, f and g are differentiable on $\mathrm{cl}(\mathrm{dom}(P))$, ∇f and ∇g are *locally Lipschitz* continuous on $\mathrm{cl}(\mathrm{dom}(P))$, and g is $\mathcal{K}$-convex, that is,

$$\alpha g(x) + (1-\alpha)g(y) - g(\alpha x + (1-\alpha)y) \in \mathcal{K}, \quad \forall x, y \in \mathbb{R}^n,\ \alpha \in [0,1].$$

It shall be mentioned that $\mathrm{dom}(P)$ is possibly *unbounded*.

Problem (2) includes a rich class of problems as a special case. For example, when $\mathcal{K} = \mathbb{R}^{m_1}_+ \times \{0\}^{m_2}$ for some m_1 and m_2, $g(x) = (g_1(x), \ldots, g_{m_1}(x), h_1(x), \ldots, h_{m_2}(x))^T$ with convex g_i's and affine h_j's, and $P(x)$ is the indicator function of a simple convex set $X \subseteq \mathbb{R}^n$, problem (2) reduces to an ordinary convex optimization problem

$$\min_{x \in X}\{f(x) : g_i(x) \le 0,\ i = 1, \ldots, m_1; h_j(x) = 0,\ j = 1, \ldots, m_2\}.$$

Numerous first-order methods were developed for solving some special cases of (2) in the literature. For example, a variant of Tseng's modified forward-backward splitting method was proposed in [14] for (2) with g being an affine map, $\mathcal{K} = \{0\}^m$, and ∇f being *Lipschitz* continuous on $\mathrm{cl}(\mathrm{dom}(P))$. Also, first-order penalty methods were proposed in [7] for (2) with g being an affine map, P being the indicator function of a simple

[3] The operation complexity of a proximal gradient method for problem (1) is measured by the amount of its fundamental operations consisting of evaluations of ∇f and proximal operator of P.

[4] $\mathrm{dist}(z, \Omega) = \min_y\{\|z - y\| : y \in \Omega\}$ for any $z \in \mathbb{R}^n$ and closed set $\Omega \subseteq \mathbb{R}^n$. In addition, an ε-residual solution x of (1) satisfying $\|x\| \le \Delta$ for some $\Delta > 0$ independent on ε is an $\mathcal{O}(\varepsilon)$-gap solution, because $F(x) - F^* \le \|x - x^*\|\mathrm{dist}(0, \partial F(x)) \le (\Delta + \|x^*\|)\varepsilon$ for any optimal solution x^* of (1). However, the converse may not be true.

compact convex set, and ∇f being *Lipschitz* continuous on this set. In addition, first-order augmented Lagrangian (AL) methods were developed in [1, 15] for (2) with g being an affine map, P having a *bounded* domain or being the indicator function of a simple *compact* convex set, and ∇f being *Lipschitz* continuous on $\mathbb{R}^n$. Also, first-order AL methods were proposed in [8, 10, 17] with $\mathcal{K} = \{0\}^m$, g being an affine map, P having *bounded* domain or being the indicator function of a simple *compact* convex set, and ∇f being *Lipschitz* continuous on this set or $\mathbb{R}^n$. For these special cases, first-order iteration complexity was established for the methods [1, 15, 17] for finding an ε-gap solution[5] of (2) and for the methods [7, 8, 14] for finding an ε-KKT type solution, which is similar to the one introduced in Definition 2 in Section 3. Since F^* is typically unknown, there is a lack of a verifiable termination criterion for the methods [1, 15, 17] to find an ε-gap solution of (2) in general. In contrast, ε-KKT type of solutions can generally be verified and the methods [7, 8, 14] are equipped with a usually verifiable termination criterion for finding an ε-KKT type solution of the aforementioned special cases of (2).

In addition to the above methods, a first-order proximal AL method was recently proposed in [12, Algorithm 2] for solving a special case of problem (2) with P having a *compact* domain and ∇f and ∇g being *Lipschitz* continuous on $\mathrm{dom}(P)$. At each iteration, this method applies a variant of Nesterov's optimal first-order method [12, Algorithm 3] to approximately solve a proximal AL subproblem and then updates the Lagrangian multiplier by a classical scheme. This method enjoys two nice features: (i) it is equipped with a verifiable termination criterion; (ii) it achieves a best-known operation complexity of $\mathcal{O}(\varepsilon^{-1}\log\varepsilon^{-1})$ for finding an ε-KKT solution[6] of such a special case of (2).

It shall be mentioned that the aforementioned methods in [1, 7, 8, 10, 12, 14, 15, 17] and their analysis rely on *boundedness* of $\mathrm{dom}(P)$ and/or *Lipschitz* continuity of ∇f and ∇g on $\mathrm{cl}(\mathrm{dom}(P))$ or $\mathbb{R}^n$. Indeed, these methods use the APG method [16] or its variant as a subproblem solver. Based on the above discussion, such a subproblem solver is not applicable or lacks complexity guarantees in general when $\mathrm{dom}(P)$ is unbounded or ∇f and ∇g are merely locally Lipschitz continuous on $\mathrm{cl}(\mathrm{dom}(P))$, because the gradient of the smooth component in the objective function of the subproblems is merely locally Lipschitz continuous. As a result, these methods are not applicable or lack complexity guarantees in general when $\mathrm{dom}(P)$ is *unbounded* or ∇f and ∇g are merely *locally Lipschitz* continuous on $\mathrm{cl}(\mathrm{dom}(P))$.

In this paper we propose a first-order proximal AL method for solving problem (2) by following the same framework as [12, Algorithm 2] except that the proximal AL subproblems are approximately solved by our APG method, namely, Algorithm 2. Though the gradient of the smooth component in the objective function of these subproblems is merely locally Lipschitz continuous, their approximate solutions can be found by our APG method with complexity guarantees. As a result, our first-order proximal AL method overcomes the aforementioned issue faced by the methods in [1, 7, 8, 10, 12, 14, 15, 17]. Besides, our method is equipped with a *verifiable* termination criterion and *almost parameter-free* except that the knowledge on convexity parameter of f is required. Moreover, we show that it achieves an operation complexity of $\mathcal{O}(\varepsilon^{-1}\log\varepsilon^{-1})$ and $\mathcal{O}(\varepsilon^{-1/2}\log\varepsilon^{-1})$ for finding an ε-KKT solution of (2) when f is convex and strongly convex, respectively.

The main contributions of our paper are summarized as follows.

- We propose and analyze APG methods for solving problem (1) under *local Lipschitz* continuity of ∇f on $\mathrm{cl}(\mathrm{dom}(P))$ for the first time. Our proposed methods are almost parameter-free, equipped with a verifiable termination criterion, and enjoy an operation complexity of $\mathcal{O}(\varepsilon^{-1/2}\log\varepsilon^{-1})$ and $\mathcal{O}(\log\varepsilon^{-1})$ for finding an ε-residual solution of (1) when f is convex and strongly convex, respectively.

- We propose and analyze a first-order proximal AL method for solving problem (2) under *local Lipschitz* continuity of ∇f and ∇g on $\mathrm{cl}(\mathrm{dom}(P))$ and possible *unboundedness* of $\mathrm{dom}(P)$ for the first time. Our proposed method is almost parameter-free, equipped with a verifiable termination criterion, and enjoys an operation complexity of $\mathcal{O}(\varepsilon^{-1}\log\varepsilon^{-1})$ and $\mathcal{O}(\varepsilon^{-1/2}\log\varepsilon^{-1})$ for finding an ε-KKT solution of (2) when f is convex and strongly convex, respectively.

The rest of this paper is organized as follows. In Subsection 1.1 we introduce some notation and terminology. In Section 2 we propose accelerated proximal gradient methods for problem (1) and study their worst-case complexity. In Section 3 we propose a first-order proximal augmented Lagrangian method for problem (2) and study its worst-case complexity. In addition, we present some preliminary numerical results and the proofs of the main results in Sections 4 and 5. Finally, we make some concluding remarks in Section 6.

[5] An ε-gap solution of problem (2) is a point x satisfying $|F(x) - \bar{F}^*| \le \varepsilon$ and $\mathrm{dist}(g(x), -\mathcal{K}) \le \varepsilon$.

[6] An ε-KKT solution of (2) is generally an $\mathcal{O}(\varepsilon)$-gap solution of (2) (see Theorems 3 and 6 of [12]). However, the converse may not be true.

1.1 Notation and terminology

The following notation will be used throughout this paper. Let $\mathbb{R}^n$ denote the Euclidean space of dimension n, $\langle\cdot,\cdot\rangle$ denote the standard inner product, and $\|\cdot\|$ stand for the Euclidean norm or its induced matrix norm. For any $\omega\in\mathbb{R}$, let $\omega_+=\max\{\omega,0\}$ and $\lceil\omega\rceil$ denote the least integer number greater than or equal to ω. Let $\mathbb{Z}_+$ denote the set of positive integers. For any $t,M\in\mathbb{Z}_+$, $\mathrm{mod}(t,M)$ denotes the remainder of t when divided by M.

For a closed convex function $P:\mathbb{R}^n\to(-\infty,\infty]$, let ∂P and $\mathrm{dom}(P)$ denote the subdifferential and domain of P, respectively. The proximal operator associated with P is denoted by prox_P, that is,

$$\mathrm{prox}_P(z)=\arg\min_{x\in\mathbb{R}^n}\left\{\frac{1}{2}\|x-z\|^2+P(x)\right\}\quad\forall z\in\mathbb{R}^n.$$

Since evaluation of $\mathrm{prox}_{\gamma P}(z)$ is often as cheap as that of $\mathrm{prox}_P(z)$, we count evaluation of $\mathrm{prox}_{\gamma P}(z)$ as one evaluation of proximal operator of P for any $\gamma>0$ and $z\in\mathbb{R}^n$. For a mapping $h:\mathbb{R}^n\to\mathbb{R}^l$, ∇h denotes the transpose of the Jacobian of h. ∇h is called L-Lipschitz continuous on a set Ω for some constant $L>0$ if $\|\nabla h(x)-\nabla h(y)\|\le L\|x-y\|$ for all $x,y\in\Omega$. In addition, ∇h is called *locally Lipschitz continuous* on Ω if for any $x\in\Omega$, there exist $L_x>0$ and an open set $\mathcal{U}_x$ containing x such that ∇h is L_x-Lipschitz continuous on $\mathcal{U}_x$.

Given a nonempty closed convex set $\Omega\subseteq\mathbb{R}^n$, $\mathrm{dist}(x,\Omega)$ stands for the Euclidean distance from x to Ω, and $\Pi_\Omega(x)$ denotes the Euclidean projection of x onto C. The normal cone of Ω at any $x\in\Omega$ is denoted by $\mathcal{N}_\Omega(x)$. For a closed convex cone $\mathcal{K}\subseteq\mathbb{R}^m$, we use $\mathcal{K}^*$ to denote the dual cone of $\mathcal{K}$, that is, $\mathcal{K}^*=\{y\in\mathbb{R}^m:\langle y,x\rangle\ge 0,\ \forall x\in\mathcal{K}\}$.

2 Accelerated proximal gradient methods for unconstrained convex optimization

In this section we consider problem (1) and propose accelerated proximal gradient (APG) methods for solving it. In particular, we aim to find an *ϵ-residual solution* of (1), which is defined below.

Definition 1. *Given any $\epsilon>0$, we say $x\in\mathbb{R}^n$ is an ϵ-residual solution of problem* (1) *if it satisfies* $\mathrm{dist}(0,\partial F(x))\le\epsilon$.

To proceed, let $\mu\ge 0$ denote the *convexity parameter* of f on $\mathrm{dom}(P)$, that is,

$$f(y)\ge f(x)+\langle\nabla f(x),y-x\rangle+\frac{\mu}{2}\|x-y\|^2,\quad\forall x\in\mathrm{dom}(P),y\in\mathbb{R}^n. \tag{3}$$

Clearly, f is strongly convex on $\mathrm{dom}(P)$ when $\mu>0$. In addition, we assume that the proximal operator associated with P can be exactly evaluated and problem (1) has at least one optimal solution. Let x^* be an arbitrary optimal solution of (1) and fixed throughout this section.

2.1 An APG method without a termination criterion for problem (1)

We propose an APG method for (1) as follows, which is a modification of [9, Algorithm 1 with a single block] by incorporating a backtracking line search scheme and an adaptive update strategy on the algorithm parameters.

Algorithm 1 An APG method without a termination criterion for problem (1)

Input: $\gamma_0 \in (0, 1/\mu]$,[7] $0 < \alpha_0 \in [\sqrt{\mu\gamma_0}, 1]$, $\delta \in (0,1)$, and $x^1 = z^1 \in \mathrm{dom}(P)$.

1: **for** $t = 1, 2, \ldots$ **do**

2: Compute

$$y^t = \left((1-\alpha_t)x^t + \alpha_t(1-\beta_t)z^t\right)/(1-\alpha_t\beta_t), \tag{4}$$

$$z^{t+1} = \arg\min_x \left\{\gamma_t[\langle \nabla f(y^t), x\rangle + P(x)] + \frac{\alpha_t}{2}\|x - \beta_t y^t - (1-\beta_t)z^t\|^2\right\}, \tag{5}$$

$$x^{t+1} = (1-\alpha_t)x^t + \alpha_t z^{t+1}, \tag{6}$$

where $\gamma_t = \gamma_0\delta^{n_t}$ and $\beta_t = \mu\gamma_t\alpha_t^{-1}$ with $\alpha_t \in (0,1]$ being the solution of

$$\gamma_{t-1}\alpha_t^2 = (1-\alpha_t)\alpha_{t-1}^2\gamma_t + \mu\alpha_t\gamma_t\gamma_{t-1}, \tag{7}$$

and n_t being the smallest non-negative integer such that

$$2\gamma_t\left(f(x^{t+1}) - f(y^t) - \langle \nabla f(y^t), x^{t+1} - y^t\rangle\right) \le \|x^{t+1} - y^t\|^2. \tag{8}$$

3: **end for**

Remark 1. *(i) Algorithm 1 is almost parameter-free except that the convexity parameter μ of f is required.*

(ii) One can observe that the fundamental operations of Algorithm 1 consist of evaluations of ∇f and proximal operator of P. Specifically, at iteration t, Algorithm 1 requires $n_t + 1$ evaluations of ∇f and proximal operator of P for finding x^{t+1} satisfying (8).

(iii) Notice from Algorithm 1 that $0 < \alpha_0 \in [\sqrt{\mu\gamma_0}, 1]$, which implies $\alpha_0 \in (0,1]$ regardless of $\mu = 0$ or $\mu > 0$. Suppose that $\alpha_{t-1} \in (0,1]$ and $\gamma_{t-1}, \gamma_t \in (0, \gamma_0]$ are given for some $t \ge 1$. Then $\alpha_t \in (0,1]$ is well defined by the equation (7). Indeed, let $\phi(\alpha) = \gamma_{t-1}\alpha^2 - (1-\alpha)\alpha_{t-1}^2\gamma_t - \mu\alpha\gamma_t\gamma_{t-1}$. Observe that $\phi(0) = -\alpha_{t-1}^2\gamma_t < 0$ and $\phi(1) = \gamma_{t-1}(1-\mu\gamma_t) \ge \gamma_{t-1}(1-\mu\gamma_0) \ge 0$ due to $\gamma_0 \in (0, 1/\mu]$. Hence, (7) has a solution in $(0,1]$ and α_t is well-defined.

We next study *well-definedness* of Algorithm 1 and also its *convergence rate* in terms of $F(x^t) - F(x^*)$. To proceed, we define

$$r_0 = \sqrt{F(x^1) - F(x^*) + \frac{\alpha_0^2}{2\gamma_0}\|x^1 - x^*\|^2}, \quad \mathcal{S} = \left\{x \in \mathrm{dom}(P) : \|x - x^*\| \le \frac{\sqrt{2\gamma_0}r_0}{\alpha_0}\right\}. \tag{9}$$

The following lemma establishes that ∇f is *Lipschitz* continuous on $\mathcal{S}$ and also on an enlarged set induced by α_0, γ_0, r_0, x^*, f and $\mathcal{S}$, albeit ∇f is *locally Lipschitz* continuous on $\mathrm{cl}(\mathrm{dom}(P))$. This result will play an important role in this section.

Lemma 1. *Let r_0 and $\mathcal{S}$ be defined in (9), and let γ_0 and α_0 be the input parameters of Algorithm 1. Then the following statements hold.*

(i) ∇f is $L_{\mathcal{S}}$-Lipschitz continuous on $\mathcal{S}$ for some constant $L_{\mathcal{S}} > 0$.

(ii) ∇f is $L_{\widehat{\mathcal{S}}}$-Lipschitz continuous on $\widehat{\mathcal{S}}$ for some constant $L_{\widehat{\mathcal{S}}} > 0$, where

$$\widehat{\mathcal{S}} = \left\{x \in \mathrm{dom}(P) : \|x - x^*\| \le (1 + \gamma_0 L_{\mathcal{S}})\frac{\sqrt{2\gamma_0}r_0}{\alpha_0}\right\}. \tag{10}$$

Proof. Notice that $\mathcal{S}$ is a convex and bounded subset in $\mathrm{dom}(P)$. By this and the local Lipschitz continuity of ∇f on $\mathrm{cl}(\mathrm{dom}(P))$, it is not hard to observe that there exists some constant $L_{\mathcal{S}} > 0$ such that ∇f is $L_{\mathcal{S}}$-Lipschitz continuous on $\mathcal{S}$. Hence, statement (i) holds and moreover the set $\widehat{\mathcal{S}}$ is well-defined. By a similar argument, one can see that statement (ii) also holds. □

The following theorem shows that Algorithm 1 is *well-defined* at each iteration. Its proof is deferred to Subsection 5.1.

[7] By convention, we define $1/0 = \infty$. Consequently, when $\mu = 0$, γ_0 can be any positive number.

Theorem 1. *Algorithm 1 is well-defined at each iteration. Moreover, $x^t, y^t, z^t \in \mathcal{S}$ and $n_t \le N$ for all $t \ge 1$, where $\mathcal{S}$ is defined in* (9) *and*

$$N = \left\lceil \frac{\log(\gamma_0 L_{\widehat{\mathcal{S}}})}{\log(1/\delta)} \right\rceil_+. \tag{11}$$

The next theorem presents a result regarding *convergence rate* of Algorithm 1, whose proof is deferred to Section 5.

