

# COMPLEXITY OF A QUADRATIC PENALTY ACCELERATED INEXACT PROXIMAL POINT METHOD FOR SOLVING LINEARLY CONSTRAINED NONCONVEX COMPOSITE PROGRAMS

WEIWEI KONG <sup>\*</sup>, JEFFERSON G. MELO <sup>†</sup>, AND RENATO D.C. MONTEIRO <sup>\*</sup>

**Abstract.** This paper analyzes the iteration-complexity of a quadratic penalty accelerated inexact proximal point method for solving linearly constrained nonconvex composite programs. More specifically, the objective function is of the form  $f + h$  where  $f$  is a differentiable function whose gradient is Lipschitz continuous and  $h$  is a closed convex function with possibly unbounded domain. The method, basically, consists of applying an accelerated inexact proximal point method for solving approximately a sequence of quadratic penalized subproblems associated to the linearly constrained problem. Each subproblem of the proximal point method is in turn approximately solved by an accelerated composite gradient (ACG) method. It is shown that the proposed scheme generates a  $\rho$ -approximate stationary point in at most  $\mathcal{O}(\rho^{-3})$  ACG iterations. Finally, numerical results showing the efficiency of the proposed method are also given.

**Key words.** quadratic penalty method, composite nonconvex program, iteration-complexity, inexact proximal point method, first-order accelerated gradient method.

**AMS subject classifications.** 47J22, 90C26, 90C30, 90C60, 65K10.

**1. Introduction.** Our main goal in this paper is to describe and establish the iteration-complexity of a quadratic penalty accelerated inexact proximal point (QP-AIPP) method for solving the linearly constrained nonconvex composite minimization problem

$$(1) \quad \min \{f(z) + h(z) : Az = b, z \in \mathbb{R}^n\}$$

where  $A \in \mathbb{R}^{l \times n}$ ,  $b \in \mathbb{R}^l$ ,  $h : \mathbb{R}^n \rightarrow (-\infty, \infty]$  is a proper lower-semicontinuous convex function and  $f$  is a real-valued differentiable (possibly nonconvex) function whose gradient is  $L_f$ -Lipschitz continuous on  $\text{dom } h$ . For given tolerances  $\hat{\rho} > 0$  and  $\hat{\eta} > 0$ , the main result of this paper shows that the QP-AIPP method, started from any point in  $\text{dom } h$  (but not necessarily satisfying  $Az = b$ ), obtains a triple  $(\hat{z}, \hat{v}, \hat{p})$  satisfying

$$(2) \quad \hat{v} \in \nabla f(\hat{z}) + \partial h(\hat{z}) + A^* \hat{p}, \quad \|\hat{v}\| \leq \hat{\rho}, \quad \|A\hat{z} - b\| \leq \hat{\eta}$$

in at most  $\mathcal{O}(\hat{\rho}^{-2}\hat{\eta}^{-1})$  accelerated composite gradient (ACG) iterations. It is worth noting that this result is obtained under the mild assumption that the optimal value of (1) is finite and hence assumes neither that  $\text{dom } h$  is bounded nor that (1) has an optimal solution.

The QP-AIPP method is based on solving penalized subproblems of the form

$$(3) \quad \min \left\{ f(z) + h(z) + \frac{c}{2} \|Az - b\|^2 : z \in \mathbb{R}^n \right\}$$

for an increasing sequence of positive penalty parameters  $c$ . These subproblems in turn are approximately solved so as to satisfy the first two conditions in (2) and the

<sup>\*</sup>School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 30332-0205. (E-mails: [wkong37@gatech.edu](mailto:wkong37@gatech.edu) and [monteiro@isye.gatech.edu](mailto:monteiro@isye.gatech.edu)). The work of Renato D.C. Monteiro was partially supported by NSF Grant CMMI-1300221, ONR Grant N00014-18-1-2077 and CNPq Grant 406250/2013-8.

<sup>†</sup>Institute of Mathematics and Statistics, Federal University of Goias, Campus II- Caixa Postal 131, CEP 74001-970, Goiânia-GO, Brazil. (E-mail: [jefferson@ufg.br](mailto:jefferson@ufg.br)). The work of this author was supported in part by CNPq Grants 406250/2013-8, 406975/2016-7 and FAPEG/GO.

QP-AIPP method terminates when  $c$  is large enough so as to guarantee that the third condition in (2) also hold. Moreover, each subproblem in turn is approximately solved by an accelerated inexact proximal point (AIPP) method which solves a sequence of prox subproblems of the form

$$(4) \quad \min \left\{ f(z) + h(z) + \frac{c}{2} \|Az - b\|^2 + \frac{1}{2\lambda} \|z - z_{k-1}\|^2 : z \in \mathbb{R}^n \right\}$$

where  $z_{k-1}$  is the previous iterate and the next one, namely  $z_k$ , is a suitable approximate solution of (4). Choosing  $\lambda$  sufficiently small ensures that the objective function of (4) is a convex composite optimization which is approximately solved by an ACG method.

More generally, the AIPP method mentioned above solves problems of the form

$$(5) \quad \phi_* := \min \{ \phi(z) := g(z) + h(z) : z \in \mathbb{R}^n \}$$

where  $h$  is as above and  $g$  is a differentiable function whose gradient is  $M$ -Lipschitz continuous on  $\text{dom } h$  and whose lower curvature is bounded below on  $\text{dom } h$  by some constant  $m \in (0, M]$ , i.e.,

$$g(u) - [g(z) + \langle \nabla g(z), u - z \rangle] \geq -\frac{m}{2} \|u - z\|^2 \quad \forall z, u \in \text{dom } h.$$

Note that the penalized subproblem (3) is a special case of (5) with  $g(z) = f(z) + (c/2)\|Az - b\|^2$ , and hence  $m = L_f$  and  $M = L_f + c\|A\|^2$ . It is well-known that the composite gradient method finds a  $\rho$ -solution of (5), i.e., a pair  $(\bar{z}, \bar{v}) \in \text{dom } h \times \mathbb{R}^n$  such that  $\bar{v} \in \nabla f(\bar{z}) + \partial h(\bar{z})$  and  $\|\bar{v}\| \leq \rho$ , in at most  $\mathcal{O}(M(\phi(z_0) - \phi_*)/\rho^2)$  composite-type iterations where  $z_0$  is the initial point. On the other hand, the AIPP method finds such solution in at most

$$(6) \quad \mathcal{O} \left( \frac{\sqrt{Mm}}{\rho^2} \min \{ \phi(z_0) - \phi_*, md_0^2 \} + \sqrt{\frac{M}{m}} \log_1^+ \left( \frac{M}{m} \right) \right)$$

composite type-iterations where  $d_0$  denotes the distance of  $z_0$  to the set of optimal solutions of (5). Hence, its complexity is better than that for the composite gradient method by a factor of  $\sqrt{M/m}$ . The main advantage of the AIPP method is that its iteration-complexity bound has a lower dependence on  $M$ , i.e., it is  $\mathcal{O}(\sqrt{M})$  instead of the  $\mathcal{O}(M)$ -dependence of the composite gradient method. Hence, the use of the AIPP method instead of the composite gradient method to solve (4) (whose associated  $M = \mathcal{O}(c)$ ) in the scheme outlined above is both theoretically and computationally appealing.

*Related works.* Under the assumption that domain of  $\phi$  is bounded, [9] presents an ACG method applied directly to (5) which obtains a  $\rho$ -approximate solution of (5) in

$$(7) \quad \mathcal{O} \left( \frac{MmD_h^2}{\rho^2} + \left( \frac{Md_0}{\rho} \right)^{2/3} \right)$$

where  $D_h$  denotes the diameter of the domain of  $h$ . Motivated by [9], other papers have proposed ACG methods for solving (5) under different assumptions on the functions  $g$  and  $h$  (see for example [5, 7, 10, 19, 28]). In particular, their analyses exploit the

lower curvature  $m$  and the work [5], which assumes  $h = 0$ , establishes a complexity which depends on  $\sqrt{M} \log M$  instead of  $M$  as in [9]. As in the latter work, our AIPP method also uses the idea of solving a sequence of convex proximal subproblems by an ACG method, but solves them in a more relaxed manner and, as a result, achieves the complexity bound (6) which improves the one in [5] by a factor of  $\log(M/\rho)$ . It should be noted that the second complexity bound in (6) in terms of  $d_0$  is new in the context of the composite nonconvex problem (5) and follows as a special case of a more general bound, namely (61), which actually unifies both bounds in (6). Moreover, in contrast to the analysis of [9], ours does not assume that  $D_h$  is finite. Also, inexact proximal point methods and HPE variants of the ones studied in [21, 30] for solving convex-concave saddle point problems and monotone variational inequalities, which inexactly solve a sequence of proximal subproblems by means of an ACG variant, were previously proposed by [11, 12, 16, 22, 27]. The behavior of an accelerated gradient method near saddle points of unconstrained instances of (5) (i.e., with  $h = 0$ ) is studied in [24].

Finally, complexity analysis of first-order quadratic penalty methods for solving special convex instances of (1) where  $h$  is an indicator function was first studied in [17] and further analyzed in [4, 20, 23]. Papers [18, 29] study the iteration-complexity of first-order augmented Lagrangian methods for solving the latter class of convex problems. The authors are not aware of earlier papers dealing with complexity analysis of quadratic penalty methods for solving nonconvex constrained optimization problems. However, [14] studies the complexity of a proximal augmented Lagrangian method for solving nonconvex instances of (1) under the very strong assumption that  $\nabla f$  is Lipschitz continuous everywhere and  $h = 0$ .

*Organization of the paper.* Subsection 1.1 contains basic definitions and notation used in the paper. Section 2 is divided into two subsections. The first one introduces the composite nonconvex optimization problem and discusses some approximate solutions criteria. The second subsection is devoted to the study of a general inexact proximal point framework to solve nonconvex optimization problems. In this subsection, we also show that a composite gradient method can be seen as an instance of the latter framework. Section 3 is divided into two subsections. The first one reviews an ACG method and its properties. Subsection 3.2 presents the AIPP method and its iteration-complexity analysis. Section 4 states and analyzes the QP-AIPP method for solving linearly constrained nonconvex composite optimization problems. Section 5 presents computational results. Section 6 gives some concluding remarks. Finally, the appendix gives the proofs of some technical results needed in our presentation.