Theorem 2. *Let $\{x^t\}$ be generated by Algorithm 1. Then for all $t \ge 1$, it holds that*

$$F(x^t) - F(x^*) \le \min\left\{ \left(1 - \sqrt{\mu \min\left\{\gamma_0, \delta L_{\widehat{\mathcal{S}}}^{-1}\right\}}\right)^{t-1},\ 4\left(2 + (t-1)\alpha_0 \sqrt{\min\left\{1, \delta\gamma_0^{-1} L_{\widehat{\mathcal{S}}}^{-1}\right\}}\right)^{-2} \right\} r_0^2. \tag{12}$$

Remark 2. *(i) Despite only assuming local Lipschitz continuity of ∇f on $\mathrm{cl}(\mathrm{dom}(P))$, Algorithm 1 enjoys a similar convergence rate as the optimal APG method [9, Algorithm 1 with a single block] which was proposed and analyzed for solving a special case of problem* (1) *with ∇f being Lipschitz continuous on $\mathbb{R}^n$.*

(ii) An adaptive gradient method was recently proposed in [13, Algorithm 1] for solving a special case of problem (1) *with $P \equiv 0$. It is a variant of classical gradient methods without acceleration and enjoys a much worse convergence rate than the one given in* (12)*. In particular, when f is convex, it has a convergence rate of $\mathcal{O}(1/t)$ (see [13, Theorem 1]).*

From theoretical perspective, it follows from Theorem 2 that Algorithm 1 enjoys an iteration complexity of $\mathcal{O}(\varepsilon^{-1/2})$ and $\mathcal{O}(\log \varepsilon^{-1})$ for finding an ε-gap solution x^t of (1) satisfying $F(x^t) - F^* \le \varepsilon$ when f is convex and strongly convex, respectively. However, since F^*, $L_{\widehat{\mathcal{S}}}^{-1}$ and r_0 are typically unknown, it is difficult to come up with a verifiable termination criterion for Algorithm 1 to find an ε-gap solution of (1). To circumvent this issue, we propose some variants of Algorithm 1 with a verifiable termination criterion in the next two subsections.

2.2 An APG method with a termination criterion for problem (1) with $\mu > 0$

In this subsection we propose an APG method with a *verifiable termination criterion* for finding an *ϵ-residual solution* of problem (1) with $\mu > 0$, namely, f being strongly convex on $\mathrm{dom}(P)$. It is a slight variant of Algorithm 1 by incorporating a termination criterion that is checked only *periodically*.

Algorithm 2 An APG method with a termination criterion for problem (1) with $\mu > 0$

Input: $\epsilon > 0$, $\gamma_0 \in (0, 1/\mu]$, $0 < \alpha_0 \in [\sqrt{\mu\gamma_0}, 1]$, $\delta \in (0,1)$, $M \in \mathbb{Z}_+$, and $x^1 = z^1 \in \mathrm{dom}(P)$.

1: **for** $t = 1, 2, \ldots$ **do**

2: Compute

$$\begin{aligned} y^t &= \left((1-\alpha_t)x^t + \alpha_t(1-\beta_t)z^t\right)/(1-\alpha_t\beta_t), \\ z^{t+1} &= \arg\min_x \left\{\gamma_t[\langle \nabla f(y^t), x\rangle + P(x)] + \frac{\alpha_t}{2}\|x - \beta_t y^t - (1-\beta_t)z^t\|^2\right\}, \\ x^{t+1} &= (1-\alpha_t)x^t + \alpha_t z^{t+1}, \end{aligned}$$

where $\gamma_t = \gamma_0\delta^{n_t}$ and $\beta_t = \mu\gamma_t\alpha_t^{-1}$ with $\alpha_t \in (0,1]$ being the solution of

$$\gamma_{t-1}\alpha_t^2 = (1-\alpha_t)\alpha_{t-1}^2\gamma_t + \mu\alpha_t\gamma_t\gamma_{t-1},$$

and n_t being the smallest non-negative integer such that

$$2\gamma_t\left(f(x^{t+1}) - f(y^t) - \langle \nabla f(y^t), x^{t+1} - y^t\rangle\right) \le \|x^{t+1} - y^t\|^2.$$

3: **if** $\mathrm{mod}(t, M) = 0$ **then**

4: Call Algorithm 3 with $(x^{t+1}, \gamma_0, \delta)$ as the input and output $(\tilde{x}^{t+1}, \tilde{\gamma}_{t+1})$.

5: Terminate the algorithm and output $\tilde{x}^{t+1}$ if

$$\|\tilde{\gamma}_{t+1}^{-1}(x^{t+1} - \tilde{x}^{t+1}) + \nabla f(\tilde{x}^{t+1}) - \nabla f(x^{t+1})\| \le \epsilon. \tag{13}$$

6: **end if**

7: **end for**

Algorithm 3 Adaptive proximal gradient iteration

Input: $v \in \mathcal{S}$ and $\tilde{\gamma}_0, \delta > 0$.

1: Compute

$$\tilde{v} = \arg\min_x \left\{ \tilde{\gamma}\langle \nabla f(v), x\rangle + \tilde{\gamma} P(x) + \frac{1}{2}\|x - v\|^2 \right\}, \tag{14}$$

where $\tilde{\gamma} = \tilde{\gamma}_0 \delta^{\tilde{n}}$ with $\tilde{n}$ being the smallest non-negative integer such that

$$2\tilde{\gamma}(f(\tilde{v}) - f(v) - \langle \nabla f(v), \tilde{v} - v\rangle) \le \|\tilde{v} - v\|^2. \tag{15}$$

2: Terminate the algorithm and output $(\tilde{v}, \tilde{\gamma})$.

Remark 3. *It is clear to see that Algorithm 2 is well-defined at each iteration and equipped with a verifiable termination criterion. In addition, it is almost parameter-free except that the convexity parameter μ of f is required.*

The following theorem presents an *iteration and operation complexity* of Algorithm 2 for finding an ϵ-residual solution of problem (1) with a strongly convex f on $\mathrm{dom}(P)$, whose proof is deferred to Subsection 5.2.

Theorem 3. *Suppose that $\mu > 0$, i.e., f is strongly convex on $\mathrm{dom}(P)$. Let ϵ, M, δ, α_0 and γ_0 be the input parameters of Algorithm 2, r_0 and $L_{\widehat{\mathcal{S}}}$ be given in (9) and Lemma 1 respectively, and let*

$$T = M + \left\lceil \frac{2\log \frac{\epsilon}{r_0\left(\sqrt{2\max\{\gamma_0^{-1}, L_{\widehat{\mathcal{S}}}\delta^{-1}\}} + \sqrt{2\gamma_0} L_{\widehat{\mathcal{S}}}\right)}}{\log\left(1 - \sqrt{\mu \min\left\{\gamma_0, \delta L_{\widehat{\mathcal{S}}}^{-1}\right\}}\right)} \right\rceil_+, \tag{16}$$

$$\bar{N} = (1 + M^{-1}) \left(M + \left\lceil \frac{2\log \frac{\epsilon}{r_0\left(\sqrt{2\max\{\gamma_0^{-1}, L_{\widehat{\mathcal{S}}}\delta^{-1}\}} + \sqrt{2\gamma_0} L_{\widehat{\mathcal{S}}}\right)}}{\log\left(1 - \sqrt{\mu \min\left\{\gamma_0, \delta L_{\widehat{\mathcal{S}}}^{-1}\right\}}\right)} \right\rceil_+ \right) \left(1 + \left\lceil \frac{\log(\gamma_0 L_{\widehat{\mathcal{S}}})}{\log(1/\delta)} \right\rceil_+ \right). \tag{17}$$

Then Algorithm 2 terminates and outputs an ϵ-residual solution of problem (1) in at most T iterations. Moreover, the total number of evaluations of ∇f and proximal operator of P performed in Algorithm 2 is no more than $\bar{N}$, respectively.

Remark 4. *It can be seen from Theorem 3 that Algorithm 2 enjoys an operation complexity of $\mathcal{O}(\log \varepsilon^{-1})$ for finding an ϵ-residual solution of problem (1) with a strongly convex f on $\mathrm{dom}(P)$.*

2.3 An APG method with a termination criterion for problem (1) with $\mu = 0$

In this subsection we propose an APG method with a *verifiable termination criterion* for finding an *ε-residual solution* of problem (1) with $\mu = 0$, namely, f being convex but not strongly convex on $\mathrm{dom}(P)$. In particular, the proposed APG method applies Algorithm 2 to a sequence of strongly convex optimization problems arising from a perturbation of problem (1).

Algorithm 4 An APG method with a termination criterion for problem (1) with $\mu = 0$

Input: $\varepsilon > 0$, $x_0 \in \mathrm{dom}(P)$, $M \in \mathbb{Z}_+$, $0 < \delta < 1$, $\rho_0 > 1$, $0 < \gamma_0 \le \rho_0$, $\alpha_0 \in [\sqrt{\gamma_0/\rho_0}, 1]$, $0 < \eta_0 \le 1$, $\zeta > 1$, $0 < \sigma < 1/\zeta$, $\rho_k = \rho_0\zeta^k$, $\eta_k = \eta_0\sigma^k$ for all $k \ge 0$.

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

2: Call Algorithm 2 with $F \leftarrow F_k$, $f \leftarrow f_k$, $\epsilon \leftarrow \eta_k$, $\mu \leftarrow \rho_k^{-1}$, $x^1 = z^1 \leftarrow x^k$ and the parameters α_0, γ_0, δ and M, and denote its output by x^{k+1}, where

$$f_k(x) = f(x) + \frac{1}{2\rho_k}\|x - x^k\|^2, \quad F_k(x) = f_k(x) + P(x). \tag{18}$$

3: Terminate the algorithm and output x^{k+1} if

$$\frac{1}{\rho_k}\|x^{k+1} - x^k\| \le \frac{\epsilon}{2},\ \eta_k \le \frac{\epsilon}{2}. \tag{19}$$

4: **end for**

Remark 5. *Algorithm 4 is parameter-free and equipped with a verifiable termination criterion. In addition, by the monotonicity of* $\{\rho_k\}$*, one has*

$$0 < \gamma_0 \le \rho_0 \le \rho_k, \qquad \sqrt{\rho_k^{-1}\gamma_0} \le \sqrt{\rho_0^{-1}\gamma_0} \le \alpha_0 \le 1.$$

Consequently, the choice of α_0 *and* γ_0 *in Algorithm 4 satisfies the requirements specified in Algorithm 2. It then follows from Theorem 3 that at the* k*th outer iteration of Algorithm 4,* x^{k+1} *must be successfully generated by Algorithm 2, which is an* η_k*-residual solution of the problem* $\min_x\{F_k(x) = f_k(x) + P(x)\}$*. Thus, it holds that*

$$\mathrm{dist}(0, \partial F_k(x^{k+1})) \le \eta_k. \tag{20}$$

We next study iteration and operation complexity of Algorithm 4 for finding an ε-residual solution of problem (1) with f being convex but not strongly convex on $\mathrm{dom}(P)$. Before proceeding, we introduce some notation that will be used subsequently. We define

$$r_0 = \|x^0 - x^*\|, \quad \theta = \sum_{i=0}^{\infty}\rho_i\eta_i = \frac{\rho_0\eta_0}{1 - \sigma\zeta}, \tag{21}$$

$$\tilde{r}_0 = \max\left\{\sqrt{2\gamma_0\alpha_0^{-2}(F(x^0) - F(x^*)) + r_0^2},\ \sqrt{2\gamma_0\alpha_0^{-2}(r_0 + \theta)\left(\eta_0 + \rho_0^{-1}(r_0 + \theta)\right) + (r_0 + \theta)^2}\right\}. \tag{22}$$

Also, we define

$$\mathcal{Q} = \{x \in \mathrm{dom}(P) : \|x - x^*\| \le \tilde{r}_0 + r_0 + \theta\}. \tag{23}$$

Let $L_{\nabla f}$ be the Lipschitz constant of ∇f on $\mathcal{Q}$ and

$$L = L_{\nabla f} + \rho_0^{-1}, \quad \widehat{\mathcal{Q}} = \{x \in \mathrm{dom}(P) : \|x - x^*\| \le (1 + \gamma_0 L)\tilde{r}_0 + r_0 + \theta\}, \quad \widehat{L} = \widehat{L}_{\nabla f} + \rho_0^{-1}, \tag{24}$$

where $\widehat{L}_{\nabla f}$ is the Lipschitz constant of ∇f on $\widehat{\mathcal{Q}}$. By the local Lipschitz continuity of ∇f on $\mathrm{cl}(\mathrm{dom}(P))$ and a similar argument as in the proof of Lemma 1, one can easily observe that L, $\widehat{L}$, $L_{\nabla f}$, $\widehat{L}_{\nabla f}$, $\mathcal{Q}$, and $\widehat{\mathcal{Q}}$ are well-defined.

The following theorem presents an *iteration and operation complexity* of Algorithm 4 for finding an *ϵ-residual solution* of problem (1) with f being convex but not strongly convex on $\mathrm{dom}(P)$, namely, a point x satisfying $\mathrm{dist}(0, \partial F(x)) \le \varepsilon$, whose proof is deferred to Subsection 5.3.

Theorem 4. *Suppose that* $\mu = 0$*, i.e., f is convex but not strongly convex on* $\mathrm{dom}(P)$*. Let* ε*, M,* δ*,* ρ_0*,* α_0*,* γ_0*,* η_0*,* ζ *and* σ *be the input parameters of Algorithm 5, and let* r_0*,* θ*,* $\tilde{r}_0$ *and* $\widehat{L}$ *be given in* (21)*,* (22) *and* (24)*, respectively. Define*

$$\widetilde{C}_1 = (1 + M^{-1})\left(1 + \left\lceil\frac{\log(\gamma_0\widehat{L})}{\log(1/\delta)}\right\rceil_+\right), \tag{25}$$

$$\widetilde{C}_2 = \frac{\sqrt{\rho_0\zeta}\widetilde{C}_1\left(\log\frac{\alpha_0^2\tilde{r}_0^2\left(\sqrt{\max\{\gamma_0^{-2}, \gamma_0^{-1}\widehat{L}\delta^{-1}\}} + \widehat{L}\right)^2}{\eta_0^2}\right)_+}{(\sqrt{\zeta} - 1)\min\left\{\sqrt{\gamma_0}, \sqrt{\delta\widehat{L}^{-1}}\right\}}, \quad \widetilde{C}_3 = \frac{2\sqrt{\rho_0\zeta}\widetilde{C}_1\log(1/\sigma)}{(\sqrt{\zeta} - 1)\min\left\{\sqrt{\gamma_0}, \sqrt{\delta\widehat{L}^{-1}}\right\}}. \tag{26}$$

Then the following statements hold.

(i) Algorithm 4 outputs an ε-residual solution of problem (1) after at most $K+1$ outer iterations, where

$$K = \left\lceil \max\left\{ \log\left(\frac{2r_0+2\theta}{\rho_0\varepsilon}\right)/\log\zeta, \frac{\log(2\eta_0/\varepsilon)}{\log(1/\sigma)} \right\} \right\rfloor_+. \tag{27}$$

(ii) The total number of evaluations of ∇f and proximal operator of P performed in Algorithm 4 is no more than $\widetilde{N}$, respectively, where

$$\begin{aligned}\widetilde{N} = {} & (M+1)\widetilde{C}_1 + (M+1)\widetilde{C}_1 \left\lceil \max\left\{ \log\left(\frac{2r_0+2\theta}{\varepsilon\rho_0}\right)/\log\zeta, \frac{\log(2\eta_0/\varepsilon)}{\log(1/\sigma)} \right\} \right\rfloor_+ \\ & + \widetilde{C}_2 \max\left\{ \sqrt{\frac{2\zeta(r_0+\theta)}{\varepsilon\rho_0}}, \sqrt{\zeta}\left(\frac{2\eta_0}{\varepsilon}\right)^{\frac{\log\zeta}{2\log(1/\sigma)}}, 1 \right\} \\ & + \widetilde{C}_3 \left\lceil \max\left\{ \log\left(\frac{2r_0+2\theta}{\varepsilon\rho_0}\right)/\log\zeta, \frac{\log(2\eta_0/\varepsilon)}{\log(1/\sigma)} \right\} \right\rfloor_+ \max\left\{ \sqrt{\frac{2\zeta(r_0+\theta)}{\varepsilon\rho_0}}, \sqrt{\zeta}\left(\frac{2\eta_0}{\varepsilon}\right)^{\frac{\log\zeta}{2\log(1/\sigma)}}, 1 \right\}. \end{aligned} \tag{28}$$

Remark 6. *Since $1 < \zeta < 1/\sigma$, it can be seen from Theorem 4 that Algorithm 4 enjoys an operation complexity of $\mathcal{O}(\varepsilon^{-1/2}\log\varepsilon^{-1})$ for finding an ε-residual solution of problem (1) with f being convex but not strongly convex on $\mathrm{dom}(P)$.*

3 A first-order proximal augmented Lagrangian method for constrained convex optimization

In this section we consider problem (2) and propose a first-order proximal augmented Lagrangian (AL) method for solving it. Let $\mu \geq 0$ denote the *convexity parameter* of f on $\mathrm{dom}(P)$, that is, (3) holds for f and μ. Before proceeding, we make the following additional assumptions for problem (2).

Assumption 1. *(a) The proximal operator associated with P and the projection onto $\mathcal{K}^*$ can be exactly evaluated.*

(b) Both problem (2) and its Lagrangian dual problem

$$\sup_{\lambda\in\mathcal{K}^*} \inf_x \{f(x) + P(x) + \langle\lambda, g(x)\rangle\} \tag{29}$$

have optimal solutions, and moreover, they share the same optimal value.

Under the assumptions on problem (2), it can be observed that (x,λ) is a pair of optimal solutions of (2) and (29) if and only if it satisfies the Karush-Kuhn-Tucker (KKT) condition

$$0 \in \begin{pmatrix} \nabla f(x) + \nabla g(x)\lambda + \partial P(x) \\ -g(x) + \mathcal{N}_{\mathcal{K}^*}(\lambda) \end{pmatrix}.$$

In general, it is difficult to find an exact optimal solution of (2) and (29). Instead, for any given $\varepsilon > 0$, we are interested in finding an ε-KKT solution (x,λ) of problems (2) and (29) that is defined below.

Definition 2. *Given any $\varepsilon > 0$, we say $(x,\lambda) \in \mathbb{R}^n \times \mathbb{R}^m$ is an ε-KKT solution of problems (2) and (29) if*

$$\mathrm{dist}(0, \nabla f(x) + \partial P(x) + \nabla g(x)\lambda) \leq \varepsilon, \quad \mathrm{dist}(g(x), \mathcal{N}_{\mathcal{K}^*}(\lambda)) \leq \varepsilon.$$

We next propose a first-order proximal AL method with a verifiable termination criterion for solving problem (2), which follows the same framework as [12, Algorithm 2] except that the proximal AL subproblems are approximately solved by our newly proposed APG method, namely, Algorithm 2. Specifically, at the kth iteration, our method applies Algorithm 2 to approximately solve the proximal AL subproblem

$$\min_x \mathcal{L}(x, \lambda^k; \rho_k) + \frac{1}{2\rho_k}\|x - x^k\|^2$$

for some $\lambda^k \in \mathcal{K}^*$ and $\rho_k > 0$, where $\mathcal{L}$ is the AL function associated with problem (2) defined as

$$\mathcal{L}(x, \lambda; \rho) = f(x) + P(x) + \frac{1}{2\rho}\left(\mathrm{dist}^2\left(\lambda + \rho g(x), -\mathcal{K}\right) - \|\lambda\|^2\right). \tag{30}$$

Algorithm 5 A first-order proximal augmented Lagrangian method for problem (2)

Input: $\varepsilon > 0$, $(x_0, \lambda_0) \in \mathrm{dom}(P) \times \mathcal{K}^*$, $M \in \mathbb{Z}_+$, $0 < \delta < 1$, $\rho_0 > (\mu + \sqrt{\mu^2 + 4})/2$, $\alpha_0 \in [\sqrt{(\mu + 1/\rho_0)/\rho_0}, 1]$, $0 < \eta_0 \le 1$, $\zeta > 1$, $0 < \sigma < 1/\zeta$, $\rho_k = \rho_0\zeta^k$, $\eta_k = \eta_0\sigma^k$ for all $k \ge 0$.