**1.1. Basic definitions and notation.** This subsection provides some basic definitions and notation used in this paper.

The set of real numbers is denoted by  $\mathbb{R}$ . The set of non-negative real numbers and the set of positive real numbers are denoted by  $\mathbb{R}_+$  and  $\mathbb{R}_{++}$ , respectively. We let  $\mathbb{R}_{++}^2 := \mathbb{R}_{++} \times \mathbb{R}_{++}$ . Let  $\mathbb{R}^n$  denote the standard  $n$ -dimensional Euclidean space with inner product and norm denoted by  $\langle \cdot, \cdot \rangle$  and  $\|\cdot\|$ , respectively. For  $t > 0$ , define  $\log_1^+(t) := \max\{\log t, 1\}$ . The diameter of a set  $D \subset \mathbb{R}^n$  is defined as  $\sup\{\|z - z'\| : z, z' \in D\}$ .

Let  $\psi : \mathbb{R}^n \rightarrow (-\infty, +\infty]$  be given. The effective domain of  $\psi$  is denoted by  $\text{dom } \psi := \{x \in \mathbb{R}^n : \psi(x) < \infty\}$  and  $\psi$  is proper if  $\text{dom } \psi \neq \emptyset$ . Moreover, a proper function  $\psi : \mathbb{R}^n \rightarrow (-\infty, +\infty]$  is  $\mu$ -strongly convex for some  $\mu \geq 0$  if

$$\psi(\alpha z + (1 - \alpha)u) \leq \alpha\psi(z) + (1 - \alpha)\psi(u) - \frac{\alpha(1 - \alpha)\mu}{2}\|z - u\|^2$$

for every  $z, u \in \text{dom } \psi$  and  $\alpha \in [0, 1]$ . If  $\psi$  is differentiable at  $\bar{z} \in \mathbb{R}^n$ , then its affine approximation  $\ell_\psi(\cdot; \bar{z})$  at  $\bar{z}$  is defined as

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

Also, for  $\varepsilon \geq 0$ , its  $\varepsilon$ -subdifferential at  $z \in \text{dom } \psi$  is denoted by

$$(9) \quad \partial_\varepsilon \psi(z) := \{v \in \mathbb{R}^n : \psi(u) \geq \psi(z) + \langle v, u - z \rangle - \varepsilon, \forall u \in \mathbb{R}^n\}.$$

The subdifferential of  $\psi$  at  $z \in \text{dom } \psi$ , denoted by  $\partial \psi(z)$ , corresponds to  $\partial_0 \psi(z)$ . The set of all proper lower semi-continuous convex functions  $\psi : \mathbb{R}^n \rightarrow (-\infty, +\infty]$  is denoted by  $\overline{\text{Conv}}(\mathbb{R}^n)$ .

The proof of the following result can be found in [13, Proposition 4.2.2].

**PROPOSITION 1.** *Let  $\psi : \mathbb{R}^n \rightarrow (-\infty, +\infty]$ ,  $z, \bar{z} \in \text{dom } \psi$  and  $v \in \mathbb{R}^n$  be given and assume that  $v \in \partial \psi(z)$ . Then,  $v \in \partial_\varepsilon \psi(\bar{z})$  where  $\varepsilon = \psi(\bar{z}) - \psi(z) - \langle v, \bar{z} - z \rangle \geq 0$ .*

**2. Inexact proximal point method for nonconvex optimization.** This section contains two subsections. The first one states the composite nonconvex optimization (CNO) problem and discusses some notions of approximate solutions. The second subsection proposes and analyzes a general framework for solving nonconvex optimization problems and shows under very mild conditions that the composite gradient method is an instance of the general framework.

**2.1. The CNO problem and corresponding approximate solutions.** This subsection describes the CNO problem which will be the main subject of our analysis in Subsection 3.2. It also describes different notions of approximate solutions for the CNO problem and discusses their relationship.

The CNO problem we are interested in is (5) where the following conditions are assumed to hold:

(A1)  $h \in \overline{\text{Conv}}(\mathbb{R}^n)$ ;

(A2)  $g$  is a differentiable function on  $\text{dom } h$  which, for some  $M \geq m > 0$ , satisfies

$$(10) \quad -\frac{m}{2} \|u - z\|^2 \leq g(u) - \ell_g(u; z) \leq \frac{M}{2} \|u - z\|^2 \quad \forall z, u \in \text{dom } h;$$

(A3)  $\phi_* > -\infty$ .

We now make a few remarks about the above assumptions. First, if  $\nabla g$  is assumed to be  $M$ -Lipschitz continuous, then (10) holds with  $m = M$ . However, our interest is in the case where  $0 < m \ll M$  since this case naturally arises in the context of penalty methods for solving linearly constrained composite nonconvex optimization problems as will be seen in Section 4. Second, it is well-known that a necessary condition for  $z^* \in \text{dom } h$  to be a local minimum of (5) is that  $z^*$  be a stationary point of  $g + h$ , i.e.,  $0 \in \nabla g(z^*) + \partial h(z^*)$ .

The latter inclusion motivates the following notion of approximate solution for problem (5): for a given tolerance  $\hat{\rho} > 0$ , a pair  $(\hat{z}, \hat{v})$  is called a  $\hat{\rho}$ -approximate solution of (5) if

$$(11) \quad \hat{v} \in \nabla g(\hat{z}) + \partial h(\hat{z}), \quad \|\hat{v}\| \leq \hat{\rho}.$$

Another notion of approximate solution that naturally arises in our analysis of the general framework of Subsection 2.2 is as follows. For a given tolerance pair  $(\bar{\rho}, \bar{\varepsilon}) \in \mathbb{R}_{++}^2$ , a quintuple  $(\lambda, z^-, z, w, \varepsilon) \in \mathbb{R}_{++} \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}_+$  is called a  $(\bar{\rho}, \bar{\varepsilon})$ -prox-approximate solution of (5) if

$$(12) \quad w \in \partial_\varepsilon \left( \phi + \frac{1}{2\lambda} \|\cdot - z^-\|^2 \right) (z), \quad \left\| \frac{1}{\lambda} (z^- - z) + w \right\| \leq \bar{\rho}, \quad \varepsilon \leq \bar{\varepsilon}.$$

Note that the first definition of approximate solution above depends on the composite structure  $(g, h)$  of  $\phi$  but the second one does not.

The next proposition, whose proof is presented in Appendix A, shows how an approximate solution as in (11) can be obtained from a prox-approximate solution by performing a composite gradient step.

**PROPOSITION 2.** *Let  $h \in \overline{\text{Conv}}(\mathfrak{R}^n)$  and  $g$  be a differentiable function on  $\text{dom } h$  whose gradient satisfies the second inequality in (10). Let  $(\bar{\rho}, \bar{\varepsilon}) \in \mathfrak{R}_{++}^2$  and a  $(\bar{\rho}, \bar{\varepsilon})$ -prox-approximate solution  $(\lambda, z^-, z, w, \varepsilon)$  be given and define*

$$(13) \quad z_g := \operatorname{argmin}_u \left\{ \ell_g(u; z) + h(u) + \frac{M + \lambda^{-1}}{2} \|u - z\|^2 \right\},$$

$$(14) \quad q_g := [M + \lambda^{-1}](z - z_g),$$

$$(15) \quad \delta_g := h(z) - h(z_g) - \langle q_g - \nabla g(z), z - z_g \rangle,$$

$$(16) \quad v_g := q_g + \nabla g(z_g) - \nabla g(z).$$

Then, the following statements hold:

(a)  $q_g \in \nabla g(z) + \partial h(z_g)$  and

$$(M + \lambda^{-1}) \|z - z_g\| = \|q_g\| \leq \bar{\rho} + \sqrt{2\bar{\varepsilon}(M + \lambda^{-1})};$$

(b)  $\delta_g \geq 0$ ,  $q_g \in \nabla g(z) + \partial_{\delta_g} h(z)$  and

$$\|q_g\|^2 + 2(M + \lambda^{-1})\delta_g \leq \left[ \bar{\rho} + \sqrt{2\bar{\varepsilon}(M + \lambda^{-1})} \right]^2;$$

(c) if  $\nabla g$  is  $M$ -Lipschitz continuous on  $\text{dom } h$ , then

$$v_g \in \nabla g(z_g) + \partial h(z_g), \quad \|v_g\| \leq 2\|q_g\| \leq 2 \left[ \bar{\rho} + \sqrt{2\bar{\varepsilon}(M + \lambda^{-1})} \right].$$

Proposition 2 shows that a prox-approximate solution yields three possible ways of measuring the quality of an approximate solution of (5). Note that the ones described in (a) and (b) do not assume  $\nabla g$  to be Lipschitz continuous while the one in (c) does. This paper only derives complexity results with respect to prox-approximate solutions and approximate solutions as in (c) but we remark that complexity results for the ones in (a) or (b) can also be obtained. Finally, we note that Lemma 20 in Appendix A provides an alternative way of constructing approximate solutions as in (a), (b) or (c) from a given prox-approximate solution.

**2.2. A general inexact proximal point framework.** This subsection introduces a general inexact proximal point (GIPP) framework for solving the CNO problem (5).

Although our main goal is to use the GIPP framework in the context of the CNO problem, we will describe it in the context of the following more general problem

$$(17) \quad \phi_* := \inf \{ \phi(z) : z \in \mathfrak{R}^n \}$$

where  $\phi : \mathfrak{R}^n \rightarrow (-\infty, \infty]$  is a proper lower semi-continuous function, and  $\phi_* > -\infty$ .

We now state the GIPP framework for computing prox-approximate solutions of (17).

---

GIPP Framework

---

- (0) Let  $\sigma \in (0, 1)$  and  $z_0 \in \text{dom } \phi$  be given, and set  $k = 1$ ;  
 (1) find a quadruple  $(\lambda_k, z_k, \tilde{v}_k, \tilde{\varepsilon}_k) \in \mathfrak{R}_{++} \times \mathfrak{R}^n \times \mathfrak{R}^n \times \mathfrak{R}_+$  satisfying

$$(18) \quad \tilde{v}_k \in \partial_{\tilde{\varepsilon}_k} \left( \lambda_k \phi + \frac{1}{2} \|\cdot - z_{k-1}\|^2 \right) (z_k),$$

$$(19) \quad \|\tilde{v}_k\|^2 + 2\tilde{\varepsilon}_k \leq \sigma \|z_{k-1} - z_k + \tilde{v}_k\|^2;$$

- (2) set  $k \leftarrow k + 1$  and go to (1).