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

2: Call Algorithm 2 with $F \leftarrow F_k$, $f \leftarrow f_k$, $\epsilon \leftarrow \eta_k$, $\gamma_0 \leftarrow \rho_k^{-1}$, $\mu \leftarrow \mu + \rho_k^{-1}$, $x^1 = z^1 \leftarrow x^k$ and the parameters α_0, δ and M, and denote its output by x^{k+1}, where

$$f_k(x) = f(x) + \frac{1}{2\rho_k}\left(\mathrm{dist}^2\left(\lambda^k + \rho_k g(x), -\mathcal{K}\right) - \|\lambda^k\|^2 + \|x - x^k\|^2\right), \quad F_k(x) = f_k(x) + P(x). \tag{31}$$

3: Set $\lambda^{k+1} = \Pi_{\mathcal{K}^*}\left(\lambda^k + \rho_k g(x^{k+1})\right)$.

4: Terminate the algorithm and output (x^{k+1}, λ^{k+1}) if

$$\frac{1}{\rho_k}\|(x^{k+1}, \lambda^{k+1}) - (x^k, \lambda^k)\| \le \frac{\epsilon}{2}, \quad \eta_k \le \frac{\epsilon}{2}. \tag{32}$$

5: **end for**

Remark 7. *(i) Algorithm 5 is equipped with a verifiable termination criterion and almost parameter-free except that the convexity parameter μ of f is required.*

(ii) Since $\rho_0 > (\mu + \sqrt{\mu^2 + 4})/2$, it follows that $\rho_0^{-1} < 1/(\mu + \rho_0^{-1})$. By this, $\alpha_0 \in [\sqrt{(\mu + 1/\rho_0)/\rho_0}, 1]$, and the monotonicity of $\{\rho_k\}$, one has

$$0 < \rho_k^{-1} \le \rho_0^{-1} < \frac{1}{\mu + \rho_0^{-1}} \le \frac{1}{\mu + \rho_k^{-1}}, \qquad \sqrt{(\mu + \rho_k^{-1})\rho_k^{-1}} \le \sqrt{(\mu + \rho_0^{-1})\rho_0^{-1}} \le \alpha_0 \le 1.$$

Consequently, the choice of α_0 and γ_0 in Algorithm 5 satisfies the requirements specified in Algorithm 2. It then follows from Theorem 3 that at the kth outer iteration of Algorithm 5, x^{k+1} must be successfully generated by Algorithm 2, which is an η_k-residual solution of the problem $\min_x\{F_k(x) = f_k(x) + P(x)\}$. Thus, it holds that

$$\mathrm{dist}(0, \partial F_k(x^{k+1})) \le \eta_k. \tag{33}$$

We next study iteration and operation complexity of Algorithm 5 for finding an ε-KKT solution of problems (2) and (29). Before proceeding, we introduce some notation that will be used subsequently.

Let (x^*, λ^*) be an arbitrary pair of optimal solutions of problems (2) and (29) and fixed throughout this section. We define

$$r_0 = \|(x^0, \lambda^0) - (x^*, \lambda^*)\|, \quad \theta = \sum_{i=0}^{\infty}\rho_i\eta_i = \frac{\rho_0\eta_0}{1 - \sigma\zeta}, \quad \widetilde{\mathcal{Q}} = \{x \in \mathrm{dom}(P) : \|x - x^*\| \le r_0 + \theta\}. \tag{34}$$

Let $\widetilde{L}_g$ be the Lipschitz constant of g on $\widetilde{\mathcal{Q}}$ and

$$\tilde{r}_0 = \max\left\{\sqrt{2\rho_0^{-1}\alpha_0^{-2}(F(x^0) - F(x^*)) + \rho_0^{-2}\alpha_0^{-2}(\|\Pi_{\mathcal{K}^*}(\lambda^0 + \rho_0 g(x^0))\|^2 + \|\lambda^0 - \lambda^*\|^2 - \|\lambda^0\|^2) + r_0^2},\right.$$
$$\left.\sqrt{2\rho_0^{-1}\alpha_0^{-2}(r_0 + \theta)\left(\eta_0 + \rho_0^{-1}(r_0 + \theta) + 2\widetilde{L}_g(\zeta + 1)(\|\lambda^*\| + r_0 + \theta) + \rho_0\alpha_0^2(r_0 + \theta)\right)}\right\}. \tag{35}$$

We define

$$\mathcal{Q} = \{x \in \mathrm{dom}(P) : \|x - x^*\| \le \tilde{r}_0 + r_0 + \theta\}. \tag{36}$$

Let $L_{\nabla f}$, $L_{\nabla g}$ and L_g be the Lipschitz constants of ∇f, ∇g and g on $\mathcal{Q}$, respectively, and let

$$C = L_{\nabla g}\sup_{x\in\mathcal{Q}}\|g(x)\| + L_g^2,\ B = L_{\nabla f} + L_{\nabla g}(\|\lambda^*\| + \sqrt{2}r_0),\ L = C + \rho_0^{-1}B + \rho_0^{-1}L_{\nabla g}\theta + \rho_0^{-2}. \tag{37}$$

We define

$$\widehat{\mathcal{Q}} = \{x \in \mathrm{dom}(P) : \|x - x^*\| \le (1+L)\tilde{r}_0 + r_0 + \theta\}. \tag{38}$$

Let $\widehat{L}_{\nabla f}$, $\widehat{L}_{\nabla g}$ and $\widehat{L}_g$ be the Lipschitz constants of ∇f, ∇g and g on $\widehat{\mathcal{Q}}$, respectively, and let

$$\widehat{C} = \widehat{L}_{\nabla g} \sup_{x\in\widehat{\mathcal{Q}}} \|g(x)\| + \widehat{L}_g^2,\ \widehat{B} = \widehat{L}_{\nabla f} + \widehat{L}_{\nabla g}(\|\lambda^*\| + \sqrt{2}r_0),\ \widehat{L} = \widehat{C} + \rho_0^{-1}\widehat{B} + \rho_0^{-1}\widehat{L}_{\nabla g}\theta + \rho_0^{-2}. \tag{39}$$

By the local Lipschitz continuity of ∇f and ∇g on $\mathrm{cl}(\mathrm{dom}(P))$ and a similar argument as in the proof of Lemma 1, one can easily observe that $\widehat{L}_g$, $L_{\nabla f}$, $L_{\nabla g}$, L_g, $\widehat{L}_{\nabla f}$, $\widehat{L}_{\nabla g}$, $\widehat{L}_g$, B, C, L, $\widehat{B}$, $\widehat{C}$, $\widehat{L}$, $\mathcal{Q}$, and $\widehat{\mathcal{Q}}$ are well-defined.

The following theorem presents an *iteration and operation complexity* of Algorithm 5 for finding an *ϵ-KKT solution* of problems (2) and (29), whose proof is deferred to Subsection 5.4.

Theorem 5. *Let ε, M, δ, ρ_0, α_0, η_0, ζ and σ be the input parameters of Algorithm 5, and let r_0, θ, $\tilde{r}_0$ and $\widehat{L}$ be given in* (34), (35) *and* (39), *respectively. Define*

$$\widehat{C}_1 = (1+M^{-1})\left(1 + \left\lceil \frac{\log \widehat{L}}{\log(1/\delta)} \right\rceil_+\right), \tag{40}$$

$$\widehat{C}_2 = \frac{\widehat{C}_1\left(\log \frac{\rho_0^2\alpha_0^2\tilde{r}_0^2\left(\sqrt{\max\{1,\widehat{L}\delta^{-1}\}}+\widehat{L}\right)^2}{\eta_0^2}\right)_+}{(\sqrt{\zeta}-1)\min\left\{1, \sqrt{\delta\widehat{L}^{-1}}\right\}}, \quad \widehat{C}_3 = \frac{2\widehat{C}_1\log(\zeta/\sigma)}{(\sqrt{\zeta}-1)\min\left\{1, \sqrt{\delta\widehat{L}^{-1}}\right\}}. \tag{41}$$

Then the following statements hold.

(i) Algorithm 5 outputs an ε-KKT solution of problems (2) *and* (29) *after at most $K+1$ outer iterations, where*

$$K = \left\lceil \max\left\{\log\left(\frac{2r_0+2\theta}{\rho_0\varepsilon}\right)/\log\zeta, \frac{\log(2\eta_0/\varepsilon)}{\log(1/\sigma)}\right\}\right\rceil_+. \tag{42}$$

(ii) If $\mu = 0$, i.e., f is convex but not strongly convex, the total number of evaluations of ∇f, ∇g, proximal operator of P and projection onto $\mathcal{K}^*$ performed in Algorithm 5 is no more than $\widehat{N}$, respectively, where

$$\begin{aligned}\widehat{N} = {} & 1 + (M+1)\widehat{C}_1 + \left(1+(M+1)\widehat{C}_1\right)\left\lceil \max\left\{\log\left(\frac{2r_0+2\theta}{\varepsilon\rho_0}\right)/\log\zeta, \frac{\log(2\eta_0/\varepsilon)}{\log(1/\sigma)}\right\}\right\rceil_+ \\ & + \widehat{C}_2\rho_0\zeta\max\left\{\frac{2\zeta(r_0+\theta)}{\varepsilon\rho_0}, \zeta\left(\frac{2\eta_0}{\varepsilon}\right)^{\frac{\log\zeta}{\log(1/\sigma)}}, 1\right\} \\ & + \widehat{C}_3\rho_0\zeta\left\lceil \max\left\{\log\left(\frac{2r_0+2\theta}{\varepsilon\rho_0}\right)/\log\zeta, \frac{\log(2\eta_0/\varepsilon)}{\log(1/\sigma)}\right\}\right\rceil_+ \max\left\{\frac{2\zeta(r_0+\theta)}{\varepsilon\rho_0}, \zeta\left(\frac{2\eta_0}{\varepsilon}\right)^{\frac{\log\zeta}{\log(1/\sigma)}}, 1\right\}.\end{aligned} \tag{43}$$

(iii) If $\mu > 0$, i.e., f is strongly convex, the total number of evaluations of ∇f, ∇g, proximal operator of P and projection onto $\mathcal{K}^*$ performed in Algorithm 5 is no more than $\check{N}$, respectively, where

$$\begin{aligned}\check{N} = {} & 1 + (M+1)\widehat{C}_1 + \left(1+(M+1)\widehat{C}_1\right)\left\lceil \max\left\{\log\left(\frac{2r_0+2\theta}{\varepsilon\rho_0}\right)/\log\zeta, \frac{\log(2\eta_0/\varepsilon)}{\log(1/\sigma)}\right\}\right\rceil_+ \\ & + \widehat{C}_2\sqrt{\frac{\rho_0\zeta}{\mu}}\max\left\{\sqrt{\frac{2\zeta(r_0+\theta)}{\varepsilon\rho_0}}, \sqrt{\zeta}\left(\frac{2\eta_0}{\varepsilon}\right)^{\frac{\log\zeta}{2\log(1/\sigma)}}, 1\right\} \\ & + \widehat{C}_3\sqrt{\frac{\rho_0\zeta}{\mu}}\left\lceil \max\left\{\log\left(\frac{2r_0+2\theta}{\varepsilon\rho_0}\right)/\log\zeta, \frac{\log(2\eta_0/\varepsilon)}{\log(1/\sigma)}\right\}\right\rceil_+ \max\left\{\sqrt{\frac{2\zeta(r_0+\theta)}{\varepsilon\rho_0}}, \sqrt{\zeta}\left(\frac{2\eta_0}{\varepsilon}\right)^{\frac{\log\zeta}{2\log(1/\sigma)}}, 1\right\}.\end{aligned} \tag{44}$$

Remark 8. *Since $1 < \zeta < 1/\sigma$, it can be seen from Theorem 5 that Algorithm 5 enjoys an operation complexity of $\mathcal{O}(\varepsilon^{-1}\log\varepsilon^{-1})$ and $\mathcal{O}(\varepsilon^{-1/2}\log\varepsilon^{-1})$ for finding an ε-KKT solution of problems* (2) *and* (29) *when f is convex and strongly convex on* $\mathrm{dom}(P)$, *respectively.*

4 Numerical results

In this section we conduct some preliminary experiments to test the performance of our proposed method (Algorithm 5), and compare it with a first-order proximal AL method (FPAL) [12], the forward-reflected-backward splitting method (FRBS) [13] and the modified forward-backward splitting method (MFBS) with an Armijo-Goldstein-type stepsize [19], respectively. All the algorithms are coded in Matlab and all the computations are performed on a desktop with a 3.60 GHz Intel i7-12700K 12-core processor and 32 GB of RAM.

4.1 Quadratically constrained quadratic programming with box constraints

In this subsection we consider quadratically constrained quadratic programming (QCQP) with box constraints

$$\begin{aligned} \min\ & \frac{1}{2}x^TAx + b^Tx \\ \text{s.t. } & \frac{1}{2}x^TB_ix + c_i^Tx + d_i \le 0, \quad i = 1,\dots,m, \\ & -1 \le x_i \le 1, \quad i = 1,\dots,n, \end{aligned} \tag{45}$$

where $A, B_1, \dots, B_m \in \mathbb{R}^{n\times n}$ are positive semidefinite matrices, $b, c_1, \dots, c_m \in \mathbb{R}^n$, and $d_1, \dots, d_m \in \mathbb{R}$.

For each dimension n, we set $m = \lceil 0.05n \rceil$ and randomly generate 10 instances of problem (45). In particular, we first generate $x^* \in [-1,1]^n$ whose entries are first independently chosen from the standard normal distribution and then projected to $[-1,1]$, and $\lambda^* \in \mathbb{R}^m_+$ whose entries are first independently chosen from the normal distribution with mean 1 and standard deviation 1 and then projected to $\mathbb{R}_+$. We then randomly generate an orthogonal matrix U by performing $U = \mathrm{orth}(\mathrm{randn}(n))$, an $n \times n$ diagonal matrix D whose diagonal entries are first independently chosen from the normal distribution with mean 0 and standard deviation 100 and then projected to $\mathbb{R}_+$, and set $A = UDU^T$. Also, we randomly generate an orthogonal matrix $\widetilde{U}$ by performing $\widetilde{U} = \mathrm{orth}(\mathrm{randn}(n))$, an $n \times n$ diagonal matrices $\widetilde{D}$ whose diagonal entries are first independently chosen from the normal distribution with mean 0 and standard deviation 0.01 and then projected to $\mathbb{R}_+$. We set $B_1 = \widetilde{U}\widetilde{D}\widetilde{U}^T$, and generate $B_i,\ i = 2,\dots,m$ in a similar vein. In addition, we generate $c_i,\ i = 1,\dots,m$ independently according the normal distribution with mean 0 and standard deviation 0.01. We finally choose b and $d_i,\ i = 1,\dots,m$ so that the KKT conditions of (45) are satisfied at (x^*, λ^*), namely (x^*, λ^*) is a KKT point of (45).

Notice that (45) is a special case of (2) with $f(x) = x^TAx/2 + b^Tx$, $P(x) = \mathcal{I}_{[-1,1]^n}(x)$, $g_i(x) = x^TB_ix/2 + c_i^Tx + d_i,\ i = 1,\dots,m$, and $\mathcal{K} = \mathbb{R}^m_+$, where $\mathcal{I}_{[-1,1]^n}(\cdot)$ is the indicator function of $[-1,1]^n$. Moreover, f and P are convex, g is $\mathcal{K}$-convex, $\mathrm{dom}(P)$ is compact, and ∇f and ∇g are (globally) Lipschitz continuous on $\mathrm{dom}(P)$. Consequently, (45) can be suitably solved by Algorithm 5 and FPAL [12]. It shall be mentioned that FPAL [12] is only applicable to (2) with $\mathrm{dom}(P)$ being compact. Our aim is to find a 10^{-2}-KKT solution of (45) by Algorithm 5 and FPAL, and compare their performance. Due to this, we terminate them once a 10^{-2}-KKT solution is found. Besides, for both methods, we choose zero vector as the initial point and set their parameters as follows.

- $(\varepsilon, M, \delta, \rho_0, \alpha_0, \eta_0, \zeta, \sigma) = (10^{-2}, 500, 0.9, 10, 1, 0.1, 2, 0.4)$ for Algorithm 5;
- $\epsilon = 10^{-2}$, $\rho_k = \rho_0\zeta^k$, $\eta_k = \eta_0\sigma^k$ with $(\rho_0, \eta_0, \zeta, \sigma) = (10, 0.1, 2, 0.4)$ for FPAL [12].

The computational results of Algorithm 5 and FPAL for the instances generated above are presented in Table 1. In detail, the value of n is listed in the first column. For each n, the average number of gradient evaluations and the average CPU time (in seconds) of Algorithm 5 and FPAL over 10 random instances are given in the rest of the columns. One can observe that our method, namely Algorithm 5, significantly outperforms FPAL in terms of average number of gradient evaluations and average CPU time. This phenomenon is not surprising because Algorithm 5 uses a local Lipschitz constant of the gradient of the smooth component of the AL functions, while FPAL uses its global Lipschitz constant that can be excessively conservative.

	Gradient evaluations		CPU time (seconds)	
n	Algorithm 5	FPAL	Algorithm 5	FPAL
100	4.97×10^3	3.96×10^3	0.20	0.21
200	4.57×10^3	5.37×10^3	6.35	12.07
300	4.47×10^3	6.23×10^3	12.53	27.88
400	4.18×10^3	8.00×10^3	33.12	93.04
500	4.18×10^3	9.75×10^3	22.65	248.28
600	4.18×10^3	1.32×10^4	58.65	617.72
700	4.08×10^3	1.22×10^4	122.52	889.71
800	4.08×10^3	1.57×10^4	186.23	1551.13
900	4.08×10^3	1.94×10^4	305.97	2737.96
1000	4.08×10^3	2.30×10^4	429.38	4398.44

Table 1: Numerical results for problem (45)

4.2 Quadratically constrained quadratic programming

In this subsection we consider the quadratically constrained quadratic programming (QCQP)

$$\begin{aligned} \min\ & \frac{1}{2}x^T Ax + b^T x \\ \text{s.t. } & \frac{1}{2}x^T B_i x + c_i^T x + d_i \le 0, \quad i = 1, \ldots, m, \end{aligned} \tag{46}$$

where $A, B_1, \ldots, B_m \in \mathbb{R}^{n\times n}$ are positive semidefinite matrices, $b, c_1, \ldots, c_m \in \mathbb{R}^n$, and $d_1, \ldots, d_m \in \mathbb{R}$.

For each dimension n, we set $m = \lceil 0.05n \rceil$ and randomly generate 10 instances of problem (46). In particular, we first generate $x^* \in \mathbb{R}^n$ with all the entries independently chosen from the standard normal distribution, and $\lambda^* \in \mathbb{R}^m_+$ whose entries are first independently chosen from the normal distribution with mean 1 and standard deviation 1 and then projected to $\mathbb{R}_+$. We then generate A and $B_i, c_i,\ i = 1, \ldots, m$ in the same manner as described in Subsection 4.1. We finally choose b and $d_i,\ i = 1, \ldots, m$ so that the KKT conditions of (46) are satisfied at (x^*, λ^*), namely (x^*, λ^*) is a KKT point of (46).