Observe that GIPP framework is not a well-specified algorithm but rather a conceptual framework consisting of (possibly many) specific instances. In particular, it does not specify how the quadruple  $(\lambda_k, z_k, \tilde{v}_k, \tilde{\varepsilon}_k)$  is computed and whether it exists. These two issues will depend on the specific instance under consideration and the properties assumed about problem (17). In this paper, we will discuss two specific instances of the above GIPP framework for solving (5), namely, the composite gradient method briefly discussed at the end of this subsection and an accelerated proximal method presented in Subsection 3.2. In both of these instances, the sequences  $\{\tilde{v}_k\}$  and  $\{\tilde{\varepsilon}_k\}$  are non-trivial (see Proposition 7 and Lemma 12(c)).

Let  $\{(\lambda_k, z_k, \tilde{v}_k, \tilde{\varepsilon}_k)\}$  be the sequence generated by an instance of the GIPP framework and consider the sequences  $\{(r_k, v_k, \varepsilon_k)\}$  defined as

$$(20) \quad (v_k, \varepsilon_k) := \frac{1}{\lambda_k} (\tilde{v}_k, \tilde{\varepsilon}_k), \quad r_k := \frac{z_{k-1} - z_k}{\lambda_k}.$$

Then, it follows from (18) that the quintuple  $(\lambda, \hat{z}, z, v, \varepsilon) = (\lambda_k, z_{k-1}, z_k, v_k, \varepsilon_k)$  satisfies the inclusion in (12) for every  $k \geq 1$ . In what follows, we will derive the iteration complexity for the quintuple  $(\lambda_k, z_{k-1}, z_k, v_k, \varepsilon_k)$  to satisfy: i) the first inequality in (12) only, namely,  $\|v_k + r_k\| \leq \bar{\rho}$ ; and ii) both inequalities in (12), namely,  $\|v_k + r_k\| \leq \bar{\rho}$  and  $\varepsilon_k \leq \bar{\varepsilon}$ , and hence a  $(\bar{\rho}, \bar{\varepsilon})$ -prox-approximate solution of (5).

Without necessarily assuming that the error condition (19) holds, the following technical but straightforward result derives bounds on  $\tilde{\varepsilon}_k$  and  $\|\tilde{v}_k + z_{k-1} - z_k\|$  in terms of the quantities

$$(21) \quad \delta_k = \delta_k(\sigma) := \frac{1}{\lambda_k} \max \{0, \|\tilde{v}_k\|^2 + 2\tilde{\varepsilon}_k - \sigma \|z_{k-1} - z_k + \tilde{v}_k\|^2\}, \quad \Lambda_k := \sum_{i=1}^k \lambda_i$$

where  $\sigma \in [0, 1)$  is a given parameter. Note that if (19) is assumed then  $\delta_k = 0$ .

LEMMA 3. Assume that the sequence  $\{(\lambda_k, z_k, \tilde{v}_k, \tilde{\varepsilon}_k)\}$  satisfies (18) and let  $\sigma \in (0, 1)$  be given. Then, for every  $k \geq 1$ , there holds

$$(22) \quad \frac{1}{\sigma \lambda_k} (\|\tilde{v}_k\|^2 + 2\tilde{\varepsilon}_k - \lambda_k \delta_k) \leq \frac{1}{\lambda_k} \|z_{k-1} - z_k + \tilde{v}_k\|^2 \leq \frac{2[\phi(z_{k-1}) - \phi(z_k)] + \delta_k}{1 - \sigma}$$

where  $\delta_k$  is as in (21).

*Proof.* First note that the inclusion in (18) is equivalent to

$$\lambda_i \phi(z) + \frac{1}{2} \|z - z_{i-1}\|^2 \geq \lambda_i \phi(z_i) + \frac{1}{2} \|z_i - z_{i-1}\|^2 + \langle \tilde{v}_i, z - z_i \rangle - \tilde{\varepsilon}_i \quad \forall z \in \mathfrak{R}^n.$$

Setting  $z = z_{i-1}$  in the above inequality and using the definition of  $\delta_i$  given in (21), we obtain

$$\begin{aligned} \lambda_i (\phi(z_{i-1}) - \phi(z_i)) &\geq \frac{1}{2} (\|z_{i-1} - z_i\|^2 + 2 \langle \tilde{v}_i, z_{i-1} - z_i \rangle - 2\tilde{\varepsilon}_i) \\ &= \frac{1}{2} [\|z_{i-1} - z_i + \tilde{v}_i\|^2 - \|\tilde{v}_i\|^2 - 2\tilde{\varepsilon}_i] \geq \frac{1}{2} [(1 - \sigma) \|z_{i-1} - z_i + \tilde{v}_i\|^2 - \lambda_i \delta_i] \end{aligned}$$

and hence the proof of the second inequality in (22) follows after simple rearrangements. The first inequality in (22) follows immediately from (21).  $\square$

LEMMA 4. *Let  $\{(\lambda_k, z_k, \tilde{v}_k, \tilde{\varepsilon}_k)\}$  be generated by an instance of the GIPP framework. Then, for every  $u \in \mathbb{R}^n$ , there holds*

$$\phi(z_k) \leq \phi(u) + \frac{1}{2(1-\sigma)\lambda_k} \|z_{k-1} - u\|^2, \quad \forall k \geq 1.$$

*Proof.* Using a simple algebraic manipulation, it is easy to see that (19) yields

$$(23) \quad \langle \tilde{v}_k, z_k - z_{k-1} \rangle + \frac{1}{\sigma} \tilde{\varepsilon}_k - \frac{1}{2} \|z_{k-1} - z_k\|^2 \leq -\frac{1-\sigma}{2\sigma} \|\tilde{v}_k\|^2.$$

Now, letting  $\theta := (1-\sigma)/\sigma > 0$ , recalling definition (9), using (18) and (23), and the fact that  $\langle v, v' \rangle \leq (\theta/2)\|v\|^2 + (1/2\theta)\|v'\|^2$  for all  $v, v' \in \mathbb{R}^n$ , we conclude that

$$\begin{aligned} \lambda_k [\phi(z_k) - \phi(u)] &\leq \frac{1}{2} \|z_{k-1} - u\|^2 + \langle \tilde{v}_k, z_k - u \rangle + \tilde{\varepsilon}_k - \frac{1}{2} \|z_k - z_{k-1}\|^2 \\ &\leq \frac{1}{2} \|z_{k-1} - u\|^2 + \langle \tilde{v}_k, z_{k-1} - u \rangle - \frac{1-\sigma}{2\sigma} \|\tilde{v}_k\|^2 \\ &\leq \frac{1}{2} \|z_{k-1} - u\|^2 + \left( \frac{\theta}{2} \|\tilde{v}_k\|^2 + \frac{1}{2\theta} \|z_{k-1} - u\|^2 \right) - \frac{1-\sigma}{2\sigma} \|\tilde{v}_k\|^2 \end{aligned}$$

and hence that the conclusion of the lemma holds due to the definition of  $\theta$ .  $\square$

Let  $z_0 \in \mathbb{R}^n$ ,  $\sigma \in (0, 1)$ , and  $\lambda \geq 0$  be given and consider the following quantity

$$(24) \quad R(\phi; \lambda) := \inf \left\{ R(u; \phi, \lambda) := \frac{1}{2} \|z_0 - u\|^2 + (1-\sigma)\lambda[\phi(u) - \phi_*] : u \in \mathbb{R}^n \right\}$$

where  $\phi_*$  is as in (17). Clearly,  $R(u; \phi, \lambda) \in \mathbb{R}_+$  for all  $u \in \text{dom } h$  and  $R(\phi; \lambda) \in \mathbb{R}_+$ .

PROPOSITION 5. *Let  $\{(\lambda_k, z_k, \tilde{v}_k, \tilde{\varepsilon}_k)\}$  be generated by an instance of the GIPP framework. Then, the following statements hold:*

(a) *for every  $k \geq 1$ ,*

$$(25) \quad \frac{1-\sigma}{2\lambda_k} \|z_{k-1} - z_k + \tilde{v}_k\|^2 \leq \phi(z_{k-1}) - \phi(z_k);$$

(b) *for every  $k > 1$ , there exists  $i \leq k$  such that*

$$(26) \quad \frac{1}{\lambda_i^2} \|z_{i-1} - z_i + \tilde{v}_i\|^2 \leq \frac{2R(\phi; \lambda_1)}{(1-\sigma)^2 \lambda_1 (\Lambda_k - \lambda_1)}$$

where  $\Lambda_k$  and  $R(\cdot; \cdot)$  are as in (21) and (24), respectively.

*Proof.* (a) The proof of (25) follows immediately from (22) and the fact that (19) is equivalent to  $\delta_k = 0$ .

(b) It follows from definitions of  $\phi_*$  and  $R(\cdot; \cdot, \cdot)$  in (17) and (24), respectively, (25) and Lemma 4 with  $k = 1$  that for all  $u \in \mathbb{R}^n$ ,

$$\begin{aligned} \frac{R(u; \phi, \lambda_1)}{(1-\sigma)\lambda_1} &= \frac{1}{2(1-\sigma)\lambda_1} \|z_0 - u\|^2 + \phi(u) - \phi_* \geq \phi(z_1) - \phi_* \geq \sum_{i=2}^k [\phi(z_{i-1}) - \phi(z_i)] \\ &\geq (1-\sigma) \sum_{i=2}^k \frac{\|z_{i-1} - z_i + \tilde{v}_i\|^2}{2\lambda_i} \geq \frac{(1-\sigma)(\Lambda_k - \lambda_1)}{2} \min_{i \leq k} \frac{1}{\lambda_i^2} \|z_{i-1} - z_i + \tilde{v}_i\|^2 \end{aligned}$$

and hence that (26) holds in view of the definition of  $R(\cdot; \cdot)$  in (24).  $\square$

Proposition 5(a) shows that GIPP enjoys the descent property (25) which many frameworks and/or algorithms for solving (17) also share. It is worth noting that, under the assumption that  $\phi$  is a KL-function, frameworks and/or algorithms sharing this property have been developed for example in [1, 2, 6, 8] where it is shown that the generated sequence  $\{z_k\}$  converges to some stationary point of (17) with a well-characterized asymptotic (but not global) convergence rate, as long as  $\{z_k\}$  has an accumulation point.

The following result, which follows immediately from Proposition 5, considers the instances of the GIPP framework in which  $\{\lambda_k\}$  is constant. For the purpose of stating it, define

$$(27) \quad d_0 := \inf\{\|z_0 - z^*\| : z^* \text{ is an optimal solution of (17)}\}.$$

Note that  $d_0 < \infty$  if and only if (17) has an optimal solution in which case the above infimum can be replaced by a minimum in view of the first assumption following (17).

COROLLARY 6. *Let  $\{(\lambda_k, z_k, \tilde{v}_k, \tilde{\varepsilon}_k)\}$  be generated by an instance the GIPP framework in which  $\lambda_k = \lambda$  for every  $k \geq 1$ , and define  $\{(v_k, \varepsilon_k, r_k)\}$  as in (20). Then, the following statements hold:*

(a) *for every  $k > 1$ , there exists  $i \leq k$  such that*

$$(28) \quad \frac{1}{\lambda^2} \|z_{i-1} - z_i + \tilde{v}_i\|^2 \leq \frac{2R(\phi; \lambda)}{\lambda^2(1-\sigma)^2(k-1)} \leq \frac{\min\left\{2[\phi(z_0) - \phi_*], \frac{d_0^2}{(1-\sigma)\lambda}\right\}}{\lambda(1-\sigma)(k-1)}$$

*where  $R(\cdot; \cdot)$  and  $d_0$  are as in (24) and (27), respectively;*

(b) *for any given tolerance  $\bar{\rho} > 0$ , the GIPP generates a quintuple  $(z^-, z, \tilde{v}, \tilde{\varepsilon})$  such that  $\|z^- - z + \tilde{v}\| \leq \lambda\bar{\rho}$  in a number of iterations bounded by*

$$(29) \quad \left\lceil \frac{2R(\phi; \lambda)}{\lambda^2(1-\sigma)^2\bar{\rho}^2} + 1 \right\rceil.$$

*Proof.* (a) The proof of the first inequality follows immediately from Proposition 5(b) and the fact that  $\lambda_k = \lambda$  for every  $k \geq 1$ . Now, note that due to (24), we have  $R(\phi; \lambda) \leq R(z_0; \phi, \lambda) = (1-\sigma)\lambda[\phi(z_0) - \phi_*]$  and  $R(\phi; \lambda) \leq R(z^*; \phi, \lambda) = \|z_0 - z^*\|^2/2$  for every optimal solution  $z^*$  of (17). The second inequality now follows from the previous observation and the definition of  $d_0$  in (27).

(b) This statement follows immediately from the first inequality in (a).  $\square$

In the above analysis, we have assumed that  $\phi$  is quite general. On the other hand, the remaining part of this subsection assumes that  $\phi$  has the composite structure as in (5), i.e.,  $\phi = g + h$  where  $g$  and  $h$  satisfy conditions (A1)-(A3) of Subsection 2.1.

We now briefly discuss some specific instances of the GIPP framework. Recall that, for given stepsize  $\lambda > 0$  and initial point  $z_0 \in \text{dom } h$ , the composite gradient method for solving the CNO problem (5) computes recursively a sequence  $\{z_k\}$  given by

$$(30) \quad z_k = \underset{z}{\operatorname{argmin}} \left\{ \ell_g(z; z_{k-1}) + \frac{1}{2\lambda} \|z - z_{k-1}\|^2 + h(z) \right\}$$

where  $\ell_g(\cdot; \cdot)$  is defined in (8). Note that if  $h$  is the indicator function of a closed convex set then the above scheme reduces to the classical projected gradient method.

The following result, whose proof is given in Appendix B, shows that the composite gradient method with  $\lambda$  sufficiently small is a special case of the GIPP framework in which  $\lambda_k = \lambda$  for all  $k$ .



PROPOSITION 7. Let  $\{z_k\}$  be generated by the composite gradient method (30) with  $\lambda \leq 1/m$  and  $\lambda < 2/M$ , and define  $\tilde{v}_k := z_{k-1} - z_k$ ,  $\lambda_k := \lambda$  and

$$(31) \quad \tilde{\varepsilon}_k := \lambda \left[ g(z_k) - \ell_g(z_k; z_{k-1}) + \frac{1}{2\lambda} \|z_k - z_{k-1}\|^2 \right].$$

Then, the quadruple  $(\lambda_k, z_k, \tilde{v}_k, \tilde{\varepsilon}_k)$  satisfies the inclusion (18) with  $\phi = g + h$ , and the relative error condition (19) with  $\sigma := (\lambda M + 2)/4$ . Thus, the composite gradient method (30) can be seen as an instance of the GIPP framework.

Under the assumption that  $\lambda < 2/M$  and  $\nabla g$  is  $M$ -Lipschitz continuous, it is well-known that the composite gradient method obtains a  $\hat{\rho}$ -approximate solution in  $\mathcal{O}([\phi(z_0) - \phi_*]/(\lambda \hat{\rho}^2))$  iterations. On the other hand, under the assumption that  $\lambda \leq 1/M$  and  $\nabla g$  is  $M$ -Lipschitz continuous, we can easily see that the above result together with Corollary 6(b) imply that the composite gradient method obtains a  $\hat{\rho}$ -approximate solution in  $\mathcal{O}(R(\phi; \lambda)/(\lambda^2 \hat{\rho}^2))$  iterations.

We now make a few general remarks about our discussion in this subsection so far. First, the condition on the stepsize  $\lambda$  of Proposition 7 forces it to be  $\mathcal{O}(1/M)$  and hence quite small whenever  $M \gg m$ . Second, Corollary 6(b) implies that the larger  $\lambda$  is, the smaller the complexity bound (29) becomes. Third, letting  $\lambda_k = \lambda$  in the GIPP framework for some  $\lambda \leq 1/m$  guarantees that the function  $\lambda_k \phi + \|\cdot - z_{k-1}\|^2/2$  which appears in (18) is convex.

In the remaining part of this subsection, we briefly outline the ideas behind an accelerated instance of the GIPP framework which chooses  $\lambda = \mathcal{O}(1/m)$ . First, note that when  $\sigma = 0$ , (18) and (19) imply that  $(\tilde{v}_k, \tilde{\varepsilon}_k) = (0, 0)$  and

$$(32) \quad 0 \in \partial \left( \lambda_k \phi + \frac{1}{2} \|\cdot - z_{k-1}\|^2 \right) (z_k).$$

and hence that  $z_k$  is an optimal solution of the prox-subproblem

$$(33) \quad z_k = \operatorname{argmin}_z \left\{ \lambda_k \phi(z) + \frac{1}{2} \|z - z_{k-1}\|^2 \right\}.$$

More generally, assuming that (19) holds for some  $\sigma > 0$  gives us an interpretation of  $z_k$ , together with  $(\tilde{v}_k, \tilde{\varepsilon}_k)$ , as being an approximate solution of (33) where its (relative) accuracy is measured by the  $\sigma$ -criterion (19). Obtaining such an approximate solution is generally difficult unless the objective function of the prox-subproblem (33) is convex. This suggests choosing  $\lambda_k = \lambda$  for some  $\lambda \leq 1/m$  which, according to a remark in the previous paragraph, ensures that  $\lambda_k \phi + (1/2) \|\cdot - z_{k-1}\|^2$  is convex for every  $k$ , and then applying an ACG method to the (convex) prox-subproblem (33) to obtain  $z_k$  and a certificate pair  $(\tilde{v}_k, \tilde{\varepsilon}_k)$  satisfying (19). An accelerated prox-instance of the GIPP framework obtained in this manner will be the subject of Subsection 3.2.

**3. Accelerated gradient methods.** The main goal of this section is to present another instance of the GIPP framework where the triples  $(z_k, \tilde{v}_k, \tilde{\varepsilon}_k)$  are obtained by applying an ACG method to the subproblem (33). It contains two subsections. The first one reviews an ACG variant for solving a composite strongly convex optimization problem and discusses some well-known and new results for it which will be useful in the analysis of the accelerated GIPP instance. Subsection 3.2 presents the accelerated GIPP instance for solving (5) and derives its corresponding iteration-complexity bound.

### 3.1. Accelerated gradient method for strongly convex optimization.

This subsection reviews an ACG variant and its convergence properties for solving the following optimization problem

$$(34) \quad \min\{\psi(x) := \psi_s(x) + \psi_n(x) : x \in \mathbb{R}^n\}$$

where the following conditions are assumed to hold

- (B1)  $\psi_n : \mathbb{R}^n \rightarrow (-\infty, +\infty]$  is a proper, closed and  $\mu$ -strongly convex function with  $\mu \geq 0$ ;
- (B2)  $\psi_s$  is a convex differentiable function on  $\text{dom } \psi_n$  which, for some  $L > 0$ , satisfies  $\psi_s(u) - \ell_{\psi_s}(u; x) \leq L\|u - x\|^2/2$  for every  $x, u \in \text{dom } \psi_n$  where  $\ell_{\psi_s}(\cdot; \cdot)$  is defined in (8).

The ACG variant ([3, 12, 25, 26, 31]) for solving (34) is as follows.