Notice that (46) is a special case of (2) with $f(x) = x^T Ax/2 + b^T x$, $P(x) = 0$, $g_i(x) = x^T B_i x/2 + c_i^T x + d_i,\ i = 1, \ldots, m$, and $\mathcal{K} = \mathbb{R}^m_+$. Clearly, f and P are convex, g is $\mathcal{K}$-convex, ∇f and ∇g are Lipschitz continuous, while $\mathrm{dom}(P) = \mathbb{R}^n$ is unbounded. As a result, (46) can be suitably solved by Algorithm 5 but not FPAL [12], since the latter method is only applicable to (2) with $\mathrm{dom}(P)$ being compact. On the other hand, it is not hard to observe that problem (46) and its dual can be solved as the monotone inclusion problem

$$0 \in F(x, \lambda) + B(x, \lambda), \tag{47}$$

where

$$F(x, \lambda) = \begin{pmatrix} \nabla f(x) + \nabla g(x)\lambda \\ -g(x) \end{pmatrix}, \quad B(x, \lambda) = \begin{pmatrix} 0 \\ \mathcal{N}_{\mathbb{R}^m_+}(\lambda) \end{pmatrix}.$$

One can also observe that F is monotone and locally Lipschitz continuous on $\mathrm{cl}(\mathrm{dom}B)$ and B is maximal monotone. As a result, problem (47) and hence (46) can be suitably solved by FRBS [13] and MFBS [19]. Our aim is to find a 10^{-2}-KKT solution of (46) by Algorithm 5, FRBS and MFBS, and compare their performance. Due to this, we terminate them once a 10^{-2}-KKT solution is found. In addition, for all the methods, we choose zero vector as the initial point and set their parameters as follows.

- $(\varepsilon, M, \delta, \rho_0, \alpha_0, \eta_0, \zeta, \sigma) = (10^{-2}, 500, 0.9, 10, 1, 0.1, 2, 0.4)$ for Algorithm 5;
- $(\lambda_0, \delta, \sigma) = (0.1, 0.5, 0.9)$ for FRBS [13];
- $(\sigma, \theta, \beta) = (0.1, 0.5, 0.9)$ for MFBS [19].

The computational results of Algorithm 5, FRBS and MFBS for the instances generated above are presented in Table 2. In detail, the value of n is listed in the first column. For each n, the average number of gradient evaluations and the average CPU time (in seconds) for these methods over 10 random instances are given in the rest of the columns. One can observe that our method, namely Algorithm 5, significantly outperforms the other

two methods in terms of average number of gradient evaluations and average CPU time. This phenomenon may not be surprising because our method enjoys a nearly optimal operation complexity while the other two methods lack complexity guarantees.

	Gradient evaluations			CPU time (seconds)		
n	Algorithm 5	FRBS	MFBS	Algorithm 5	FRBS	MFBS
100	5.02×10^3	1.89×10^5	1.62×10^5	0.17	1.80	1.40
200	5.01×10^3	1.38×10^5	1.36×10^5	7.16	80.49	77.47
300	4.68×10^3	1.11×10^5	1.02×10^5	12.10	147.19	132.95
400	4.28×10^3	9.76×10^4	8.33×10^4	32.52	387.91	323.89
500	4.08×10^3	7.12×10^4	6.16×10^4	23.18	147.25	125.38
600	4.18×10^3	7.37×10^4	6.07×10^4	53.51	427.64	346.31
700	4.08×10^3	6.59×10^4	5.33×10^4	101.14	637.33	518.73
800	4.08×10^3	5.69×10^4	4.72×10^4	152.60	782.77	629.80
900	4.09×10^3	5.46×10^4	4.54×10^4	236.99	1359.02	1138.14
1000	4.08×10^3	4.63×10^4	3.92×10^4	384.96	1854.11	1568.45

Table 2: Numerical results for problem (46)

5 Proof of the main results

In this section we provide a proof of our main results presented in Sections 2 and 3, which are particularly Theorems 1-5.

5.1 Proof of the main results in Subsection 2.1

In this subsection we first establish several technical lemmas and then use them to prove Theorems 1 and 2.

Lemma 2. *Suppose that α_t, β_t and γ_t are generated by Algorithm 1 for some $t \geq 1$. Then the following statements hold.*

(i) $\sqrt{\mu\gamma_t} \leq \alpha_t \leq 1$ *and* $\alpha_t^2\gamma_t^{-1} \leq \alpha_{t-1}^2\gamma_{t-1}^{-1}$.

(ii) $\beta_t = \mu\gamma_t\alpha_t^{-1} \in [0, 1]$.

Proof. (i) We first prove by induction that $\sqrt{\mu\gamma_i} \leq \alpha_i \leq 1$ for all $1 \leq i \leq t$. Indeed, notice from Algorithm 1 that $\sqrt{\mu\gamma_0} \leq \alpha_0 \leq 1$. Suppose that $\sqrt{\mu\gamma_{i-1}} \leq \alpha_{i-1} \leq 1$ for some $1 \leq i < t$. By this, (7), and $\alpha_i \in (0, 1]$, one has

$$\gamma_{i-1}\alpha_i^2 = (1-\alpha_i)\alpha_{i-1}^2\gamma_i + \mu\alpha_i\gamma_i\gamma_{i-1} \geq (1-\alpha_i)\mu\gamma_{i-1}\gamma_i + \mu\alpha_i\gamma_i\gamma_{i-1} = \mu\gamma_i\gamma_{i-1},$$

which together with $\gamma_{i-1} > 0$ yields $\sqrt{\mu\gamma_i} \leq \alpha_i \leq 1$. Hence, the induction is completed and $\sqrt{\mu\gamma_t} \leq \alpha_t \leq 1$ holds as desired.

We next show that $\alpha_t^2\gamma_t^{-1} \leq \alpha_{t-1}^2\gamma_{t-1}^{-1}$. Indeed, by $\sqrt{\mu\gamma_{t-1}} \leq \alpha_{t-1}$, $\gamma_{t-1}, \gamma_t > 0$, and (7), one has

$$\gamma_{t-1}\alpha_t^2 = (1-\alpha_t)\alpha_{t-1}^2\gamma_t + \mu\alpha_t\gamma_t\gamma_{t-1} \leq (1-\alpha_t)\alpha_{t-1}^2\gamma_t + \alpha_t\gamma_t\alpha_{t-1}^2 = \gamma_t\alpha_{t-1}^2,$$

which implies that the conclusion holds.

(ii) Notice from Algorithm 1 that $\beta_t = \mu\gamma_t\alpha_t^{-1}$. By this and statement (i), one has

$$0 \leq \beta_t = \mu\gamma_t\alpha_t^{-1} \leq \sqrt{\mu\gamma_t} \leq 1.$$

□

Lemma 3. *Suppose that x^{t+1}, y^t and z^{t+1} are generated by Algorithm 1 for some $t \geq 1$. Then for all $x \in \mathrm{dom}(P)$ and $P'(z^{t+1}) \in \partial P(z^{t+1})$, we have*

$$\gamma_t\langle P'(z^{t+1}), z^{t+1} - x\rangle \leq \gamma_t\langle \nabla f(y^t), x - z^{t+1}\rangle + \frac{1}{2}\alpha_t\beta_t\|x - y^t\|^2 + \frac{1}{2}\alpha_t(1-\beta_t)\|x - z^t\|^2 - \frac{1}{2}\alpha_t\|x - z^{t+1}\|^2 + R_t, \quad (48)$$

where

$$R_t = \frac{1}{2}\mu\gamma_t(\alpha_t^{-1} - 1)\|x^t - y^t\|^2 - \frac{1}{2\alpha_t}\|x^{t+1} - y^t\|^2. \quad (49)$$

Proof. By the optimality condition of (5), one has

$$\langle \gamma_t \nabla f(y^t) + \gamma_t P'(z^{t+1}) + \alpha_t(z^{t+1} - \beta_t y^t - (1-\beta_t)z^t), x - z^{t+1}\rangle \geq 0$$

for all $x \in \mathrm{dom}(P)$ and $P'(z^{t+1}) \in \partial P(z^{t+1})$. It follows from this relation that

$$\begin{aligned}
\gamma_t\langle P'(z^{t+1}), z^{t+1} - x\rangle &\leq \gamma_t\langle \nabla f(y^t), x - z^{t+1}\rangle + \alpha_t\langle z^{t+1} - \beta_t y^t - (1-\beta_t)z^t, x - z^{t+1}\rangle \\
&= \gamma_t\langle \nabla f(y^t), x - z^{t+1}\rangle + \alpha_t\beta_t\langle z^{t+1} - y^t, x - z^{t+1}\rangle + \alpha_t(1-\beta_t)\langle z^{t+1} - z^t, x - z^{t+1}\rangle \\
&= \gamma_t\langle \nabla f(y^t), x - z^{t+1}\rangle + \frac{1}{2}\alpha_t\beta_t\left(\|x - y^t\|^2 - \|x - z^{t+1}\|^2 - \|y^t - z^{t+1}\|^2\right) \\
&\quad + \frac{1}{2}\alpha_t(1-\beta_t)\left(\|x - z^t\|^2 - \|x - z^{t+1}\|^2 - \|z^t - z^{t+1}\|^2\right) \\
&= \gamma_t\langle \nabla f(y^t), x - z^{t+1}\rangle + \frac{1}{2}\alpha_t\beta_t\|x - y^t\|^2 + \frac{1}{2}\alpha_t(1-\beta_t)\|x - z^t\|^2 \\
&\quad - \frac{1}{2}\alpha_t\|x - z^{t+1}\|^2 + Q_t,
\end{aligned} \tag{50}$$

where

$$Q_t = -\frac{1}{2}\alpha_t\beta_t\|y^t - z^{t+1}\|^2 - \frac{1}{2}\alpha_t(1-\beta_t)\|z^t - z^{t+1}\|^2. \tag{51}$$

We next show that $Q_t \leq R_t$. Indeed, it follows from (4) that

$$x^t - y^t = \alpha_t(1-\alpha_t)^{-1}(1-\beta_t)(y^t - z^t), \tag{52}$$

which together with (6) implies that

$$\begin{aligned}
x^{t+1} - y^t &= (1-\alpha_t)x^t + \alpha_t z^{t+1} - y^t = (1-\alpha_t)(x^t - y^t) + \alpha_t z^{t+1} - \alpha_t y^t \\
&\overset{(52)}{=} \alpha_t(1-\beta_t)(y^t - z^t) + \alpha_t z^{t+1} - \alpha_t y^t = \alpha_t\left(z^{t+1} - \beta_t y^t - (1-\beta_t)z^t\right).
\end{aligned} \tag{53}$$

Using this relation, $\beta_t \in [0,1]$, and the convexity of $\|\cdot\|^2$, we obtain

$$\alpha_t^{-2}\|x^{t+1} - y^t\|^2 \overset{(53)}{=} \|z^{t+1} - \beta_t y^t - (1-\beta_t)z^t\|^2 \leq \beta_t\|z^{t+1} - y^t\|^2 + (1-\beta_t)\|z^{t+1} - z^t\|^2.$$

By this, (49), (51), and $\alpha_t \in (0,1]$, one has

$$\begin{aligned}
2\alpha_t^{-1}(Q_t - R_t) =& -\beta_t\|y^t - z^{t+1}\|^2 - (1-\beta_t)\|z^t - z^{t+1}\|^2 + \alpha_t^{-2}\|x^{t+1} - y^t\|^2 - \mu\gamma_t\alpha_t^{-2}(1-\alpha_t)\|x^t - y^t\|^2 \\
\leq& -\beta_t\|y^t - z^{t+1}\|^2 - (1-\beta_t)\|z^t - z^{t+1}\|^2 + \alpha_t^{-2}\|x^{t+1} - y^t\|^2 \leq 0,
\end{aligned}$$

which along with $\alpha_t > 0$ implies that $Q_t \leq R_t$.

The conclusion of this Lemma directly follows from (50) and $Q_t \leq R_t$. □

Lemma 4. *Suppose that x^{t+1}, y^t and z^{t+1} are generated by Algorithm 1 for some $t \geq 1$. Then for any $x \in \mathrm{dom}(P)$, we have*

$$F(x^{t+1}) - F(x) + \frac{\alpha_t^2}{2\gamma_t}\|x - z^{t+1}\|^2 \leq \prod_{i=1}^{t}(1-\alpha_i)\left(F(x^1) - F(x) + \frac{\alpha_0^2}{2\gamma_0}\|x - x^1\|^2\right). \tag{54}$$

Proof. By (6), (48), and the convexity of P, one has that for all $P'(z^{t+1}) \in \partial P(z^{t+1})$,

$$\begin{aligned}
\gamma_t\alpha_t^{-1}P(x^{t+1}) &\leq \gamma_t\alpha_t^{-1}\left((1-\alpha_t)P(x^t) + \alpha_t P(z^{t+1})\right) = \gamma_t(\alpha_t^{-1} - 1)P(x^t) + \gamma_t P(z^{t+1}) \\
&\leq \gamma_t(\alpha_t^{-1} - 1)P(x^t) + \gamma_t P(x) + \gamma_t\langle P'(z^{t+1}), z^{t+1} - x\rangle \\
&\overset{(48)}{\leq} \gamma_t(\alpha_t^{-1} - 1)P(x^t) + \gamma_t P(x) + \gamma_t\langle \nabla f(y^t), x - z^{t+1}\rangle \\
&\quad + \frac{1}{2}\alpha_t(1-\beta_t)\|x - z^t\|^2 - \frac{1}{2}\alpha_t\|x - z^{t+1}\|^2 + \frac{1}{2}\alpha_t\beta_t\|x - y^t\|^2 + R_t.
\end{aligned} \tag{55}$$

By (3), (6), $\alpha_t \in (0,1]$, and $\gamma_t > 0$, one has that for all $x \in \mathrm{dom}(P)$,

$$\begin{aligned}
&\gamma_t\alpha_t^{-1}f(y^t) + \gamma_t\alpha_t^{-1}\langle\nabla f(y^t), x^{t+1} - y^t\rangle + \gamma_t\langle\nabla f(y^t), x - z^{t+1}\rangle \\
&\overset{(6)}{=} \gamma_t\alpha_t^{-1}f(y^t) + \gamma_t\alpha_t^{-1}\langle\nabla f(y^t), (1-\alpha_t)x^t + \alpha_t z^{t+1} - y^t\rangle + \gamma_t\langle\nabla f(y^t), x - z^{t+1}\rangle \\
&= \gamma_t\alpha_t^{-1}f(y^t) + \gamma_t(\alpha_t^{-1}-1)\langle\nabla f(y^t), x^t - y^t\rangle + \gamma_t\langle\nabla f(y^t), x - y^t\rangle \\
&= \gamma_t(\alpha_t^{-1}-1)\left(f(y^t) + \langle\nabla f(y^t), x^t - y^t\rangle\right) + \gamma_t\left(f(y^t) + \langle\nabla f(y^t), x - y^t\rangle\right) \\
&\overset{(3)}{\le} \gamma_t(\alpha_t^{-1}-1)\left(f(x^t) - \frac{1}{2}\mu\|x^t - y^t\|^2\right) + \gamma_t\left(f(x) - \frac{1}{2}\mu\|x - y^t\|^2\right) \\
&= \gamma_t(\alpha_t^{-1}-1)f(x^t) + \gamma_t f(x) - \frac{1}{2}\mu\gamma_t(\alpha_t^{-1}-1)\|x^t - y^t\|^2 - \frac{1}{2}\mu\gamma_t\|x - y^t\|^2. \qquad (56)
\end{aligned}$$

Using (8), (55) and (56), we have

$$\begin{aligned}
\gamma_t\alpha_t^{-1}F(x^{t+1}) &\overset{(8)}{\le} \gamma_t\alpha_t^{-1}f(y^t) + \gamma_t\alpha_t^{-1}\langle\nabla f(y^t), x^{t+1} - y^t\rangle + \frac{1}{2\alpha_t}\|x^{t+1} - y^t\|^2 + \gamma_t\alpha_t^{-1}P(x^{t+1}) \\
&\overset{(55)}{\le} \gamma_t\alpha_t^{-1}f(y^t) + \gamma_t\alpha_t^{-1}\langle\nabla f(y^t), x^{t+1} - y^t\rangle + \gamma_t\langle\nabla f(y^t), x - z^{t+1}\rangle + \frac{1}{2\alpha_t}\|x^{t+1} - y^t\|^2 \\
&\quad + \gamma_t(\alpha_t^{-1}-1)P(x^t) + \gamma_t P(x) + \frac{1}{2}\alpha_t(1-\beta_t)\|x - z^t\|^2 - \frac{1}{2}\alpha_t\|x - z^{t+1}\|^2 \\
&\quad + \frac{1}{2}\alpha_t\beta_t\|x - y^t\|^2 + R_t \\
&\overset{(56)}{\le} \gamma_t(\alpha_t^{-1}-1)F(x^t) + \gamma_t F(x) + \frac{1}{2}(\alpha_t\beta_t - \mu\gamma_t)\|x - y^t\|^2 - \frac{1}{2}\mu\gamma_t(\alpha_t^{-1}-1)\|x^t - y^t\|^2 \\
&\quad + \frac{1}{2\alpha_t}\|x^{t+1} - y^t\|^2 + \frac{1}{2}\alpha_t(1-\beta_t)\|x - z^t\|^2 - \frac{1}{2}\alpha_t\|x - z^{t+1}\|^2 + R_t \\
&= \gamma_t(\alpha_t^{-1}-1)F(x^t) + \gamma_t F(x) + \frac{1}{2}\alpha_t(1-\beta_t)\|x - z^t\|^2 - \frac{1}{2}\alpha_t\|x - z^{t+1}\|^2, \qquad (57)
\end{aligned}$$

where the equality follows from (49) and $\beta_t = \mu\gamma_t\alpha_t^{-1}$. In addition, it follows from (7) and $\beta_t = \mu\gamma_t\alpha_t^{-1}$ that

$$\gamma_{t-1}\alpha_t^2(1-\beta_t) = \gamma_{t-1}\alpha_t^2 - \gamma_{t-1}\alpha_t^2\beta_t = \gamma_{t-1}\alpha_t^2 - \mu\alpha_t\gamma_t\gamma_{t-1} \overset{(7)}{=} (1-\alpha_t)\alpha_{t-1}^2\gamma_t. \qquad (58)$$

In view of (57) and (58), one has

$$\begin{aligned}
F(x^{t+1}) - F(x) + \frac{\alpha_t^2}{2\gamma_t}\|x - z^{t+1}\|^2 &\overset{(57)}{\le} (1-\alpha_t)\left(F(x^t) - F(x)\right) + \frac{\alpha_t^2(1-\beta_t)}{2\gamma_t}\|x - z^t\|^2 \\
&\overset{(58)}{=} (1-\alpha_t)\left(F(x^t) - F(x) + \frac{\alpha_{t-1}^2}{2\gamma_{t-1}}\|x - z^t\|^2\right).
\end{aligned}$$

The conclusion of this lemma immediately follows from the above inequality and $z^1 = x^1$. □

Suppose that x^t and z^t are generated by Algorithm 1 for some $t \ge 1$. For any $0 < \gamma \le \gamma_0$, we define

$$y^t(\gamma) = \left((1-\alpha(\gamma))x^t + \alpha(\gamma)(1-\beta(\gamma))z^t\right)/\left(1-\alpha(\gamma)\beta(\gamma)\right), \qquad (59)$$

$$z^{t+1}(\gamma) = \arg\min_x \left\{\gamma\langle\nabla f(y^t(\gamma)), x\rangle + \gamma P(x) + \frac{\alpha(\gamma)}{2}\|x - \beta(\gamma)y^t(\gamma) - (1-\beta(\gamma))z^t\|^2\right\}, \qquad (60)$$

$$x^{t+1}(\gamma) = (1-\alpha(\gamma))x^t + \alpha(\gamma)z^{t+1}(\gamma), \qquad (61)$$

where $\beta(\gamma) = \mu\gamma\alpha(\gamma)^{-1}$ and $\alpha(\gamma) \in (0,1]$ satisfies

$$\gamma_{t-1}\alpha(\gamma)^2 = (1-\alpha(\gamma))\alpha_{t-1}^2\gamma + \mu\gamma\gamma_{t-1}\alpha(\gamma). \qquad (62)$$

Lemma 5. *Let $\mathcal{S}$ and $\widehat{\mathcal{S}}$ be defined in (9) and (10). Suppose that $x^t, z^t \in \mathcal{S}$, and $y^t(\gamma)$ and $x^{t+1}(\gamma)$ are defined in (59) and (61) for some $t \ge 1$. Then $y^t(\gamma) \in \mathcal{S}$ and $x^{t+1}(\gamma) \in \widehat{\mathcal{S}}$ for all $0 < \gamma \le \gamma_0$.*

Proof. Fix any $0 < \gamma \le \gamma_0$. By the optimality condition of problems (1) and (60), one has

$$\langle \gamma \nabla f(y^t(\gamma)) + \gamma P'(z^{t+1}(\gamma)) + \alpha(\gamma)(z^{t+1}(\gamma) - \beta(\gamma) y^t(\gamma) - (1-\beta(\gamma)) z^t), x^* - z^{t+1}(\gamma) \rangle \ge 0,$$
$$\langle \gamma \nabla f(x^*) + \gamma P'(x^*), z^{t+1}(\gamma) - x^* \rangle \ge 0,$$

where $P'(z^{t+1}(\gamma)) \in \partial P(z^{t+1}(\gamma))$ and $P'(x^*) \in \partial P(x^*)$. Letting $w = \beta(\gamma) y^t(\gamma) + (1-\beta(\gamma)) z^t$ and using the above two inequalities and the convexity of P, we obtain

$$\langle \alpha(\gamma)(z^{t+1}(\gamma) - w) + \gamma(\nabla f(y^t(\gamma)) - \nabla f(x^*)), x^* - z^{t+1}(\gamma) \rangle \ge \gamma \langle P'(z^{t+1}(\gamma)) - P'(x^*), z^{t+1}(\gamma) - x^* \rangle \ge 0,$$

which yields

$$\begin{aligned} \alpha(\gamma) \|z^{t+1}(\gamma) - x^*\|^2 &\le \langle \alpha(\gamma)(x^* - w) + \gamma(\nabla f(y^t(\gamma)) - \nabla f(x^*)), x^* - z^{t+1}(\gamma) \rangle \\ &\le \|\alpha(\gamma)(x^* - w) + \gamma(\nabla f(y^t(\gamma)) - \nabla f(x^*))\| \|z^{t+1}(\gamma) - x^*\|. \end{aligned} \tag{63}$$

In addition, recall from Lemma 2 that $\sqrt{\mu \gamma_{t-1}} \le \alpha_{t-1} \le 1$. By this, $\alpha(\gamma) \in (0,1]$, (62), and a similar argument as in the proof of Lemma 2(ii), one can see that $\beta(\gamma) \in [0,1]$. It then follows from this, (59), the expression of w, and $x^t, z^t \in \mathcal{S}$ that $y^t(\gamma), w \in \mathcal{S}$. By these, $\alpha(\gamma) > 0$, (9), (63), and Lemma 1, one has

$$\begin{aligned} \alpha(\gamma)\|z^{t+1}(\gamma) - x^*\| &\overset{(63)}{\le} \|\alpha(\gamma)(w - x^*) + \gamma(\nabla f(y^t(\gamma)) - \nabla f(x^*))\| \le \alpha(\gamma)\|w - x^*\| + \gamma \|\nabla f(y^t(\gamma)) - \nabla f(x^*)\| \\ &\le \alpha(\gamma)\|w - x^*\| + \gamma L_{\mathcal{S}} \|y^t(\gamma) - x^*\| \overset{(9)}{\le} (\alpha(\gamma) + \gamma L_{\mathcal{S}}) \frac{\sqrt{2\gamma_0} r_0}{\alpha_0}. \end{aligned}$$

Using this, (9), (61), $\alpha(\gamma) \in (0,1]$, $x^t \in \mathcal{S}$, and $\gamma \le \gamma_0$, we obtain that

$$\begin{aligned} \|x^{t+1}(\gamma) - x^*\| &\overset{(61)}{\le} (1 - \alpha(\gamma))\|x^t - x^*\| + \alpha(\gamma)\|z^{t+1}(\gamma) - x^*\| \\ &\overset{(9)}{\le} (1 - \alpha(\gamma)) \frac{\sqrt{2\gamma_0} r_0}{\alpha_0} + (\alpha(\gamma) + \gamma L_{\mathcal{S}}) \frac{\sqrt{2\gamma_0} r_0}{\alpha_0} \\ &\le (1 + \gamma_0 L_{\mathcal{S}}) \frac{\sqrt{2\gamma_0} r_0}{\alpha_0}. \end{aligned}$$

It then follows from the last relation and (10) that $x^{t+1}(\gamma) \in \widehat{\mathcal{S}}$. □

For the convenience of our subsequent analysis, we define

$$\lambda_0 = 1, \quad \lambda_t = \prod_{i=1}^{t} (1 - \alpha_i). \tag{64}$$

Lemma 6. *Let $\mathcal{S}$ and N be defined in (9) and (11). Suppose that $x^t, z^t \in \mathcal{S}$ for some $t \ge 1$. Then x^{t+1}, y^t and z^{t+1} are successfully generated by Algorithm 1 at iteration t with $n_t \le N$, and moreover, $x^{t+1}, y^t, z^{t+1} \in \mathcal{S}$.*

Proof. Let $\gamma = \gamma_0 \delta^N$ and $y^t(\gamma)$ and $x^{t+1}(\gamma)$ be defined in (59) and (61). By $\delta \in (0,1)$ and (11), one can observe that $0 < \gamma \le \gamma_0$ and $\gamma \le L_{\widehat{\mathcal{S}}}^{-1}$. Using these, $x^t, z^t \in \mathcal{S}$, and Lemma 5, we see that $x^{t+1}(\gamma) \in \widehat{\mathcal{S}}$ and $y^t(\gamma) \in \mathcal{S} \subseteq \widehat{\mathcal{S}}$, where $\widehat{\mathcal{S}}$ is defined in (10). It then follows from $\gamma \le L_{\widehat{\mathcal{S}}}^{-1}$ and Lemma 1(ii) that

$$2\gamma \left(f(x^{t+1}(\gamma)) - f(y^t(\gamma)) - \langle \nabla f(y^t(\gamma)), x^{t+1}(\gamma) - y^t(\gamma) \rangle \right) \le \gamma L_{\widehat{\mathcal{S}}} \|x^{t+1}(\gamma) - y^t(\gamma)\|^2 \le \|x^{t+1}(\gamma) - y^t(\gamma)\|^2.$$

This together with the definition of n_t in Algorithm 1 implies that $n_t \le N$. It then follows that x^{t+1}, y^t and z^{t+1} are successfully generated by Algorithm 1.

Since $x^t, z^t \in \mathcal{S}$ and $y^t = y^t(\gamma_t)$ for some $0 < \gamma_t \le \gamma_0$, it follows from Lemma 5 that $y^t \in \mathcal{S}$. We next show that $x^{t+1}, z^{t+1} \in \mathcal{S}$. Indeed, by (7) and (64), one has

$$\lambda_t \overset{(64)}{=} (1 - \alpha_t)\lambda_{t-1} \overset{(7)}{=} \frac{\gamma_{t-1}\alpha_t^2 - \mu \alpha_t \gamma_t \gamma_{t-1}}{\alpha_{t-1}^2 \gamma_t} \lambda_{t-1} \le \frac{\gamma_{t-1}\alpha_t^2}{\alpha_{t-1}^2 \gamma_t} \lambda_{t-1},$$

which along with $\lambda_0 = 1$ implies that $\gamma_t\lambda_t/\alpha_t^2 \le \gamma_0/\alpha_0^2$. Using this, (54) and (64), we obtain that

$$\begin{aligned}\|z^{t+1} - x^*\|^2 &\le \frac{2\gamma_t}{\alpha_t^2}\left(F(x^{t+1}) - F(x^*) + \frac{\alpha_t^2}{2\gamma_t}\|z^{t+1} - x^*\|^2\right)\\ &\le \frac{2\gamma_t\lambda_t}{\alpha_t^2}\left(F(x^1) - F(x^*) + \frac{\alpha_0^2}{2\gamma_0}\|z^1 - x^*\|^2\right)\\ &\le \frac{2\gamma_0}{\alpha_0^2}\left(F(x^1) - F(x^*) + \frac{\alpha_0^2}{2\gamma_0}\|z^1 - x^*\|^2\right),\end{aligned}$$

which together with (9) implies that $z^{t+1} \in \mathcal{S}$. It then follows from this and (6) that $x^{t+1} \in \mathcal{S}$. □

We are now ready to prove Theorems 1 and 2.

Proof of Theorem 1. We prove this theorem by induction. Indeed, notice from Algorithm 1 that $z^1 = x^1 \in \mathcal{S}$. It then follows from Lemma 6 that x^2, y^1 and z^2 are successfully generated with $n_1 \le N$ and $x^2, y^1, z^2 \in \mathcal{S}$. Now, suppose that x^t, y^{t-1} and z^t are already generated with $n_{t-1} \le N$ and $x^t, y^{t-1}, z^t \in \mathcal{S}$. It then follows from Lemma 6 that x^{t+1}, y^t and z^{t+1} are successfully generated with $n_t \le N$ and $x^{t+1}, y^t, z^{t+1} \in \mathcal{S}$. Hence, the induction is complete and the conclusion of this theorem holds. □

Proof of Theorem 2. Observe from (64) that $\lambda_i = (1-\alpha_i)\lambda_{i-1} < \lambda_{i-1}$ for all $i \ge 1$. In addition, recall from the proof of Lemma 6 that $\gamma_i\lambda_i/\alpha_i^2 \le \gamma_0/\alpha_0^2$ for all $i \ge 1$. By these relations, one has

$$\frac{1}{\sqrt{\lambda_i}} - \frac{1}{\sqrt{\lambda_{i-1}}} = \frac{\lambda_{i-1} - \lambda_i}{\sqrt{\lambda_{i-1}\lambda_i}(\sqrt{\lambda_{i-1}} + \sqrt{\lambda_i})} \ge \frac{\lambda_{i-1} - \lambda_i}{2\lambda_{i-1}\sqrt{\lambda_i}} = \frac{\alpha_i}{2\sqrt{\lambda_i}} \ge \frac{1}{2}\alpha_0\sqrt{\gamma_i/\gamma_0} \qquad \forall i \ge 1.$$

Summing up the above inequalities for $i = 1, 2, \ldots, t$ and using $\lambda_0 = 1$, we obtain

$$\frac{1}{\sqrt{\lambda_t}} - 1 \ge \frac{1}{2}\alpha_0\sum_{i=1}^t\sqrt{\gamma_i/\gamma_0} \quad \Rightarrow \quad \lambda_t \le 4\left(2 + \alpha_0\sum_{i=1}^t\sqrt{\gamma_i/\gamma_0}\right)^{-2}. \tag{65}$$

Also, observe from (64) and Lemma 2(i) that

$$\lambda_t = \prod_{i=1}^t(1-\alpha_i) \le \prod_{i=1}^t(1 - \sqrt{\mu\gamma_i}). \tag{66}$$

In addition, recall from Theorem 1 that $n_i \le N$, which together with (11) implies that $\gamma_i = \gamma_0\delta^{n_i} \ge \min\{\gamma_0, \delta/L_{\widehat{\mathcal{S}}}\}$ for all $i \ge 1$. By this, (65) and (66), one has

$$\lambda_t \le \min\left\{\left(1 - \sqrt{\mu\min\left\{\gamma_0, \delta L_{\widehat{\mathcal{S}}}^{-1}\right\}}\right)^t,\ 4\left(2 + t\alpha_0\sqrt{\min\left\{1, \delta\gamma_0^{-1}L_{\widehat{\mathcal{S}}}^{-1}\right\}}\right)^{-2}\right\} \qquad \forall t \ge 1.$$

The conclusion of Theorem 2 then directly follows from this relation, (64) and (54) with $x = x^*$. □

5.2 Proof of the main results in Subsection 2.2

In this subsection we first establish two technical lemmas and then use them to prove Theorem 3.

Lemma 7. *Let γ_0, δ be given in Algorithm 2 and N be defined in* (11). *Suppose that (v, γ_0, δ) is the input for Algorithm 3 for any $v \in \mathcal{S}$. Then $(\tilde{v}, \tilde{\gamma})$ is successfully generated by Algorithm 3 with $\tilde{n} \le N$, $\tilde{v} \in \widehat{\mathcal{S}}$ and $\tilde{\gamma} \ge \min\{\gamma_0, \delta/L_{\widehat{\mathcal{S}}}\}$.*

Proof. For any $0 < \gamma \le \gamma_0$, let

$$\tilde{v}(\gamma) = \arg\min_x\left\{\gamma\langle\nabla f(v), x - v\rangle + \gamma P(x) + \frac{1}{2}\|x - v\|^2\right\}. \tag{67}$$

By the optimality condition of (1) and (67) and a similar argument as for (63), one has

$$\|\tilde{v}(\gamma) - x^*\| \le \|v - x^* - \gamma(\nabla f(v) - \nabla f(x^*))\|.$$

Using this, $v \in \mathcal{S}$, (9), and Lemma 1(i), we obtain

$$\|\tilde{v}(\gamma) - x^*\| \le \|v - x^*\| + \gamma L_{\mathcal{S}}\|v - x^*\| \le (1 + \gamma_0 L_{\mathcal{S}})\,\frac{\sqrt{2\gamma_0} r_0}{\alpha_0} \qquad \forall 0 < \gamma \le \gamma_0.$$

This along with the definition of $\widehat{\mathcal{S}}$ in (10) implies that $\tilde{v}(\gamma) \in \widehat{\mathcal{S}}$ for all $0 < \gamma \le \gamma_0$. Now, let $\gamma = \gamma_0\delta^N$. By $\delta \in (0,1)$ and (11), one can observe that $0 < \gamma \le \gamma_0$ and $\gamma \le L_{\widehat{\mathcal{S}}}^{-1}$. It then follows that $\tilde{v}(\gamma) \in \widehat{\mathcal{S}}$. By these, $v \in \mathcal{S} \subseteq \widehat{\mathcal{S}}$ and Lemma 1(ii), one has

$$2\gamma(f(\tilde{v}(\gamma)) - f(v) - \langle \nabla f(v), \tilde{v}(\gamma) - v\rangle) \le \gamma L_{\widehat{\mathcal{S}}}\|\tilde{v}(\gamma) - v\|^2 \le \|\tilde{v}(\gamma) - v\|^2.$$

These together with (11) and the definition of $\tilde{n}$ in Algorithm 3 implies that $(\tilde{v}, \tilde{\gamma})$ is successfully generated by Algorithm 3 with $\tilde{n} \le N$, and moreover,

$$\gamma_0 \ge \tilde{\gamma} = \gamma_0\delta^{\tilde{n}} \ge \gamma_0\delta^N \ge \min\{\gamma_0, \delta/L_{\widehat{\mathcal{S}}}\}, \quad \tilde{v} = \tilde{v}(\tilde{\gamma}) \in \widehat{\mathcal{S}}.$$

□

Lemma 8. *Suppose that x^{t+1} and $(\tilde{x}^{t+1}, \tilde{\gamma}_{t+1})$ are generated in Algorithm 2 for some $t \ge 1$. Then we have*

$$\begin{aligned}\mathrm{dist}(0, \partial F(\tilde{x}^{t+1})) &\le \|\tilde{\gamma}_{t+1}^{-1}(x^{t+1} - \tilde{x}^{t+1}) + \nabla f(\tilde{x}^{t+1}) - \nabla f(x^{t+1})\| \\ &\le \left(\sqrt{2\max\{\gamma_0^{-1}, L_{\widehat{\mathcal{S}}}\delta^{-1}\}} + \sqrt{2\gamma_0}L_{\widehat{\mathcal{S}}}\right)\sqrt{F(x^{t+1}) - F(x^*)}, \end{aligned} \tag{68}$$

where $L_{\widehat{\mathcal{S}}}$ is given in Lemma 1, and γ_0 and δ are the input parameters of Algorithm 1.