---

#### ACG Method

---

- (0) Let a pair of functions  $(\psi_s, \psi_n)$  as in (34) and initial point  $x_0 \in \text{dom } \psi_n$  be given, and set  $y_0 = x_0$ ,  $A_0 = 0$ ,  $\Gamma_0 \equiv 0$  and  $j = 0$ ;
- (1) compute

$$\begin{aligned} A_{j+1} &= A_j + \frac{\mu A_j + 1 + \sqrt{(\mu A_j + 1)^2 + 4L(\mu A_j + 1)A_j}}{2L}, \\ \tilde{x}_j &= \frac{A_j}{A_{j+1}}x_j + \frac{A_{j+1} - A_j}{A_{j+1}}y_j, \quad \Gamma_{j+1} = \frac{A_j}{A_{j+1}}\Gamma_j + \frac{A_{j+1} - A_j}{A_{j+1}}\ell_{\psi_s}(\cdot; \tilde{x}_j), \\ y_{j+1} &= \arg\min_y \left\{ \Gamma_{j+1}(y) + \psi_n(y) + \frac{1}{2A_{j+1}}\|y - y_0\|^2 \right\}, \\ x_{j+1} &= \frac{A_j}{A_{j+1}}x_j + \frac{A_{j+1} - A_j}{A_{j+1}}y_{j+1}; \end{aligned}$$

- (2) compute

$$\begin{aligned} u_{j+1} &= \frac{y_0 - y_{j+1}}{A_{j+1}}, \\ \eta_{j+1} &= \psi(x_{j+1}) - \Gamma_{j+1}(y_{j+1}) - \psi_n(y_{j+1}) - \langle u_{j+1}, x_{j+1} - y_{j+1} \rangle; \end{aligned}$$

- (3) set  $j \leftarrow j + 1$  and go to (1).
- 

Some remarks about the ACG method follow. First, the main core and usually the common way of describing an iteration of the ACG method is as in step 1. Second, the extra sequences  $\{u_j\}$  and  $\{\eta_j\}$  computed in step 2 will be used to develop a stopping criterion for the ACG method when the latter is called as a subroutine in the context of the AIPP method stated in Subsection 3.2. Third, the ACG method in which  $\mu = 0$  is a special case of a slightly more general one studied by Tseng in [31] (see Algorithm 3 of [31]). The analysis of the general case of the ACG method in which  $\mu \geq 0$  was studied in [12, Proposition 2.3].

The next proposition summarizes the basic properties of the ACG method.

**PROPOSITION 8.** *Let  $\{(A_j, \Gamma_j, x_j, u_j, \eta_j)\}$  be the sequence generated by the ACG method applied to (34) where  $(\psi_s, \psi_n)$  is a given pair of data functions satisfying (B1) and (B2) with  $\mu \geq 0$ . Then, the following statements hold*

(a) for every  $j \geq 1$ , we have  $\Gamma_j \leq \psi_s$  and

$$(35) \quad \psi(x_j) \leq \min_x \left\{ \Gamma_j(x) + \psi_n(x) + \frac{1}{2A_j} \|x - x_0\|^2 \right\},$$

$$(36) \quad A_j \geq \frac{1}{L} \max \left\{ \frac{j^2}{4}, \left( 1 + \sqrt{\frac{\mu}{4L}} \right)^{2(j-1)} \right\};$$

(b) for every solution  $x^*$  of (34), we have

$$(37) \quad \psi(x_j) - \psi(x^*) \leq \frac{1}{2A_j} \|x^* - x_0\|^2 \quad \forall j \geq 1;$$

(c) for every  $j \geq 1$ , we have

$$(38) \quad u_j \in \partial_{\eta_j}(\psi_s + \psi_n)(x_j), \quad \|A_j u_j + x_j - x_0\|^2 + 2A_j \eta_j \leq \|x_j - x_0\|^2.$$

*Proof.* For the proofs of (a) and (b) see [12, Proposition 2.3].

(c) It follows from the optimality condition for  $y_j$  and the definition of  $u_j$  that  $u_j \in \partial(\Gamma_j + \psi_n)(y_j)$ , for all  $j \geq 1$ . Hence Proposition 1 yields

$$(39) \quad (\Gamma_j + \psi_n)(x) \geq (\Gamma_j + \psi_n)(x_j) + \langle u_j, x - x_j \rangle - \tilde{\eta}_j, \quad \forall x \in \mathbb{R}^n,$$

where  $\tilde{\eta}_j = (\Gamma_j + \psi_n)(x_j) - (\Gamma_j + \psi_n)(y_j) - \langle u_j, x_j - y_j \rangle \geq 0$ . Thus the inclusion in (38) follows from (39), the first statement in (a), and the fact that  $\eta_j = \tilde{\eta}_j + \psi_s(x_j) - \Gamma_j(x_j)$ . Now, in order to prove the inequality in (38), note that  $y_0 = x_0$  and that the definitions of  $u_j$  and  $\eta_j$  yield

$$(40) \quad \|A_j u_j + x_j - x_0\|^2 - \|x_j - x_0\|^2 = \|y_j - y_0\|^2 + 2\langle y_0 - y_j, x_j - y_0 \rangle$$

$$(41) \quad 2A_j \eta_j = 2A_j [\psi(x_j) - (\Gamma_j + \psi_n)(y_j)] + 2\langle y_0 - y_j, y_j - x_j \rangle.$$

Then adding the above two identities we obtain

$$\begin{aligned} \|A_j u_j + x_j - x_0\|^2 + 2A_j \eta_j - \|x_j - x_0\|^2 &= 2A_j [\psi(x_j) - (\Gamma_j + \psi_n)(y_j)] - \|y_j - y_0\|^2 \\ &\leq 2A_j \left[ \psi(x_j) - \left( (\Gamma_j + \psi_n)(y_j) + \frac{1}{2A_j} \|y_j - y_0\|^2 \right) \right]. \end{aligned}$$

Hence, the inequality in (38) follows from the last inequality, the definition of  $y_j$  and (35).  $\square$

The main role of the ACG variant of this subsection is to find an approximate solution  $z_k$  of the subproblem (18) together with a certificate pair  $(\tilde{v}_k, \tilde{\varepsilon}_k)$  satisfying (18) and (19). Indeed, since (33) with  $\lambda$  sufficiently small is a special case of (34), we can apply the ACG method with  $x_0 = z_{k-1}$  to obtain the triple  $(z_k, \tilde{v}_k, \tilde{\varepsilon}_k)$  satisfying (18) and (19).

The following result essentially analyzes the iteration-complexity to compute the aforementioned triple.

**LEMMA 9.** *Let  $\{(A_j, x_j, u_j, \eta_j)\}$  be the sequence generated by the ACG method applied to (34) where  $(\psi_s, \psi_n)$  is a given pair of data functions satisfying (B1) and (B2) with  $\mu \geq 0$ . Then, for any  $\sigma > 0$  and index  $j$  such that  $A_j \geq 2(1 + \sqrt{\sigma})^2/\sigma$ , we have*

$$(42) \quad \|u_j\|^2 + 2\eta_j \leq \sigma \|x_0 - x_j + u_j\|^2.$$

As a consequence, the ACG method obtains a triple  $(x, u, \eta) = (x_j, u_j, \eta_j)$  satisfying

$$u \in \partial_\eta(\psi_s + \psi_n)(x) \quad \|u\|^2 + 2\eta \leq \sigma \|x_0 - x + u\|^2$$

in at most  $\left\lceil 2\sqrt{2L}(1 + \sqrt{\sigma})/\sqrt{\sigma} \right\rceil$  iterations.

*Proof.* Using the triangle inequality for norms, the relation  $(a + b)^2 \leq 2(a^2 + b^2)$  for all  $a, b \in \mathfrak{R}$ , and the inequality in (38), we obtain

$$\begin{aligned} \|u_j\|^2 + 2\eta_j &\leq \max\{1/A_j^2, 1/(2A_j)\}(\|A_j u_j\|^2 + 4A_j \eta_j) \\ &\leq \max\{1/A_j^2, 1/(2A_j)\}(2\|A_j u_j + x_j - x_0\|^2 + 2\|x_j - x_0\|^2 + 4A_j \eta_j) \\ &\leq \max\{(2/A_j)^2, 2/A_j\}\|x_j - x_0\|^2 \leq \frac{\sigma}{(1 + \sqrt{\sigma})^2} \|x_j - x_0\|^2 \end{aligned}$$

where the last inequality is due to  $A_j \geq 2(1 + \sqrt{\sigma})^2/\sigma$ . On the other hand, the triangle inequality for norms and simple calculations yield

$$\|x_j - x_0\|^2 \leq (1 + \sqrt{\sigma})\|x_0 - x_j + u_j\|^2 + \left(1 + \frac{1}{\sqrt{\sigma}}\right) \|u_j\|^2.$$

Combining the previous estimates, we obtain

$$(43) \quad \|u_j\|^2 + 2\eta_j \leq \frac{\sigma}{1 + \sqrt{\sigma}} \|x_0 - x_j + u_j\|^2 + \frac{\sqrt{\sigma}}{1 + \sqrt{\sigma}} \|u_j\|^2$$

which easily implies (42). Now if  $j \geq \left\lceil 2\sqrt{2L}(1 + \sqrt{\sigma})/\sqrt{\sigma} \right\rceil$  then it follows from (36) that  $A_j \geq 2(1 + \sqrt{\sigma})^2/\sigma$  and hence, due to the first statement of the lemma, (42) holds. The last conclusion combined with the inclusion in (38) prove the last statement of the lemma.  $\square$

Note that Proposition 8 and Lemma 9 hold for any  $\mu \geq 0$ . On the other hand, the next two results hold only for  $\mu > 0$  and derive some important relations satisfied by two distinct iterates of the ACG method. They will be used later on in Subsection 3.2 to analyze the refinement phase (step 3) of the AIPP method stated there.

LEMMA 10. Let  $\{(A_j, x_j, u_j, \eta_j)\}$  be generated by the ACG method applied to (34) where  $(\psi_s, \psi_n)$  is a given pair of data functions satisfying (B1) and (B2) with  $\mu > 0$ . Then,

$$(44) \quad \left(1 - \frac{1}{\sqrt{A_j \mu}}\right) \|x^* - x_0\| \leq \|x_j - x_0\| \leq \left(1 + \frac{1}{\sqrt{A_j \mu}}\right) \|x^* - x_0\| \quad \forall j \geq 1,$$

where  $x^*$  is the unique solution of (34). As a consequence, for all indices  $i, j \geq 1$  such that  $A_i \mu > 1$ , we have

$$(45) \quad \|x_j - x_0\| \leq \left(\frac{1 + \frac{1}{\sqrt{A_j \mu}}}{1 - \frac{1}{\sqrt{A_i \mu}}}\right) \|x_i - x_0\|.$$

*Proof.* First note that condition (B1) combined with (34) imply that  $\psi$  is  $\mu$ -strongly convex. Hence, it follows from (37) that

$$\frac{\mu}{2}\|x_j - x^*\|^2 \leq \psi(x_j) - \psi(x^*) \leq \frac{1}{2A_j}\|x^* - x_0\|^2$$

and hence that

$$(46) \quad \|x_j - x^*\| \leq \frac{1}{\sqrt{A_j\mu}}\|x^* - x_0\|.$$

The inequalities

$$\|x^* - x_0\| - \|x_j - x^*\| \leq \|x_j - x_0\| \leq \|x_j - x^*\| + \|x^* - x_0\|,$$

which are due to the triangle inequality for norms, together with (46) clearly implies (44). The last statement of the lemma follows immediately from (44).  $\square$

As a consequence of Lemma 10, the following result obtains several important relations on certain quantities corresponding to two arbitrary iterates of the ACG method.