Proof. Notice that $(\tilde{x}^{t+1}, \tilde{\gamma}_{t+1})$ is the output of Algorithm 3 with $(x^{t+1}, \gamma_0, \delta)$ as the input. By Lemma 7, one has that $\tilde{x}^{t+1} \in \widehat{\mathcal{S}}$ and $\gamma_0 \ge \tilde{\gamma}_{t+1} \ge \min\{\gamma_0, \delta/L_{\widehat{\mathcal{S}}}\}$. Also, it follows from (14) and (15) with $v = x^{t+1}$, $\tilde{v} = \tilde{x}^{t+1}$ and $\tilde{\gamma} = \tilde{\gamma}_{t+1}$ that

$$\tilde{x}^{t+1} = \arg\min_x \left\{\tilde{\gamma}_{t+1}\langle \nabla f(x^{t+1}), x\rangle + \tilde{\gamma}_{t+1}P(x) + \frac{1}{2}\|x - x^{t+1}\|^2\right\}, \tag{69}$$

$$2\tilde{\gamma}_{t+1}(f(\tilde{x}^{t+1}) - f(x^{t+1}) - \langle \nabla f(x^{t+1}), \tilde{x}^{t+1} - x^{t+1}\rangle) \le \|\tilde{x}^{t+1} - x^{t+1}\|^2. \tag{70}$$

By the optimality condition of (69), it can be easily shown that

$$\tilde{\gamma}_{t+1}^{-1}(x^{t+1} - \tilde{x}^{t+1}) + \nabla f(\tilde{x}^{t+1}) - \nabla f(x^{t+1}) \in \partial F(\tilde{x}^{t+1}), \tag{71}$$

$$\tilde{\gamma}_{t+1}\langle \nabla f(x^{t+1}), \tilde{x}^{t+1}\rangle + \tilde{\gamma}_{t+1}P(\tilde{x}^{t+1}) \le \tilde{\gamma}_{t+1}\langle \nabla f(x^{t+1}), x^{t+1}\rangle + \tilde{\gamma}_{t+1}P(x^{t+1}) - \|\tilde{x}^{t+1} - x^{t+1}\|^2. \tag{72}$$

By (70) and (72), one has

$$\begin{aligned}\tilde{\gamma}_{t+1}F(\tilde{x}^{t+1}) &\overset{(70)}{\le} \tilde{\gamma}_{t+1}P(\tilde{x}^{t+1}) + \tilde{\gamma}_{t+1}f(x^{t+1}) + \tilde{\gamma}_{t+1}\langle \nabla f(x^{t+1}), \tilde{x}^{t+1} - x^{t+1}\rangle + \frac{1}{2}\|\tilde{x}^{t+1} - x^{t+1}\|^2 \\ &\overset{(72)}{\le} \tilde{\gamma}_{t+1}F(x^{t+1}) - \frac{1}{2}\|\tilde{x}^{t+1} - x^{t+1}\|^2,\end{aligned}$$

which yields $\|\tilde{x}^{t+1} - x^{t+1}\| \le \sqrt{2\tilde{\gamma}_{t+1}(F(x^{t+1}) - F(\tilde{x}^{t+1}))}$. This together with (71), $\tilde{x}^{t+1} \in \widehat{\mathcal{S}}$, $\gamma_0 \ge \tilde{\gamma}_{t+1} \ge \min\{\gamma_0, \delta/L_{\widehat{\mathcal{S}}}\}$, and Lemma 1(ii) implies

$$\begin{aligned}\mathrm{dist}(0, \partial F(\tilde{x}^{t+1})) &\le \|\tilde{\gamma}_{t+1}^{-1}(x^{t+1} - \tilde{x}^{t+1}) + \nabla f(\tilde{x}^{t+1}) - \nabla f(x^{t+1})\| \le (\tilde{\gamma}_{t+1}^{-1} + L_{\widehat{\mathcal{S}}})\|\tilde{x}^{t+1} - x^{t+1}\| \\ &\le \left(\sqrt{2\tilde{\gamma}_{t+1}^{-1}} + \sqrt{2\tilde{\gamma}_{t+1}}L_{\widehat{\mathcal{S}}}\right)\sqrt{F(x^{t+1}) - F(\tilde{x}^{t+1})} \\ &\le \left(\sqrt{2\max\{\gamma_0^{-1}, L_{\widehat{\mathcal{S}}}\delta^{-1}\}} + \sqrt{2\gamma_0}L_{\widehat{\mathcal{S}}}\right)\sqrt{F(x^{t+1}) - F(x^*)}.\end{aligned}$$

□

We are now ready to prove Theorem 3.

Proof of Theorem 3. Suppose for contradiction that Algorithm 2 does not terminate within T iterations. It then follows that x^{t+1} and $\tilde{x}^{t+1}$ must be generated in Algorithm 2 for some $T-M<t\le T$ with $\mathrm{mod}(t,M)=0$. In addition, observe that (12) also holds for Algorithm 2. By $t>T-M$, (12), (16) and (68), one has

$$\begin{aligned}
\|\tilde{\gamma}_{t+1}^{-1}(x^{t+1}-\tilde{x}^{t+1})+\nabla f(\tilde{x}^{t+1})-\nabla f(x^{t+1})\| &\overset{(68)}{\le} \left(\sqrt{2\max\{\gamma_0^{-1},L_{\widehat{\mathcal{S}}}\delta^{-1}\}}+\sqrt{2\gamma_0}L_{\widehat{\mathcal{S}}}\right)\sqrt{F(x^{t+1})-F(x^*)}\\
&\overset{(12)}{\le} r_0\left(\sqrt{2\max\{\gamma_0^{-1},L_{\widehat{\mathcal{S}}}\delta^{-1}\}}+\sqrt{2\gamma_0}L_{\widehat{\mathcal{S}}}\right)\left(1-\sqrt{\mu\min\left\{\gamma_0,\delta L_{\widehat{\mathcal{S}}}^{-1}\right\}}\right)^{t/2}\\
&< r_0\left(\sqrt{2\max\{\gamma_0^{-1},L_{\widehat{\mathcal{S}}}\delta^{-1}\}}+\sqrt{2\gamma_0}L_{\widehat{\mathcal{S}}}\right)\left(1-\sqrt{\mu\min\left\{\gamma_0,\delta L_{\widehat{\mathcal{S}}}^{-1}\right\}}\right)^{(T-M)/2} \overset{(16)}{\le} \epsilon.
\end{aligned}$$

which implies that Algorithm 2 terminates at iteration t and leads to a contradiction. Consequently, Algorithm 2 must terminate at some iteration $t\le T$ and output $\tilde{x}^{t+1}$ that satisfies (13). By this and Lemma 8, one can see that $\mathrm{dist}(0,\partial F(\tilde{x}^{t+1}))\le\epsilon$ and hence $\tilde{x}^{t+1}$ is an ϵ-residual solution of problem (1).

In addition, one can observe from Algorithm 2 that (i) evaluations of ∇f and proximal operator of P are performed in the backtracking line search procedure (see step 2) and Algorithm 3 (see step 4); (ii) the total number of iterations of Algorithm 2 is at most T; (iii) n_t backtracking trials are performed in each iteration t and each of them requires one evaluation of ∇f and proximal operator of P; (iv) the total number of calls of Algorithm 3 in Algorithm 2 is at most T/M and each call requires at most N evaluations of ∇f and proximal operator of P (see Algorithm 3 and Lemma 7), where N is given in (11). By this observation and Theorem 1, one can see that the total number of evaluations of ∇f and proximal operator of P performed in Algorithm 2 is no more than $\bar{N}$, respectively. □

5.3 Proof of the main results in Subsection 2.3

In this subsection we first establish several technical lemmas and then use them to prove Theorem 4.

Let $\{x^k\}_{k\in\mathbb{K}}$ denote all the iterates generated by Algorithm 4, where $\mathbb{K}$ is a subset of consecutive nonnegative integers starting from 0. We define $\mathbb{K}-1=\{k-1:k\in\mathbb{K}\}$. For any $0\le k\in\mathbb{K}-1$, let f_k and F_k be defined in (18). Also, let x_*^k be defined as

$$x_*^k=\arg\min_x F_k(x). \tag{73}$$

Recall that α_0, γ_0 and $\{\rho_k\}$ are the input parameters of Algorithm 4, and $L_{\nabla f}$ and $\widehat{L}_{\nabla f}$ are the Lipschitz constant of ∇f on $\mathcal{Q}$ and $\widehat{\mathcal{Q}}$, respectively. Let

$$L_k=L_{\nabla f}+\rho_k^{-1},\quad \widehat{L}_k=\widehat{L}_{\nabla f}+\rho_k^{-1}, \tag{74}$$

$$\bar{r}_k=\sqrt{F_k(x^k)-F_k(x_*^k)+\frac{\alpha_0^2}{2\gamma_0}\|x^k-x_*^k\|^2}, \tag{75}$$

$$\mathcal{S}_k=\left\{x\in\mathrm{dom}(P):\|x-x_*^k\|\le\alpha_0^{-1}\sqrt{2\gamma_0}\bar{r}_k\right\}, \tag{76}$$

$$\widehat{\mathcal{S}}_k=\left\{x\in\mathrm{dom}(P):\|x-x_*^k\|\le(1+\gamma_0L_k)\,\alpha_0^{-1}\sqrt{2\gamma_0}\bar{r}_k\right\}. \tag{77}$$

Since $L_{\nabla f}$ and $\widehat{L}_{\nabla f}$ are respectively the Lipschitz constant of ∇f on $\mathcal{Q}$ and $\widehat{\mathcal{Q}}$, it then follows from (18) that ∇f_k is L_k- and $\widehat{L}_k$-Lipschitz continuous on $\mathcal{Q}$ and $\widehat{\mathcal{Q}}$, respectively. In addition, by the definition of L and $\widehat{L}$ in (24) and the monotonicity of $\{\rho_k\}$, one has

$$L_k=L_{\nabla f}+\rho_k^{-1}\le L,\quad \widehat{L}_k=\widehat{L}_{\nabla f}+\rho_k^{-1}\le\widehat{L}. \tag{78}$$

Lemma 9. *Let x_*^k be defined in* (73)*. Then the following statements hold.*

$$\|x^k-x_*^k\|^2+\|x_*^k-x^*\|^2\le\|x^k-x^*\|^2\quad\forall 0\le k\in\mathbb{K}-1, \tag{79}$$

$$\|x^k-x^{k-1}\|\le\|x^0-x^*\|+\sum_{i=0}^{k-1}\rho_i\eta_i,\quad \|x^k-x^*\|\le\|x^0-x^*\|+\sum_{i=0}^{k-1}\rho_i\eta_i\quad\forall 1\le k\in\mathbb{K}. \tag{80}$$

Proof. One can observe that Algorithm 4 is an inexact proximal point algorithm (PPA) [18] applied to the monotone inclusion problem $0 \in \mathcal{T}(x)$, where $\mathcal{T} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a maximal monotone set-valued operator defined as

$$\mathcal{T}(x) = \begin{cases} \partial F(x) & \text{if } x \in \text{dom}(P), \\ \emptyset & \text{otherwise,} \end{cases} \qquad \forall x \in \mathbb{R}^n.$$

In addition, one can observe from (20) and (73) that $\text{dist}(0, \mathcal{T}(x^{k+1}) + \rho_k^{-1}(x^{k+1} - x^k)) \leq \eta_k$ and $x_*^k = (I + \rho_k \mathcal{T})^{-1}(x^k)$. It then follows from [18, Proposition 3] that

$$\|x^{k+1} - (I + \rho_k \mathcal{T})^{-1}(x^k)\| \leq \rho_k \eta_k \quad \forall k \in \mathbb{K} - 1. \tag{81}$$

By this, $0 \in \mathcal{T}(x^*)$, $x_*^k = (I + \rho_k \mathcal{T})^{-1}(x^k)$ and [18, Proposition 1], one can see that (79) holds. In addition, (80) follows from (81) and [12, Lemma 3]. □

As a consequence of Lemma 9 and the definition of r_0 and θ in (21), one has that

$$\|x^0 - x_*^0\| \leq r_0, \ \ \|x^k - x^*\| \leq r_0 + \theta, \ \ \|x^k - x_*^k\| \leq r_0 + \theta, \ \ \|x^k - x^{k-1}\| \leq r_0 + \theta, \quad \forall 1 \leq k \in \mathbb{K}. \tag{82}$$

Lemma 10. *Let $\tilde{r}_0$ and $\bar{r}^k$ be defined in (22) and (75). Then for all $0 \leq k \in \mathbb{K} - 1$, we have*

$$\bar{r}_k^2 \leq \alpha_0^2 \tilde{r}_0^2 / (2\gamma_0). \tag{83}$$

Proof. We first prove that (83) holds for $k = 0$, that is, $\bar{r}_0^2 \leq \alpha_0^2 \tilde{r}_0^2/(2\gamma_0)$. By (1), (18) and the definition of x^*, one has

$$F_0(x_*^0) = F(x_*^0) + \frac{1}{2\rho_0}\|x_*^0 - x^0\|^2 \geq F(x^*), \quad F_0(x^0) = F(x^0).$$

It then follows from these, (22), (75), and (82) that

$$\bar{r}_0^2 \overset{(75)}{=} F_0(x^0) - F_0(x_*^0) + \frac{\alpha_0^2}{2\gamma_0}\|x^0 - x_*^0\|^2 \leq F(x^0) - F(x^*) + \frac{\alpha_0^2 r_0^2}{2\gamma_0} \overset{(22)}{\leq} \frac{\alpha_0^2 \tilde{r}_0^2}{2\gamma_0}.$$

We next show that (83) holds for all $1 \leq k \in \mathbb{K} - 1$. It follows from (18) and (20) that there exists $P'(x^k) \in \partial P(x^k)$ such that

$$F'_{k-1}(x^k) = \nabla f(x^k) + \rho_{k-1}^{-1}(x^k - x^{k-1}) + P'(x^k) \in \partial F_{k-1}(x^k), \quad \|F'_{k-1}(x^k)\| \leq \eta_{k-1}. \tag{84}$$

Also, we have

$$\nabla f(x^k) + P'(x^k) \in \partial F_k(x^k),$$

which together with (84) yields

$$F'_{k-1}(x^k) - \rho_{k-1}^{-1}(x^k - x^{k-1}) \in \partial F_k(x^k). \tag{85}$$

By the convexity of F, $\eta_{k-1} \leq \eta_0$, $\rho_{k-1} \geq \rho_0$, (82) and (85), one has

$$\begin{aligned} F_k(x^k) - F_k(x_*^k) &\overset{(85)}{\leq} \langle F'_{k-1}(x^k) - \rho_{k-1}^{-1}(x^k - x^{k-1}), x^k - x_*^k \rangle \leq (\|F'_{k-1}(x^k)\| + \rho_{k-1}^{-1}\|x^k - x^{k-1}\|)\|x^k - x_*^k\| \\ &\overset{(82)}{\leq} \eta_0(r_0 + \theta) + \rho_0^{-1}(r_0 + \theta)^2. \end{aligned}$$

This together with (75) and (82) yields

$$\bar{r}_k^2 = \ F_k(x^k) - F_k(x_*^k) + \frac{\alpha_0^2}{2\gamma_0}\|x^k - x_*^k\|^2 \leq \eta_0(r_0 + \theta) + \rho_0^{-1}(r_0 + \theta)^2 + \frac{\alpha_0^2(r_0 + \theta)^2}{2\gamma_0}.$$

By this relation and the definition of $\tilde{r}_0$ in (22), one can see that (83) holds for all $1 \leq k \in \mathbb{K} - 1$. □

Lemma 11. *Let f_k, L_k, $\widehat{L}_k$, $\mathcal{S}_k$ and $\widehat{\mathcal{S}}_k$ be respectively defined in (18), (74), (76) and (77). Then for all $0 \leq k \in \mathbb{K} - 1$, ∇f_k is Lipschitz continuous on $\mathcal{S}_k$ and $\widehat{\mathcal{S}}_k$ with Lipschitz constants L_k and $\widehat{L}_k$, respectively.*

Proof. Let $\mathcal{Q}$ and $\widehat{\mathcal{Q}}$ be defined in (23) and (24). We first show that $\mathcal{S}_k \subseteq \mathcal{Q}$ and $\widehat{\mathcal{S}}_k \subseteq \widehat{\mathcal{Q}}$ for all $0 \le k \in \mathbb{K}-1$. To this end, fix any $0 \le k \in \mathbb{K}-1$. By (76), (82) and (83), one has that for all $x \in \mathcal{S}_k$,

$$\|x - x^*\| \le \|x - x_*^k\| + \|x_*^k - x^*\| \overset{(76)}{\le} \alpha_0^{-1}\sqrt{2\gamma_0}\bar{r}_k + \|x_*^k - x^*\| \le \tilde{r}_0 + r_0 + \theta,$$

where the last inequality follows from (82) and (83). This together with (23) implies that $\mathcal{S}_k \subseteq \mathcal{Q}$. In addition, using (77), (78), (82) and (83), we obtain that for all $x \in \widehat{\mathcal{S}}_k$,

$$\begin{aligned}\|x - x^*\| &\le \|x - x_*^k\| + \|x_*^k - x^*\| \overset{(77)}{\le} (1+\gamma_0 L_k)\,\alpha_0^{-1}\sqrt{2\gamma_0}\bar{r}_k + \|x_*^k - x^*\| \\ &\le (1+\gamma_0 L_k)\tilde{r}_0 + r_0 + \theta \overset{(78)}{\le} (1+\gamma_0 L)\tilde{r}_0 + r_0 + \theta,\end{aligned}$$

which along with (24) implies that $\widehat{\mathcal{S}}_k \subseteq \widehat{\mathcal{Q}}$.

Recall that ∇f_k is L_k- and $\widehat{L}_k$-Lipschitz continuous on $\mathcal{Q}$ and $\widehat{\mathcal{Q}}$, respectively. The conclusion of this lemma then follows from this fact and the relations $\mathcal{S}_k \subseteq \mathcal{Q}$ and $\widehat{\mathcal{S}}_k \subseteq \widehat{\mathcal{Q}}$ for all $0 \le k \in \mathbb{K}-1$. □

Lemma 12. *Let N_k denote the number of evaluations of ∇f and proximal operator of P performed by Algorithm 2 at the kth outer iteration of Algorithm 4. Then for all $0 \le k \in \mathbb{K}-1$, it holds that*

$$N_k \le \widetilde{C}_1 \left(M + 1 + \frac{\left(\log \frac{\alpha_0^2 \tilde{r}_0^2 \left(\sqrt{\max\{\gamma_0^{-2}, \gamma_0^{-1}\widehat{L}\delta^{-1}\}} + \widehat{L} \right)^2}{\eta_k^2} \right)_+}{\sqrt{\rho_k^{-1} \min\left\{\gamma_0, \delta\widehat{L}^{-1}\right\}}} \right), \tag{86}$$

where M, δ, α_0, γ_0, $\{\rho_k\}$ and $\{\eta_k\}$ are the input parameters of Algorithm 4, and $\tilde{r}_0$, $\widehat{L}$ and $\widetilde{C}_1$ are given in (22), (24) *and* (25)*, respectively.*

Proof. Notice that at the kth outer iteration of Algorithm 4, Algorithm 2 is called to find an η_k-residual solution x^{k+1} of the problem $\min_x \{f_k(x) + P(x)\}$ with the inputs $\epsilon \leftarrow \eta_k$, $\mu \leftarrow \rho_k^{-1}$ and $x^1 = z^1 \leftarrow x^k$. In view of (75), (76), (77), Lemma 11 and Theorem 3, one can replace $(r_0, \mu, \epsilon, L_{\widehat{\mathcal{S}}})$ in (17) by $(\bar{r}_k, \rho_k^{-1}, \eta_k, \widehat{L}_k)$ respectively and obtain that

$$\begin{aligned}N_k &\le (1+M^{-1}) \left(M + \left\lceil \frac{2\log \frac{\eta_k}{\bar{r}_k\left(\sqrt{2\max\{\gamma_0^{-1}, \widehat{L}_k\delta^{-1}\}} + \sqrt{2\gamma_0}\widehat{L}_k\right)}}{\log\left(1 - \sqrt{\rho_k^{-1}\min\left\{\gamma_0, \delta\widehat{L}_k^{-1}\right\}}\right)} \right\rceil_+ \right) \left(1 + \left\lceil \frac{\log(\gamma_0\widehat{L}_k)}{\log(1/\delta)} \right\rceil_+ \right) \\ &\le (1+M^{-1}) \left(M + 1 + \frac{\left(\log \frac{2\bar{r}_k^2 \left(\sqrt{\max\{\gamma_0^{-1}, \widehat{L}_k\delta^{-1}\}} + \sqrt{\gamma_0}\widehat{L}_k \right)^2}{\eta_k^2} \right)_+}{-\log\left(1 - \sqrt{\rho_k^{-1}\min\left\{\gamma_0, \delta\widehat{L}_k^{-1}\right\}}\right)} \right) \left(1 + \left\lceil \frac{\log(\gamma_0\widehat{L}_k)}{\log(1/\delta)} \right\rceil_+ \right) \\ &\le (1+M^{-1}) \left(M + 1 + \frac{\left(\log \frac{2\gamma_0\bar{r}_k^2 \left(\sqrt{\max\{\gamma_0^{-2}, \gamma_0^{-1}\widehat{L}_k\delta^{-1}\}} + \widehat{L}_k \right)^2}{\eta_k^2} \right)_+}{\sqrt{\rho_k^{-1}\min\left\{\gamma_0, \delta\widehat{L}_k^{-1}\right\}}} \right) \left(1 + \left\lceil \frac{\log(\gamma_0\widehat{L}_k)}{\log(1/\delta)} \right\rceil_+ \right),\end{aligned}$$

where the last inequality follows from the fact that $-\log(1-\xi) \ge \xi$ for any $\xi \in (0,1)$. By the above inequality, (25), (78) and (83), one can see that (86) holds. □

We are now ready to prove Theorem 4.