LEMMA 11. *Let  $\{(A_j, x_j, u_j, \eta_j)\}$  be generated by the ACG method applied to (34) where  $(\psi_s, \psi_n)$  is a given pair of data functions satisfying (B1) and (B2) with  $\mu > 0$ . Let  $i$  be an index such that  $A_i \geq \max\{8, 9/\mu\}$ . Then, for every  $j \geq i$ , we have*

$$(47) \quad \|x_j - x_0\| \leq 2\|x_i - x_0\|, \quad \|u_j\| \leq \frac{4}{A_j}\|x_i - x_0\|, \quad \eta_j \leq \frac{2}{A_j}\|x_i - x_0\|^2,$$

$$(48) \quad \|x_0 - x_j + u_j\| \leq \left(4 + \frac{8}{A_j}\right)\|x_0 - x_i + u_i\|, \quad \eta_j \leq \frac{8\|x_0 - x_i + u_i\|^2}{A_j}.$$

*Proof.* The first inequality in (47) follows from (45) and the assumption that  $A_i\mu \geq 9$ . Now, using the inequality in (38) and the triangle inequality for norms, we easily see that

$$\|u_j\| \leq \frac{2}{A_j}\|x_j - x_0\|, \quad \eta_j \leq \frac{1}{2A_j}\|x_j - x_0\|^2$$

which, combined with the first inequality in (47), prove the second and the third inequalities in (47). Noting that  $A_i \geq 8$  by assumption, Lemma 9 implies that (42) holds with  $\sigma = 1$  and  $j = i$ , and hence that

$$(49) \quad \|u_i\| \leq \|x_0 - x_i + u_i\|.$$

Using the triangle inequality, the first two inequalities in (47) and relation (49), we conclude that

$$\begin{aligned} \|x_0 - x_j + u_j\| &\leq \|x_0 - x_j\| + \|u_j\| \leq \left(2 + \frac{4}{A_j}\right)\|x_0 - x_i\| \\ &\leq \left(2 + \frac{4}{A_j}\right)(\|x_0 - x_i + u_i\| + \|u_i\|) \leq \left(4 + \frac{8}{A_j}\right)\|x_0 - x_i + u_i\|, \end{aligned}$$

and that the first inequality in (48) holds. Now, the last inequality in (47), combined with the triangle inequality for norms and the relation  $(a + b)^2 \leq 2(a^2 + b^2)$ , imply that

$$\eta_j \leq \frac{2}{A_j}\|x_0 - x_i\|^2 \leq \frac{4}{A_j}(\|x_0 - x_i + u_i\|^2 + \|u_i\|^2).$$

Hence, in view of (49), the last inequality in (48) follows.  $\square$

**3.2. The AIPP method.** This subsection introduces and analyzes the AIPP method to compute approximate solutions of the CNO problem (5). The main results of this subsection are Theorem 13 and Corollary 14 which analyze the iteration-complexity of the AIPP method to obtain approximate solutions of the CNO problem in the sense of (12) and (11), respectively.

We start by stating the AIPP method.

---

AIPP Method

---

- (0) Let  $z_0 \in \text{dom } h$ ,  $\sigma \in (0, 1)$ , a pair  $(m, M)$  satisfying (10), a scalar  $0 < \lambda \leq 1/(2m)$  and a tolerance pair  $(\bar{\rho}, \bar{\varepsilon}) \in \mathfrak{R}_{++}^2$  be given, and set  $k = 1$ ;
- (1) perform at least  $\lceil 6\sqrt{2\lambda M + 1} \rceil$  iterations of the ACG method started from  $z_{k-1}$  and with

$$(50) \quad \psi_s = \psi_s^k := \lambda g + \frac{1}{4} \|\cdot - z_{k-1}\|^2, \quad \psi_n = \psi_n^k := \lambda h + \frac{1}{4} \|\cdot - z_{k-1}\|^2$$

to obtain a triple  $(x, u, \eta) \in \mathfrak{R}^n \times \mathfrak{R}^n \times \mathfrak{R}_+$  satisfying

$$(51) \quad u \in \partial_\eta \left( \lambda(g + h) + \frac{1}{2} \|\cdot - z_{k-1}\|^2 \right) (x), \quad \|u\|^2 + 2\eta \leq \sigma \|z_{k-1} - x + u\|^2;$$

- (2) if

$$(52) \quad \|z_{k-1} - x + u\| \leq \frac{\lambda \bar{\rho}}{5},$$

then go to (3); otherwise set  $(z_k, \tilde{v}_k, \tilde{\varepsilon}_k) = (x, u, \eta)$ ,  $k \leftarrow k + 1$  and go to (1);

- (3) restart the previous call to the ACG method in step 1 to find an iterate  $(\tilde{x}, \tilde{u}, \tilde{\eta})$  satisfying (51) with  $(x, u, \eta)$  replaced by  $(\tilde{x}, \tilde{u}, \tilde{\eta})$  and the extra condition

$$(53) \quad \tilde{\eta}/\lambda \leq \bar{\varepsilon}$$

and set  $(z_k, \tilde{v}_k, \tilde{\varepsilon}_k) = (\tilde{x}, \tilde{u}, \tilde{\eta})$ ; finally, output  $(\lambda, z^-, z, w, \varepsilon)$  where

$$(z^-, z, w, \varepsilon) = (z_{k-1}, z_k, \tilde{v}_k/\lambda, \tilde{\varepsilon}_k/\lambda).$$


---

Some comments about the AIPP method are in order. First, the ACG iterations performed in steps 1 and 3 are referred to as the inner iterations of the AIPP method. Second, in view of the last statement of Lemma 9 with  $(\psi_s, \psi_n)$  given by (50), the ACG method obtains a triple  $(x, u, \eta)$  satisfying (51). Observe that Proposition 8(c) implies that every triple  $(x, u, \eta)$  generated by the ACG method satisfies the inclusion in (51) and hence only the inequality in (51) needs to be checked for termination. Third, the consecutive loops consisting of steps 1 and 2 (or, steps 1, 2 and 3 in the last loop) are referred to as the outer iterations of the AIPP method. In view of (51), they can be viewed as iterations of the GIPP framework applied to the CNO problem (5). Fourth, instead of running the ACG method by at least the constant number of iterations described in step 1, one could run the more practical variant which stops (usually, much earlier) whenever the second inequality in (51) is satisfied. We omit the tedious analysis and more complicated description of this AIPP variant, but

remark that its iteration complexity is the same as the one studied in this subsection. Finally, the last loop supplements steps 1 and 2 with step 3 whose goal is to obtain a triple  $(\tilde{x}, \tilde{u}, \tilde{\eta})$  with a possibly smaller  $\tilde{\eta}$  while preserving the quality of the quantity  $\|z_{k-1} - \tilde{x} + \tilde{u}\|$  which at its start is bounded by  $\lambda\bar{\rho}/5$  and, throughout its inner iterations, can be shown to be bounded by  $\lambda\bar{\rho}$  (see the first inequality in (58)).

The next proposition summarizes some facts about the AIPP method.

LEMMA 12. *The following statements about the AIPP method hold:*

- (a) *at every outer iteration, the call to the ACG method in step 1 finds a triple  $(x, u, \eta)$  satisfying (51) in at most*

$$(54) \quad \left\lceil \max \left\{ \frac{2(1 + \sqrt{\sigma})}{\sqrt{\sigma}}, 6 \right\} \sqrt{2\lambda M + 1} \right\rceil$$

*inner iterations;*

- (b) *at the last outer iteration, the extra number of ACG iterations to obtain the triple  $(\tilde{x}, \tilde{v}, \tilde{\eta})$  is bounded by*

$$(55) \quad \left\lceil 2\sqrt{2\lambda M + 1} \log_1^+ \left( \frac{2\bar{\rho}\sqrt{(2\lambda M + 1)\lambda}}{5\sqrt{\varepsilon}} \right) + 1 \right\rceil;$$

- (c) *the AIPP method is a special implementation of the GIPP framework in which  $\lambda_k = \lambda$  for every  $k \geq 1$ ;*

- (d) *the number of outer iterations performed by the AIPP method is bounded by*

$$(56) \quad \left\lceil \frac{25R(\phi; \lambda)}{(1 - \sigma)^2 \lambda^2 \bar{\rho}^2} + 1 \right\rceil$$

*where  $R(\cdot; \cdot)$  is as defined in (24).*

*Proof.* (a) First note that the function  $\psi_s$  defined in (50) satisfies condition (B2) of Subsection 3.1 with  $L = \lambda M + 1/2$ , in view of (10). Hence, it follows from the last statement of Lemma 9 that the ACG method obtains a triple  $(x, u, \eta)$  satisfying (51) in at most

$$(57) \quad \left\lceil \frac{2\sqrt{2}(1 + \sqrt{\sigma})}{\sqrt{\sigma}} \sqrt{\lambda M + 1/2} \right\rceil$$

inner iterations. Hence, (a) follows from the above conclusion and the fact that the ACG method performs at least  $\left\lceil 6\sqrt{2\lambda M + 1} \right\rceil$  inner iterations, in view of step 1.

(b) Consider the triple  $(x, u, \eta)$  obtained in step 1 during the last outer iteration of the AIPP method. In view of step 1, there exists an index  $i \geq \left\lceil 6\sqrt{2\lambda M + 1} \right\rceil$  such that  $(x, u, \eta)$  is the  $i$ -iterate of the ACG method started from  $x_0 = z_{k-1}$  applied to problem (34) with  $\psi_s$  and  $\psi_n$  as in (50). Noting that the functions  $\psi_n$  and  $\psi_s$  satisfy conditions (B1) and (B2) of Subsection 3.1 with  $\mu = 1/2$  and  $L = \lambda M + 1/2$  (see (10)) and using the above inequality on the index  $i$  and relation (36), we conclude that  $A_i \geq 18 = \max\{8, 9/\mu\}$ , and hence that  $i$  satisfies the assumption of Lemma 11. It then follows from (48), (52) and (36) that the continuation of the ACG method as in step 3 generates a triple  $(\tilde{x}, \tilde{u}, \tilde{\eta}) = (x_j, u_j, \eta_j)$  satisfying

$$(58) \quad \|z_{k-1} - \tilde{x} + \tilde{u}\| \leq \left( 4 + \frac{8}{A_j} \right) \frac{\lambda\bar{\rho}}{5} \leq \lambda\bar{\rho}, \quad \tilde{\eta} \leq \frac{8\lambda^2\bar{\rho}^2}{25A_j} \leq \frac{8L\lambda^2\bar{\rho}^2}{25\left(1 + \sqrt{\frac{\mu}{4L}}\right)^{2(j-1)}}.$$

Noting the stopping criterion (53) and using the last inequality above, the fact that  $\mu = 1/2$  and  $L = \lambda M + 1/2$ , and the relation that  $\log(1+t) \geq t/2$  for all  $t \in [0, 1]$ , we can easily see that (b) holds.

(c) This statement is obvious.

(d) This statement follows by combining (c), the stopping criterion (52), and Corollary 6(b) with  $\bar{\rho}$  replaced by  $\bar{\rho}/5$ .  $\square$

Next we state one of our main results of this paper which derives the iteration-complexity of the AIPP method to obtain prox-approximate solutions of the CNO problem in the sense of (12). Recall that the AIPP method assumes that  $\lambda \leq 1/(2m)$ .

**THEOREM 13.** *Under assumptions (A1)-(A3), the AIPP method terminates with a prox-solution  $(\lambda, z^-, z, w, \varepsilon)$  within*

$$(59) \quad \mathcal{O} \left\{ \sqrt{\lambda M + 1} \left[ \frac{R(\phi; \lambda)}{\lambda^2 \bar{\rho}^2} + \log_1^+ \left( \frac{\bar{\rho} \sqrt{(\lambda M + 1)\lambda}}{\sqrt{\bar{\varepsilon}}} \right) \right] \right\}$$

inner iterations where  $R(\cdot; \cdot)$  is as defined in (24).

*Proof.* It follows from the second statement following the AIPP method and the definition of  $(\lambda, z^-, z, w, \varepsilon)$  in step 3 that the quintuple  $(\lambda, z^-, z, w, \varepsilon)$  satisfies the inclusion in (12). Now, the bound in (59) follows by multiplying the bounds in Lemma 54(a) and (b), and adding the result to the bound in Lemma 54(d).

Before stating the next result, we make two remarks about the above result. First, even though our main interest is in the case where  $m \leq M$  (see assumption (A.2)), bound (59) also holds for the case in which  $m > M$ . Second, the AIPP version in which  $\lambda = 1/(2m)$  yields the the best complexity bound under the reasonable assumption that, inside the squared bracket in (59), the first term is larger than the second one.

The following result describes the inner iteration complexity of the AIPP method with  $\lambda = 1/(2m)$  to compute approximate solutions of (5) in the sense of (11).

**COROLLARY 14.** *Assume that (A1)-(A3) hold and let a tolerance  $\hat{\rho} > 0$  be given. Also, let  $(\lambda, z^-, z, w, \varepsilon)$  be the output obtained by the AIPP method with inputs  $\lambda = 1/(2m)$  and  $(\bar{\rho}, \bar{\varepsilon})$  defined as*

$$(60) \quad (\bar{\rho}, \bar{\varepsilon}) := \left( \frac{\hat{\rho}}{4}, \frac{\hat{\rho}^2}{32(M + 2m)} \right).$$

Then the following statements hold:

(a) the AIPP method terminates in at most

$$(61) \quad \mathcal{O} \left\{ \sqrt{\frac{M}{m}} \left( \frac{m^2 R(\phi; \lambda)}{\bar{\rho}^2} + \log_1^+ \left( \frac{M}{m} \right) \right) \right\}$$

inner iterations where  $R(\cdot; \cdot)$  is as in (24).

(b) if  $\nabla g$  is  $M$ -Lipschitz continuous, then the pair  $(\hat{z}, \hat{v}) = (z_g, v_g)$  computed according to (13) and (16) is a  $\hat{\rho}$ -approximate solution of (5), i.e., (11) holds.

*Proof.* (a) This statement follows immediately from Theorem 13 with  $\lambda = 1/(2m)$  and  $(\bar{\rho}, \bar{\varepsilon})$  as in (60) and the fact that  $m \leq M$  due to (A2).

(b) First note that Theorem 13 implies that the AIPP output  $(\lambda, z^-, z, w, \varepsilon)$  satisfies criterion (12) with  $(\bar{\rho}, \bar{\varepsilon})$  as in (60). Since (60) also implies that

$$\hat{\rho} = 2 \left[ \bar{\rho} + \sqrt{2\bar{\varepsilon}(M + 2m)} \right],$$



the conclusion of (b) follows from Proposition 2(c) and the fact that  $\lambda = 1/2m$ .  $\square$

We now make a few remarks about the iteration-complexity bound (61) and its relationship to two other ones obtained in the literature under the reasonable assumption that the term  $\mathcal{O}(1/\hat{\rho}^2)$  in (61) dominates the other one, i.e., bound (61) reduces to

$$\mathcal{O}\left(\frac{m\sqrt{Mm}R(\phi;\lambda)}{\hat{\rho}^2}\right).$$

First, using the definition of  $R(\phi;\lambda)$ , it is easy to see that the above bound is majorized by the one in (6) (see the proof of Corollary 6(a)). Second, since the iteration-complexity bound for the composite gradient method with  $\lambda = 1/M$  is  $\mathcal{O}(M(\phi(z_0) - \phi_*)/\hat{\rho}^2)$  (see the discussion following Proposition 7), we conclude that (6), and hence (61), is better than the first bound by a factor of  $(M/m)^{1/2}$ . Third, bound (6), and hence (61), is also better than the one established in Corollary 2 of [9] for an ACG method applied directly to the nonconvex problem (5), namely (7), by at least a factor of  $(M/m)^{1/2}$ . Note that the ACG method of [9] assumes that the diameter  $D_h$  of  $\text{dom } h$  is bounded while the AIPP method does not.

**4. The QP-AIPP method.** This section presents the QP-AIPP method for obtaining approximate solutions of the linearly constrained nonconvex composite optimization problem (1) in the sense of (2).

Throughout this section, it is assumed that (1) satisfies the following conditions:

- (C1)  $h \in \overline{\text{Conv}}(\mathbb{R}^n)$ ,  $A \neq 0$  and  $\mathcal{F} := \{z \in \text{dom } h : Az = b\} \neq \emptyset$ ;
- (C2)  $f$  is a differentiable function on  $\text{dom } h$  and there exist scalars  $0 < m_f \leq L_f$  such that for every  $u, z \in \text{dom } h$ ,

$$(62) \quad \|\nabla f(z) - \nabla f(u)\| \leq L_f \|z - u\|,$$

$$(63) \quad \frac{-m_f}{2} \|u - z\|^2 + \ell_f(u; z) \leq f(u);$$

- (C3) there exists  $\hat{c} \geq 0$  such that  $\hat{\varphi}_{\hat{c}} > -\infty$  where

$$(64) \quad \hat{\varphi}_{\hat{c}} := \inf_z \left\{ \varphi_c(z) := (f + h)(z) + \frac{\hat{c}}{2} \|Az - b\|^2 : z \in \mathbb{R}^n \right\}, \quad \forall c \in \mathbb{R};$$

We make two remarks about conditions (C1)-(C3). First, (C1) and (C3) imply that the optimal value of (1) is finite but not necessarily achieved. Second, (C3) is quite natural in the sense that the penalty approach underlying the QP-AIPP method would not make sense without it. Finally, (62) implies that

$$(65) \quad \frac{-L_f}{2} \|u - z\|^2 \leq f(u) - \ell_f(u; z) \leq \frac{L_f}{2} \|u - z\|^2, \quad \forall z, u \in \text{dom } h.$$

and hence that (63) automatically holds with  $m_f = L_f$ , i.e., (63) is redundant when  $m_f = L_f$ . Our analysis in this section also considers the case in which a scalar  $0 < m_f < M_f$  satisfying (63) is known.

Given a tolerance pair  $(\hat{\rho}, \hat{\eta}) \in \mathbb{R}_{++}^2$ , a triple  $(\hat{z}, \hat{v}, \hat{p})$  is said to be a  $(\hat{\rho}, \hat{\eta})$ -approximate solution of (1) if it satisfies (2). Clearly, a  $(\hat{\rho}, \hat{\eta})$ -approximate solution  $(\hat{z}, \hat{v}, \hat{p})$  for the case in which  $(\hat{\rho}, \hat{\eta}) = (0, 0)$  means that  $0 = \hat{v} \in \nabla f(\hat{z}) + \partial h(\hat{z}) + A^* \hat{p}$  and  $A\hat{z} = b$ , and hence that  $(\hat{z}, \hat{p})$  is a first-order stationary pair of (1).

The QP-AIPP method is essentially a quadratic penalty approach where the AIPP method is applied to the penalty subproblem (64) associated with (1) for a fixed  $c > 0$

or for  $c$  taking values on an increasing sequence  $\{c_k\}$  converging to infinity. Note that (64) is a particular case of (5) in which

$$(66) \quad g = g_c := f + \frac{c}{2} \|A(\cdot) - b\|^2.$$

Moreover, we easily see that (63) and (65) imply that  $\nabla g_c$  satisfies condition (10) with  $(m, M) = (m_f, L_f + c\|A\|^2)$ .

Lemmas 15 and 17 below describe how a  $\bar{\rho}$ -approximate solution of (64) in the sense of (11) yields a  $(\hat{\rho}, \hat{\eta})$ -approximate solution of (1) whenever  $c$  is sufficiently large. Lemma 16 introduces an important quantity associated with the penalized problem (64) which plays a fundamental role in expressing the inner iteration complexity of the QP-AIPP method stated below for the case in which  $\text{dom } h$  is not necessarily bounded (see Theorem 18). It also establishes a few technical inequalities involving this quantity, one of which plays an important role in the proof of Theorem 18 and the statement of Lemma 17.