Proof of Theorem 4. (i) Let K be defined in (27). We first show that Algorithm 4 terminates after at most $K+1$ outer iterations. Indeed, suppose for contradiction that it runs for more than $K+1$ outer iterations. It then follows that (19) does not hold for $k=K$. On the other hand, by (27), (82), $\rho_K=\rho_0\zeta^K$ and $\eta_K=\eta_0\sigma^K$, one has

$$\frac{1}{\rho_K}\|x^{K+1}-x^K\| \le \frac{r_0+\theta}{\rho_0\zeta^K} \overset{(27)}{\le} \frac{\varepsilon}{2}, \qquad \eta_K=\eta_0\sigma^K \overset{(27)}{\le} \frac{\varepsilon}{2},$$

and hence (19) holds for $k=K$, which leads to a contradiction. Hence, there exists some $0\le k\le K$ such that (19) holds and Algorithm 4 terminates and outputs x^{k+1}. We next show that x^{k+1} is an ε-residual solution of problem (1). Indeed, it follows from (18) and (19) that

$$\begin{aligned}\mathrm{dist}(0,\partial F(x^{k+1})) &\le \mathrm{dist}(0,\partial F(x^{k+1})+\rho_k^{-1}(x^{k+1}-x^k))+\rho_k^{-1}\|x^{k+1}-x^k\| \\ &\overset{(18)}{=} \mathrm{dist}(0,\partial F_k(x^{k+1}))+\rho_k^{-1}\|x^{k+1}-x^k\| \le \eta_k+\rho_k^{-1}\|x^{k+1}-x^k\| \overset{(19)}{\le} \varepsilon,\end{aligned}$$

and hence the output x^{k+1} of Algorithm 4 is an ε-residual solution of problem (1).

(ii) Let K and $\widetilde{N}$ be defined in (27) and (28), and let N_k denote the number of evaluations of ∇f and proximal operator of P performed by Algorithm 2 at the kth outer iteration of Algorithm 4. By this and statement (i) of this theorem, one can observe that the total number of evaluations of ∇f and proximal operator of P performed in Algorithm 4 is no more than $\sum_{k=0}^{|\mathbb{K}|-2} N_k$. As a result, to prove statement (ii) of this theorem, it suffices to show that $\sum_{k=0}^{|\mathbb{K}|-2} N_k \le \widetilde{N}$. Indeed, in view of (26), (27), (86), $|\mathbb{K}|-2\le K$, $\rho_k=\rho_0\zeta^k$ and $\eta_k=\eta_0\sigma^k$, one has

$$\begin{aligned}\sum_{k=0}^{|\mathbb{K}|-2} N_k &\le \widetilde{C}_1\sum_{k=0}^{K}\left(M+1+\frac{\sqrt{\rho_k}\left(\log\frac{\alpha_0^2\tilde{r}_0^2\left(\sqrt{\max\{\gamma_0^{-2},\gamma_0^{-1}\widehat{L}\delta^{-1}\}}+\widehat{L}\right)^2}{\eta_k^2}\right)_+}{\min\left\{\sqrt{\gamma_0},\sqrt{\delta\widehat{L}^{-1}}\right\}}\right) \\ &= \widetilde{C}_1\sum_{k=0}^{K}\left(M+1+\frac{\sqrt{\rho_0}\sqrt{\zeta}^k\left(-2k\log\sigma+\log\frac{\alpha_0^2\tilde{r}_0^2\left(\sqrt{\max\{\gamma_0^{-2},\gamma_0^{-1}\widehat{L}\delta^{-1}\}}+\widehat{L}\right)^2}{\eta_0^2}\right)_+}{\min\left\{\sqrt{\gamma_0},\sqrt{\delta\widehat{L}^{-1}}\right\}}\right) \\ &\le \widetilde{C}_1\left((M+1)(K+1)+\frac{\sqrt{\rho_0}\sqrt{\zeta}^{K+1}\left(-2K\log\sigma+\log\frac{\alpha_0^2\tilde{r}_0^2\left(\sqrt{\max\{\gamma_0^{-2},\gamma_0^{-1}\widehat{L}\delta^{-1}\}}+\widehat{L}\right)^2}{\eta_0^2}\right)_+}{(\sqrt{\zeta}-1)\min\left\{\sqrt{\gamma_0},\sqrt{\delta\widehat{L}^{-1}}\right\}}\right) \le \widetilde{N},\end{aligned}$$

where the first inequality follows from (86), the second inequality is due to $\sum_{k=0}^{K}\sqrt{\zeta}^k \le \sqrt{\zeta}^{K+1}/(\sqrt{\zeta}-1)$ and $\sum_{k=0}^{K} k\sqrt{\zeta}^k \le K\sqrt{\zeta}^{K+1}/(\sqrt{\zeta}-1)$, and the last inequality follows from (26), (27) and (28). □

5.4 Proof of the main results in Section 3

In this subsection we first establish several technical lemmas and then use them to prove Theorem 5.

Let $\{(x^k,\lambda^k)\}_{k\in\mathbb{K}}$ denote all the iterates generated by Algorithm 5, where $\mathbb{K}$ is a subset of consecutive nonnegative integers starting from 0. We define $\mathbb{K}-1=\{k-1: k\in\mathbb{K}\}$. For any $0\le k\in\mathbb{K}-1$, let f_k and F_k be defined in (31). In addition, let (x_*^k,λ_*^k) be defined as

$$x_*^k=\arg\min_x F_k(x), \qquad \lambda_*^k=\Pi_{\mathcal{K}^*}\left(\lambda^k+\rho_k g(x_*^k)\right). \tag{87}$$

Recall that α_0, $\{\rho_k\}$ and $\{\eta_k\}$ are the input parameters of Algorithm 5, $\mathcal{Q}$, B, C, $\widehat{\mathcal{Q}}$, $\widehat{B}$ and $\widehat{C}$ are respectively given in (36), (37), (38) and (39), and $L_{\nabla g}$ and $\widehat{L}_{\nabla g}$ are the Lipschitz constant of ∇g on $\mathcal{Q}$ and $\widehat{\mathcal{Q}}$, respectively.

Let

$$L_k = C\rho_k + B + L_{\nabla g}\sum_{i=0}^{k-1}\rho_i\eta_i + \rho_k^{-1}, \quad \widehat{L}_k = \widehat{C}\rho_k + \widehat{B} + \widehat{L}_{\nabla g}\sum_{i=0}^{k-1}\rho_i\eta_i + \rho_k^{-1}, \tag{88}$$

$$\bar{r}_k = \sqrt{F_k(x^k) - F_k(x_*^k) + \frac{1}{2}\rho_k\alpha_0^2\|x^k - x_*^k\|^2}, \tag{89}$$

$$\mathcal{S}_k = \left\{x \in \mathrm{dom}(P) : \|x - x_*^k\| \le \alpha_0^{-1}\sqrt{2\rho_k^{-1}}\bar{r}_k\right\}, \tag{90}$$

$$\widehat{\mathcal{S}}_k = \left\{x \in \mathrm{dom}(P) : \|x - x_*^k\| \le \left(1 + L_k\rho_k^{-1}\right)\alpha_0^{-1}\sqrt{2\rho_k^{-1}}\bar{r}_k\right\}. \tag{91}$$

The following lemma states some properties of the function f_k, whose proof is similar to that of [12, Lemma 7] and thus omitted.

Lemma 13. *Let f_k, $\mathcal{Q}$, $\widehat{\mathcal{Q}}$, L_k and $\widehat{L}_k$ be respectively defined in* (31), (36), (38) *and* (88). *Then f_k is convex and continuously differentiable on* $\mathrm{dom}(P)$, *and moreover, ∇f_k is Lipschitz continuous on $\mathcal{Q}$ and $\widehat{\mathcal{Q}}$ with Lipschitz constants L_k and $\widehat{L}_k$, respectively.*

The next lemma establishes some properties of (x^k, λ^k) and (x_*^k, λ_*^k).

Lemma 14. *Let (x_*^k, λ_*^k) be defined in* (87). *Then the following statements hold.*

$$\|(x^k,\lambda^k) - (x_*^k,\lambda_*^k)\|^2 + \|(x_*^k,\lambda_*^k) - (x^*,\lambda^*)\|^2 \le \|(x^k,\lambda^k) - (x^*,\lambda^*)\|^2 \quad \forall 0 \le k \in \mathbb{K} - 1, \tag{92}$$

$$\|(x^k,\lambda^k) - (x^{k-1},\lambda^{k-1})\| \le \|(x^0,\lambda^0) - (x^*,\lambda^*)\| + \sum_{i=0}^{k-1}\rho_i\eta_i \quad \forall 1 \le k \in \mathbb{K}, \tag{93}$$

$$\|(x^k,\lambda^k) - (x^*,\lambda^*)\| \le \|(x^0,\lambda^0) - (x^*,\lambda^*)\| + \sum_{i=0}^{k-1}\rho_i\eta_i \quad \forall 1 \le k \in \mathbb{K}. \tag{94}$$

Proof. It is well-known (e.g., see [18, 12]) that Algorithm 5 is an inexact proximal point algorithm (PPA) applied to the monotone inclusion problem $0 \in \mathcal{T}_l(x,\lambda)$, where l is the Lagrangian function of problem (2), and $\mathcal{T}_l$ is a maximal monotone set-valued operator defined as

$$\mathcal{T}_l : (x,\lambda) \to \{(v,u) \in \Re^n \times \Re^m : (v,-u) \in \partial l(x,\lambda)\}, \quad \forall (x,\lambda) \in \Re^n \times \Re^m.$$

It then follows from (33), (87), and [12, Lemma 5] that

$$(x_*^k,\lambda_*^k) = \mathcal{J}_{\rho_k}(x^k,\lambda^k), \qquad \|(x^{k+1},\lambda^{k+1}) - \mathcal{J}_{\rho_k}(x^k,\lambda^k)\| \le \rho_k\eta_k, \quad \forall k \in \mathbb{K} - 1. \tag{95}$$

where $\mathcal{J}_{\rho_k} = (\mathcal{I} + \rho_k\mathcal{T}_l)^{-1}$. By the first relation in (95), $0 \in \mathcal{T}_l(x^*,\lambda^*)$, and the maximal monotonicity of $\mathcal{T}_l$, it follows from [18, Proposition 1] that (92) holds. In addition, (93) and (94) follow from the second relation in (95) and [12, Lemma 3]. □

As a consequence of Lemma 14 and the definition of r_0 and θ in (34), one has that

$$\|x^0 - x_*^0\| \le r_0, \ \ \|x^k - x^*\| \le r_0 + \theta, \ \ \|\lambda^k - \lambda^*\| \le r_0 + \theta, \ \ \|x^k - x_*^k\| \le r_0 + \theta, \ \ \|x^k - x^{k-1}\| \le r_0 + \theta \quad \forall 1 \le k \in \mathbb{K}. \tag{96}$$

Lemma 15. *Let $\tilde{r}_0$ and $\bar{r}^k$ be defined in* (35) *and* (89). *Then for all $0 \le k \in \mathbb{K} - 1$, we have*

$$\bar{r}_k^2 \le \alpha_0^2\tilde{r}_0^2\rho_k/2. \tag{97}$$

Proof. We first prove that (97) holds for $k = 0$, that is, $\bar{r}_0^2 \le \alpha_0^2\tilde{r}_0^2\rho_0/2$. Indeed, let l be the Lagrangian function of problem (2). By (30), (31) and (87), one has

$$\begin{aligned} F_0(x_*^0) &= \mathcal{L}(x_*^0,\lambda^0;\rho_0) + \frac{1}{2\rho_0}\|x_*^0 - x^0\|^2 \ge \mathcal{L}(x_*^0,\lambda^0;\rho_0) = \max_{\lambda\in\mathbb{R}^m}\left\{l(x_*^0,\lambda) - \frac{1}{2\rho_0}\|\lambda - \lambda^0\|^2\right\} \\ &\ge l(x_0^*,\lambda^*) - \frac{1}{2\rho_0}\|\lambda^0 - \lambda^*\|^2 \ge F(x^*) - \frac{1}{2\rho_0}\|\lambda^0 - \lambda^*\|^2, \end{aligned}$$

where the second equality follows from [12, Lemma 2]. Also, we have

$$F_0(x^0) = \mathcal{L}(x^0, \lambda^0; \rho_0) = F(x^0) + \frac{1}{2\rho_0}\left(\|\Pi_{\mathcal{K}^*}(\lambda^0 + \rho_0 g(x^0))\|^2 - \|\lambda^0\|^2\right).$$

It then follows from these, (35), (89), and (96) that

$$\begin{aligned}\bar{r}_0^2 &\overset{(89)}{=} F_0(x^0) - F_0(x_*^0) + \frac{1}{2}\rho_0\alpha_0^2\|x^0 - x_*^0\|^2 \\ &\le F(x^0) - F(x^*) + \frac{1}{2\rho_0}\left(\|\Pi_{\mathcal{K}^*}(\lambda^0 + \rho_0 g(x^0))\|^2 + \|\lambda^0 - \lambda^*\|^2 - \|\lambda^0\|^2\right) + \frac{1}{2}\rho_0\alpha_0^2 r_0^2 \overset{(35)}{\le} \alpha_0^2\tilde{r}_0^2\rho_0/2.\end{aligned}$$

We next show that (97) holds for all $1 \le k \in \mathbb{K} - 1$. Indeed, observe that $\|\lambda^k\| = \mathrm{dist}(\lambda^{k-1} + \rho_{k-1}g(x^k), -\mathcal{K})$ and $\|\Pi_{\mathcal{K}^*}(\lambda^k + \rho_k g(x^k))\| = \mathrm{dist}(\lambda^k + \rho_k g(x^k), -\mathcal{K})$. Using these, $\rho_k = \rho_0\zeta^k$, and (96), we have

$$\begin{aligned}\|\Pi_{\mathcal{K}^*}(\lambda^k + \rho_k g(x^k)) - \lambda^k\| &\le \mathrm{dist}(\lambda^k + \rho_k g(x^k), -\mathcal{K}) + \|\lambda^k\| = \rho_k \mathrm{dist}\left(\frac{\lambda^k}{\rho_k} + g(x^k), -\mathcal{K}\right) + \|\lambda^k\| \\ &\le \rho_k \mathrm{dist}\left(\frac{\lambda^k}{\rho_k} - \frac{\lambda^{k-1}}{\rho_{k-1}}, -\mathcal{K}\right) + \rho_k \mathrm{dist}\left(\frac{\lambda^{k-1}}{\rho_{k-1}} + g(x^k), -\mathcal{K}\right) + \|\lambda^k\| \\ &\le \rho_k \left\|\frac{\lambda^k}{\rho_k} - \frac{\lambda^{k-1}}{\rho_{k-1}}\right\| + \frac{\rho_k}{\rho_{k-1}}\mathrm{dist}\left(\lambda^{k-1} + \rho_{k-1}g(x^k), -\mathcal{K}\right) + \|\lambda^k\| \\ &= \rho_k \left\|\frac{\lambda^k}{\rho_k} - \frac{\lambda^{k-1}}{\rho_{k-1}}\right\| + \left(\frac{\rho_k}{\rho_{k-1}} + 1\right)\|\lambda^k\| \le \frac{\rho_k}{\rho_{k-1}}\|\lambda^{k-1}\| + \left(\frac{\rho_k}{\rho_{k-1}} + 2\right)\|\lambda^k\| \\ &\le 2(\zeta+1)(\|\lambda^*\| + r_0 + \theta). \end{aligned} \tag{98}$$

It follows from (31) and (33) that there exists $P'(x^k) \in \partial P(x^k)$ such that

$$F'_{k-1}(x^k) = \nabla f(x^k) + \nabla g(x^k)\Pi_{\mathcal{K}^*}(\lambda^{k-1} + \rho_{k-1}g(x^k)) + \rho_{k-1}^{-1}(x^k - x^{k-1}) + P'(x^k) \in \partial F_{k-1}(x^k), \quad \|F'_{k-1}(x^k)\| \le \eta_{k-1}. \tag{99}$$

Also, we have

$$\nabla f(x^k) + \nabla g(x^k)\Pi_{\mathcal{K}^*}(\lambda^k + \rho_k g(x^k)) + P'(x^k) \in \partial F_k(x^k),$$

which together with (99) yields

$$F'_{k-1}(x^k) - \rho_{k-1}^{-1}(x^k - x^{k-1}) + \nabla g(x^k)\left(\Pi_{\mathcal{K}^*}(\lambda^k + \rho_k g(x^k)) - \Pi_{\mathcal{K}^*}(\lambda^{k-1} + \rho_{k-1}g(x^k))\right) \in \partial F_k(x^k). \tag{100}$$

In addition, observe from (34) and (96) that $x^k \in \widetilde{\mathcal{Q}}$. Also, note that F_k is convex and g is $\widetilde{L}_g$-Lipschitz continuous on $\widetilde{\mathcal{Q}}$. By these, (98), (99), (100), and the monotonicity of $\{\rho_k\}$ and $\{\eta_k\}$, one has

$$\begin{aligned}F_k(x^k) - F_k(x_*^k) &\overset{(100)}{\le} \langle F'_{k-1}(x^k), x^k - x_*^k\rangle - \rho_{k-1}^{-1}\langle x^k - x^{k-1}, x^k - x_*^k\rangle \\ &\quad + \langle \nabla g(x^k)(\Pi_{\mathcal{K}^*}(\lambda^k + \rho_k g(x^k)) - \Pi_{\mathcal{K}^*}(\lambda^{k-1} + \rho_{k-1}g(x^k))), x^k - x_*^k\rangle \\ &\le \|F'_{k-1}(x^k)\|\|x^k - x_*^k\| + \rho_{k-1}^{-1}\|x^k - x^{k-1}\|\|x^k - x_*^k\| \\ &\quad + \|\nabla g(x^k)\|\|\Pi_{\mathcal{K}^*}(\lambda^k + \rho_k g(x^k)) - \Pi_{\mathcal{K}^*}(\lambda^{k-1} + \rho_{k-1}g(x^k))\|\|x^k - x_*^k\| \\ &= \|F'_{k-1}(x^k)\|\|x^k - x_*^k\| + \rho_{k-1}^{-1}\|x^k - x^{k-1}\|\|x^k - x_*^k\| \\ &\quad + \|\nabla g(x^k)\|\|\Pi_{\mathcal{K}^*}(\lambda^k + \rho_k g(x^k)) - \lambda^k\|\|x^k - x_*^k\| \\ &\le \eta_0(r_0 + \theta) + \rho_0^{-1}(r_0 + \theta)^2 + 2\widetilde{L}_g(\zeta+1)(\|\lambda^*\| + r_0 + \theta)(r_0 + \theta),\end{aligned}$$

where the last inequality follows from (96) and (98). Then we have

$$\begin{aligned}&\frac{2\bar{r}_k^2}{\rho_k\alpha_0^2} = \frac{2}{\rho_k\alpha_0^2}\left(F_k(x^k) - F_k(x_*^k) + \rho_k\alpha_0^2\|x^k - x_*^k\|^2\right) \le \frac{2}{\rho_0\alpha_0^2}\left(F_k(x^k) - F_k(x_*^k)\right) + 2\|x^k - x_*^k\|^2 \\ \le&\frac{2}{\rho_0\alpha_0^2}\left(\eta_0(r_0 + \theta) + \rho_0^{-1}(r_0 + \theta)^2 + 2\widetilde{L}_g(\zeta+1)(\|\lambda^*\| + r_0 + \theta)(r_0 + \theta)\right) + 2(r_0 + \theta)^2 \\ =&\frac{2(r_0 + \theta)}{\rho_0\alpha_0^2}\left(\eta_0 + \rho_0^{-1}(r_0 + \theta) + 2\widetilde{L}_g(\zeta+1)(\|\lambda^*\| + r_0 + \theta) + \rho_0\alpha_0^2(r_0 + \theta)\right).\end{aligned}$$

By this relation and the definition of $\tilde{r}_0$ in (35), one can see that (97) holds for all $1 \le k \in \mathbb{K} - 1$. □

Lemma 16. *Let f_k, L_k, $\widehat{L}_k$, $\mathcal{S}_k$ and $\widehat{\mathcal{S}}_k$ be respectively defined in* (31), (88), (90) *and* (91). *Then for all $0 \le k \in \mathbb{K}-1$, ∇f_k is Lipschitz continuous on $\mathcal{S}_k$ and $\widehat{\mathcal{S}}_k$ with Lipschitz constants L_k and $\widehat{L}_k$, respectively.*

Proof. Let $\mathcal{Q}$ and $\widehat{\mathcal{Q}}$ be defined in (36) and (38). We first show that $\mathcal{S}_k \subseteq \mathcal{Q}$ and $\widehat{\mathcal{S}}_k \subseteq \widehat{\mathcal{Q}}$ for all $0 \le k \in \mathbb{K}-1$. To this end, fix any $0 \le k \in \mathbb{K}-1$. By (90), (96) and (97), one has that for all $x \in \mathcal{S}_k$,

$$\|x - x^*\| \le \|x - x_*^k\| + \|x_*^k - x^*\| \overset{(90)}{\le} \alpha_0^{-1}\sqrt{2\rho_k^{-1}}\bar{r}_k + \|x_*^k - x^*\| \le \tilde{r}_0 + r_0 + \theta,$$

where the last inequality follows from (96) and (97). This together with (36) implies that $\mathcal{S}_k \subseteq \mathcal{Q}$. In addition, by (34), (37), (88) and $\rho_k \ge \rho_0$, one has

$$\rho_k^{-1}L_k \overset{(88)}{=} C + \rho_k^{-1}B + \rho_k^{-1}L_{\nabla g}\sum_{i=0}^{k-1}\rho_i\eta_i + \rho_k^{-2} \overset{(34)}{\le} C + \rho_0^{-1}B + \rho_0^{-1}L_{\nabla g}\theta + \rho_0^{-2} \overset{(37)}{=} L.$$

Using this, (91) and (97), we obtain that for all $x \in \widehat{\mathcal{S}}_k$,

$$\|x - x^*\| \le \|x - x_*^k\| + \|x_*^k - x^*\| \overset{(91)}{\le} \left(1 + L_k\rho_k^{-1}\right)\alpha_0^{-1}\sqrt{2\rho_k^{-1}}\bar{r}_k + \|x_*^k - x^*\| \le (1+L)\tilde{r}_0 + r_0 + \theta.$$

which along with (38) implies that $\widehat{\mathcal{S}}_k \subseteq \widehat{\mathcal{Q}}$.

The conclusion of this lemma then follows from Lemma 13 and the fact that $\mathcal{S}_k \subseteq \mathcal{Q}$ and $\widehat{\mathcal{S}}_k \subseteq \widehat{\mathcal{Q}}$ for all $0 \le k \in \mathbb{K}-1$. □

Lemma 17. *Let N_k denote the number of evaluations of ∇f, ∇g, proximal operator of P and projection onto $\mathcal{K}^*$ performed by Algorithm 2 at the kth outer iteration of Algorithm 5. Then for all $0 \le k \in \mathbb{K}-1$, it holds that*

$$N_k \le \widehat{C}_1\left(M + 1 + \frac{\left(\log\frac{\rho_k^2\alpha_0^2\tilde{r}_0^2\left(\sqrt{\max\{1,\widehat{L}\delta^{-1}\}}+\widehat{L}\right)^2}{\eta_k^2}\right)_+}{\sqrt{(\mu+\rho_k^{-1})\rho_k^{-1}\min\left\{1,\delta\widehat{L}^{-1}\right\}}}\right), \tag{101}$$

where M, δ, α_0, $\{\rho_k\}$ and $\{\eta_k\}$ are the input parameters of Algorithm 5, and $\tilde{r}_0$, $\widehat{L}$ and $\widehat{C}_1$ are given in (35), (39) *and* (40), *respectively.*

Proof. By (34), (39), (88) and $\rho_k \ge \rho_0$, one has

$$\rho_k^{-1}\widehat{L}_k \overset{(88)}{=} \widehat{C} + \rho_k^{-1}\widehat{B} + \rho_k^{-1}\widehat{L}_{\nabla g}\sum_{i=0}^{k-1}\rho_i\eta_i + \rho_k^{-2} \overset{(34)}{\le} \widehat{C} + \rho_0^{-1}\widehat{B} + \rho_0^{-1}\widehat{L}_{\nabla g}\theta + \rho_0^{-2} \overset{(39)}{=} \widehat{L}. \tag{102}$$

Notice that at the kth outer iteration of Algorithm 5, Algorithm 2 is called to find an η_k-residual solution x^{k+1} of the problem $\min_x\{f_k(x) + P(x)\}$ with the inputs $\epsilon \leftarrow \eta_k$, $\gamma_0 \leftarrow \rho_k^{-1}$, $\mu \leftarrow \mu + \rho_k^{-1}$ and $x^1 = z^1 \leftarrow x^k$. Moreover, when applied to this problem, the proximal step (5) of Algorithm 2 requires one evaluation of ∇f, ∇g, proximal operator of P and projection onto $\mathcal{K}^*$, respectively. In view of this, (89), (90), (91), Lemma 16 and Theorem 3, one can replace $(r_0, \gamma_0, \mu, \epsilon, L_{\widehat{\mathcal{S}}})$ in (17) by $(\bar{r}_k, \rho_k^{-1}, \mu + \rho_k^{-1}, \eta_k, \widehat{L}_k)$ respectively and obtain that

$$\begin{aligned} N_k &\le (1+M^{-1})\left(M + \left\lceil\frac{2\log\frac{\eta_k}{\bar{r}_k\left(\sqrt{2\max\{\rho_k,\widehat{L}_k\delta^{-1}\}}+\sqrt{2\rho_k^{-1}\widehat{L}_k}\right)}}{\log\left(1-\sqrt{(\mu+\rho_k^{-1})\min\left\{\rho_k^{-1},\delta\widehat{L}_k^{-1}\right\}}\right)}\right\rceil_+\right)\left(1 + \left\lceil\frac{\log(\rho_k^{-1}\widehat{L}_k)}{\log(1/\delta)}\right\rceil_+\right) \\ &\le (1+M^{-1})\left(M + 1 + \frac{\left(\log\frac{2\rho_k\bar{r}_k^2\left(\sqrt{\max\{1,\rho_k^{-1}\widehat{L}_k\delta^{-1}\}}+\rho_k^{-1}\widehat{L}_k\right)^2}{\eta_k^2}\right)_+}{-\log\left(1-\sqrt{(\mu+\rho_k^{-1})\rho_k^{-1}\min\left\{1,\delta\rho_k\widehat{L}_k^{-1}\right\}}\right)}\right)\left(1 + \left\lceil\frac{\log(\rho_k^{-1}\widehat{L}_k)}{\log(1/\delta)}\right\rceil_+\right) \end{aligned}$$

$$\leq (1+M^{-1})\left(M+1+\frac{\left(\log\frac{2\rho_k\bar{r}_k^2\left(\sqrt{\max\{1,\rho_k^{-1}\widehat{L}_k\delta^{-1}\}}+\rho_k^{-1}\widehat{L}_k\right)^2}{\eta_k^2}\right)_+}{\sqrt{(\mu+\rho_k^{-1})\rho_k^{-1}\min\left\{1,\delta\rho_k\widehat{L}_k^{-1}\right\}}}\right)\left(1+\left\lceil\frac{\log(\rho_k^{-1}\widehat{L}_k)}{\log(1/\delta)}\right\rceil_+\right),$$

where the last inequality follows from the fact that $-\log(1-\xi)\geq\xi$ for any $\xi\in(0,1)$. By the above inequality, (97) and (102), one can see that (101) holds. □

We are now ready to prove Theorem 5.

Proof of Theorem 5. (i) Let K be defined in (42). We first show that Algorithm 5 terminates after at most $K+1$ outer iterations. Indeed, suppose for contradiction that it runs for more than $K+1$ outer iterations. It then follows that (32) does not hold for $k=K$. On the other hand, by (34), (93), (42), $\rho_K=\rho_0\zeta^K$ and $\eta_K=\eta_0\sigma^K$, one has

$$\frac{1}{\rho_K}\|(x^{K+1},\lambda^{K+1})-(x^K,\lambda^K)\|\leq\frac{r_0+\theta}{\rho_0\zeta^K}\overset{(42)}{\leq}\frac{\varepsilon}{2},\qquad\eta_K=\eta_0\sigma^K\overset{(42)}{\leq}\frac{\varepsilon}{2},$$

and hence (32) holds for $k=K$, which leads to a contradiction. In addition, the output of Algorithm 5 is an ε-KKT solution of problems (2) and (29) due to [12, Theorem 4].

(ii) Suppose that $\mu=0$, i.e., f is convex but not strongly convex on $\mathrm{dom}(P)$. Let K and $\widehat{N}$ be defined in (42) and (43). Also, let N_k denote the number of evaluations of ∇f, ∇g, proximal operator of P and projection onto $\mathcal{K}^*$ performed by Algorithm 2 at the kth outer iteration of Algorithm 5. In addition to these evaluations, one projection onto $\mathcal{K}^*$ is performed at step 3 of Algorithm 5 each iteration. By these and statement (i) of this theorem, one can observe that the total number of evaluations of ∇f, ∇g, proximal operator of P and projection onto $\mathcal{K}^*$ performed in Algorithm 5 is no more than $\sum_{k=0}^{|\mathbb{K}|-2}(N_k+1)$. As a result, to prove statement (ii) of this theorem, it suffices to show that $\sum_{k=0}^{|\mathbb{K}|-2}(N_k+1)\leq\widehat{N}$. Indeed, in view of (41), (42), (101), $|\mathbb{K}|-2\leq K$, $\mu=0$, $\rho_k=\rho_0\zeta^k$ and $\eta_k=\eta_0\sigma^k$, one has

$$\begin{aligned}\sum_{k=0}^{|\mathbb{K}|-2}(N_k+1)&\leq K+1+\widehat{C}_1\sum_{k=0}^{K}\left(M+1+\frac{\rho_k\left(\log\frac{\rho_k^2\alpha_0^2\tilde{r}_0^2\left(\sqrt{\max\{1,\widehat{L}\delta^{-1}\}}+\widehat{L}\right)^2}{\eta_k^2}\right)_+}{\min\left\{1,\sqrt{\delta\widehat{L}^{-1}}\right\}}\right)\\&=K+1+\widehat{C}_1\sum_{k=0}^{K}\left(M+1+\frac{\rho_0\zeta^k\left(2k\log\frac{\zeta}{\sigma}+\log\frac{\rho_0^2\alpha_0^2\tilde{r}_0^2\left(\sqrt{\max\{1,\widehat{L}\delta^{-1}\}}+\widehat{L}\right)^2}{\eta_0^2}\right)_+}{\min\left\{1,\sqrt{\delta\widehat{L}^{-1}}\right\}}\right)\\&\leq K+1+\widehat{C}_1\left((M+1)(K+1)+\frac{\rho_0\zeta^{K+1}\left(2K\log\frac{\zeta}{\sigma}+\log\frac{\rho_0^2\alpha_0^2\tilde{r}_0^2\left(\sqrt{\max\{1,\widehat{L}\delta^{-1}\}}+\widehat{L}\right)^2}{\eta_0^2}\right)_+}{(\zeta-1)\min\left\{1,\sqrt{\delta\widehat{L}^{-1}}\right\}}\right)\leq\widehat{N},\end{aligned}$$

where the first inequality follows from (101) and $\mu=0$, the second inequality is due to $\sum_{k=0}^{K}\zeta^k\leq\zeta^{K+1}/(\zeta-1)$ and $\sum_{k=0}^{K}k\zeta^k\leq K\zeta^{K+1}/(\zeta-1)$, and the last equality follows from (41), (42) and (43).

(iii) Suppose that $\mu>0$, namely, f is strongly convex on $\mathrm{dom}(P)$. Similar to the proof of statement (ii) of this theorem, it suffices to show that $\sum_{k=0}^{|\mathbb{K}|-2}(N_k+1)\leq\check{N}$. Indeed, in view of (41), (42), (101), $|\mathbb{K}|-2\leq K$,

$\mu > 0$, $\rho_k = \rho_0\zeta^k$ and $\eta_k = \eta_0\sigma^k$, one has

$$\begin{aligned}
\sum_{k=0}^{|\mathbb{K}|-2}(N_k+1) &\le K+1+\widehat{C}_1\sum_{k=0}^{K}\left(M+1+\frac{\sqrt{\frac{\rho_k}{\mu}}\left(\log\frac{\rho_k^2\alpha_0^2\tilde{r}_0^2\left(\sqrt{\max\{1,\widehat{L}\delta^{-1}\}}+\widehat{L}\right)^2}{\eta_k^2}\right)_+}{\min\left\{1,\sqrt{\delta\widehat{L}^{-1}}\right\}}\right)\\
&= K+1+\widehat{C}_1\sum_{k=0}^{K}\left(M+1+\frac{\sqrt{\frac{\rho_0}{\mu}}\sqrt{\zeta}^k\left(2k\log\frac{\zeta}{\sigma}+\log\frac{\rho_0^2\alpha_0^2\tilde{r}_0^2\left(\sqrt{\max\{1,\widehat{L}\delta^{-1}\}}+\widehat{L}\right)^2}{\eta_0^2}\right)_+}{\min\left\{1,\sqrt{\delta\widehat{L}^{-1}}\right\}}\right)\\
&\le K+1+\widehat{C}_1\left((M+1)(K+1)+\frac{\sqrt{\frac{\rho_0}{\mu}}\sqrt{\zeta}^{K+1}\left(2K\log\frac{\zeta}{\sigma}+\log\frac{\rho_0^2\alpha_0^2\tilde{r}_0^2\left(\sqrt{\max\{1,\widehat{L}\delta^{-1}\}}+\widehat{L}\right)^2}{\eta_0^2}\right)_+}{(\sqrt{\zeta}-1)\min\left\{1,\sqrt{\delta\widehat{L}^{-1}}\right\}}\right)\le\check{N},
\end{aligned}$$

where the first inequality follows from (101) and $\mu > 0$, the second inequality is due to $\sum_{k=0}^{K}\sqrt{\zeta}^k \le \sqrt{\zeta}^{K+1}/(\sqrt{\zeta}-1)$ and $\sum_{k=0}^{K}k\sqrt{\zeta}^k \le K\sqrt{\zeta}^{K+1}/(\sqrt{\zeta}-1)$, and the last equality follows from (41), (42) and (44). □

6 Concluding remarks

The development and analysis of accelerated first-order methods in this paper are based on the assumption that the proximal subproblems associated with P can be exactly solved. Nevertheless, it is not hard to modify them by using a suitable inexact solution of the proximal subproblems instead.

Recently, a class of problems in the form of (1) with f being relatively smooth convex was considered in the literature (e.g., see [2, 5, 11]). Interestingly, this class consists of some problems in which ∇f is not locally Lipschitz continuous on $\mathrm{cl}(\mathrm{dom}(P))$, for example, the problem with P being the simplex and f containing the entropy function and being relatively smooth to the entropy function. It shall however be mentioned that this class generally does not include the problems considered in this paper. For example, it does not contain problem (1) with f being a convex high-degree polynomial function and P being the indicator function of the nonnegative orthant. Yet, this problem belongs to the class considered in this paper. As future research, it would be interesting to investigate whether the methods studied in this paper can be extended to relatively smooth convex optimization.

References

[1] N. S. Aybat and G. Iyengar. An augmented Lagrangian method for conic convex programming, 2013. arXiv preprint arXiv:1302.6322.

[2] H. H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.

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

[4] O. Fercoq and P. Richtárik. Accelerated, parallel, and proximal coordinate descent. *SIAM Journal on Optimization*, 25:1997–2023, 2015.

[5] F. Hanzely, P. Richtarik, and L. Xiao. Accelerated Bregman proximal gradient methods for relatively smooth convex optimization. *Computational Optimization and Applications*, 79:405–440, 2021.

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

[7] G. Lan and R. D. C. Monteiro. Iteration-complexity of first-order penalty methods for convex programming. *Mathematical Programming*, 138:115–139, 2013.

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

[9] Q. Lin, Z. Lu, and L. Xiao. An accelerated randomized proximal coordinate gradient method and its application to regularized empirical risk minimization. *SIAM Journal on Optimization*, 25:2244–2273, 2015.

[10] Y. F. Liu, X. Liu, and S. Ma. On the non-ergodic convergence rate of an inexact augmented Lagrangian framework for composite convex programming. *Mathematics of Operations Research*, 44:632–650, 2019.

[11] H. Lu, R. M. Freund, and Y. Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

[12] Z. Lu and Z. Zhou. Iteration complexity of first-order augmented Lagrangian methods for convex conic programming. 2023. To appear in *SIAM journal on optimization*.

[13] Y. Malitsky and M. K. Tam. A forward-backward splitting method for monotone inclusions without cocoercivity. *SIAM Journal on Optimization*, 30(2):1451–1472, 2020.

[14] R. D. Monteiro and B. F. Svaiter. Complexity of variants of Tseng's modified FB splitting and Korpelevich's methods for hemivariational inequalities with applications to saddle-point and convex optimization problems. *SIAM Journal on Optimization*, 21(4):1688–1720, 2011.

[15] I. Necoara, A. Patrascu, and F. Glineur. Complexity of first-order inexact Lagrangian and penalty methods for conic convex programming. *Optimization Methods and Software*, 34:305–335, 2019.

[16] Y. E. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140:125–161, 2013.

[17] A. Patrascu, I. Necoara, and T. D. Quoc. Adaptive inexact fast augmented Lagrangian methods for constrained convex optimization. *Optimization Letters*, 11:609–626, 2017.

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

[19] P. Tseng. A modified forward-backward splitting method for maximal monotone mappings. *SIAM Journal on Control and Optimization*, 38(2):431–446, 2000.

[20] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. Manuscript, May 2008.