LEMMA 15. *Let  $(c, \hat{\rho}) \in \mathbb{R}_{++}^2$  be given and let  $(\hat{z}, \hat{v})$  be a  $\hat{\rho}$ -approximate solution of (64) in the sense of (11) with  $g = g_c$  where  $g_c$  is as in (66). Then the triple  $(\hat{z}, \hat{v}, \hat{p})$  where  $\hat{p} := c(A\hat{z} - b)$  satisfies the inclusion and the first inequality in (2).*

*Proof.* Since  $(\hat{z}, \hat{v})$  is a  $\hat{\rho}$ -approximate solution of (64), we have  $\hat{v} \in \nabla g_c(\hat{z}) + \partial h(\hat{z})$  and  $\|\hat{v}\| \leq \hat{\rho}$ . Hence the result follows from the definition of  $\hat{p}$  and the fact that  $\nabla g_c(\hat{z}) = \nabla g(\hat{z}) + A^*(c(A\hat{z} - b)) = \nabla g(\hat{z}) + A^*\hat{p}$ .  $\square$

The above result is quite general in the sense that it holds for any  $c > 0$ . We will now show that, by choosing  $c$  sufficiently large, we can actually guarantee that  $(\hat{z}, \hat{v}, \hat{p})$  of Lemma 15 also satisfies the second inequality in (2) as long as  $(\hat{z}, \hat{v})$  is generated by an instance of the GIPP framework. We first establish the following technical result.

LEMMA 16. *Let  $\hat{c}$  be as in (C3) and define*

$$(67) \quad R_c(\lambda) := \inf\{R(u; \varphi_c, \lambda) : u \in \mathcal{F}\} \quad \forall (c, \lambda) \in \mathbb{R}_+^2$$

where  $\mathcal{F}$ ,  $R(\cdot; \cdot, \cdot)$  and  $\varphi_c$  are as defined in (C1), (24) and (64), respectively. Then, for every  $c \geq \hat{c}$  and  $\lambda \geq \hat{\lambda} \in \mathbb{R}_+$ , we have

$$(68) \quad 0 \leq R(u; \varphi_c, \lambda) \leq R(u; \varphi_{\hat{c}}, \hat{\lambda}) < \infty \quad \forall u \in \mathcal{F},$$

$$(69) \quad 0 \leq R_c(\lambda) \leq R_{\hat{c}}(\hat{\lambda}) < \infty.$$

Moreover, if (1) has an optimal solution  $z^*$ , then

$$(70) \quad R_c(\lambda) \leq \frac{1}{2} \|z_0 - z^*\|^2 + (1 - \sigma)\lambda[\hat{\varphi}_* - \hat{\varphi}_c]$$

where  $\hat{\varphi}_*$  denotes the optimum value of (1).

*Proof.* Using (64) and assumption (C3), it is easy to see that for every  $c \geq \hat{c}$ , we have  $\hat{\varphi}_c \geq \hat{\varphi}_{\hat{c}} > -\infty$  and  $\varphi_c(u) = \varphi_{\hat{c}}(u) = (f + h)(u)$  for every  $u \in \mathcal{F}$ . Hence, the conclusion of the lemma follows immediately from (67) and the definition of  $R(\cdot; \cdot, \cdot)$  in (24).  $\square$

We are now ready to describe the feasibility behavior of a GIPP instance applied to (64).

LEMMA 17. *Assume that  $\{(\lambda_k, z_k, \tilde{v}_k, \tilde{\varepsilon}_k)\}$  is a sequence generated by an instance of the GIPP framework with input  $\sigma \in (0, 1)$  and  $z_0 \in \text{dom } h$  and with  $\phi = \varphi_c$  for*

some  $c > \hat{c}$  where  $\hat{c}$  is as in (C3) and  $\varphi_c$  is as in (64). Also, let  $\hat{\eta} \in \mathfrak{R}_{++}$  be given and define

$$(71) \quad T_{\hat{\eta}}(\lambda) := \frac{2R_{\hat{c}}(\lambda)}{\hat{\eta}^2(1-\sigma)\lambda} + \hat{c} \quad \forall \lambda \in \mathfrak{R}_{++}$$

where  $R_{\hat{c}}(\cdot)$  is as defined in (67). Then for every  $\hat{z} \in \mathfrak{R}^n$  such that  $\varphi_c(\hat{z}) \leq \varphi_c(z_1)$ , we have

$$(72) \quad \|A\hat{z} - b\|^2 \leq \frac{[T_{\hat{\eta}}(\lambda_1) - \hat{c}]\hat{\eta}^2}{c - \hat{c}}.$$

As a consequence, if  $c \geq T_{\hat{\eta}}(\lambda_1)$  then

$$\|Az_k - b\| \leq \hat{\eta}, \quad \forall k \geq 1.$$

*Proof.* First note that the definitions of  $\varphi_c$  and  $\hat{\varphi}_c$  in (64) imply that for every  $c > 0$ ,

$$(73) \quad \varphi_c(u) = \varphi_{\hat{c}}(u) + \frac{c - \hat{c}}{2} \|Au - b\|^2 \geq \hat{\varphi}_{\hat{c}} + \frac{c - \hat{c}}{2} \|Au - b\|^2 \quad \forall u \in \mathfrak{R}^n.$$

Now, let  $\hat{z} \in \mathfrak{R}^n$  be such that  $\varphi_c(\hat{z}) \leq \varphi_c(z_1)$ . Lemma 4 with  $\phi = \varphi_c$  and  $k = 1$ , the previous inequality on  $\hat{z}$ , and (73) with  $u = \hat{z}$ , then imply that for every  $u \in \mathcal{F}$ ,

$$\frac{\|z_0 - u\|^2}{2(1-\sigma)\lambda_1} + \varphi_c(u) \geq \varphi_c(z_1) \geq \varphi_c(\hat{z}) \geq \hat{\varphi}_{\hat{c}} + \frac{c - \hat{c}}{2} \|A\hat{z} - b\|^2.$$

Since  $\varphi_c(u) = \varphi_{\hat{c}}(u)$  for every  $u \in \mathcal{F}$ , it then follows from the above inequality and the definition of  $R(\cdot; \cdot, \cdot)$  in (24) that

$$\|A\hat{z} - b\|^2 \leq \frac{2R(u; \varphi_{\hat{c}}, \lambda_1)}{(c - \hat{c})(1 - \sigma)\lambda_1} \quad \forall u \in \mathcal{F}.$$

Since the above inequality holds for every  $u \in \mathcal{F}$ , it then follows from (67) and the definition of  $T_{\hat{\eta}}(\cdot)$  in (71) that the first conclusion of the lemma holds. Now, since by assumption  $\{(\lambda_k, z_k, \tilde{v}_k, \tilde{\varepsilon}_k)\}$  is generated by an instance of the GIPP framework with  $\phi = \varphi_c$ , it follows from (25) that  $\varphi_c(z_k) \leq \varphi_c(z_1)$  for every  $k \geq 1$ , and hence that the second conclusion of the lemma follows immediately from the first one together with the assumption that  $c \geq T_{\hat{\eta}}(\lambda_1)$ .  $\square$

We now make some remarks about the above result. First, it does not assume that  $\mathcal{F}$ , and hence  $\text{dom } h$ , is bounded. Also, it does not even assume that (1) has an optimal solution. Second, it implies that all iterates (excluding the starting one) generated by an instance of the GIPP framework applied to (64) satisfy the feasibility requirement (i.e., the last inequality) in (2) as long as  $c$  is sufficiently large, i.e.,  $c \geq T_{\hat{\eta}}(\lambda_1)$  where  $T_{\hat{\eta}}(\cdot)$  is as in (71). Third, since the quantity  $R_{\hat{c}}(\lambda_1)$ , which appears in the definition of  $T_{\hat{\eta}}(\lambda_1)$  in (71), is difficult to estimate, a simple way of choosing a penalty parameter  $c$  such that  $c \geq T_{\hat{\eta}}(\lambda_1)$  is not apparent. The QP-AIPP method described below solves instead a sequence of penalized subproblems (64) for increasing values of  $c$  (i.e., updated according to  $c \leftarrow 2c$ ). Moreover, despite solving a sequence of penalized subproblems, it is shown that its overall ACG iteration complexity is the same as the one for the ideal method corresponding to solving (64) with  $c = T_{\hat{\eta}}(\lambda_1)$ .

We are ready to state the QP-AIPP method.

---

#### QP-AIPP Method

---

- (0) Let  $z_0 \in \text{dom } h$ ,  $\sigma \in (0, 1)$ ,  $L_f$  satisfying (65),  $m_f$  satisfying (63), and a tolerance pair  $(\hat{\rho}, \hat{\eta}) \in \mathbb{R}_{++}^2$  be given, and set

$$(74) \quad \lambda = \frac{1}{2m_f}, \quad c = \hat{c} + \frac{L_f}{\|A\|^2};$$

- (1) apply the AIPP method with inputs  $z_0, \sigma, \lambda$ ,

$$(75) \quad (m, M) = (m_f, L_f + c\|A\|^2),$$

and  $(\bar{\rho}, \bar{\varepsilon})$  as in (60) to find a  $(\bar{\rho}, \bar{\varepsilon})$ -prox approximate solution  $(\lambda, z^-, z, w, \varepsilon)$  of problem (5) (according to (12)) with  $g := g_c$  and  $g_c$  as in (66);

- (2) use  $(\lambda, z^-, z, w, \varepsilon)$  to compute  $(z_g, v_g)$  as in Proposition 2 with  $g = g_c$ ;  
 (3) if  $\|Az_g - b\| > \hat{\eta}$  then set  $c = 2c$  and go to (1); otherwise, stop and output  $(\hat{z}, \hat{v}, \hat{p}) = (z_g, v_g, c(Az_g - b))$ .
- 

Every loop of the QP-AIPP method invokes in its step 1 the AIPP method of Subsection 3.2 to compute a  $(\bar{\rho}, \bar{\varepsilon})$ -prox approximate solution of (5). The latter method in turn uses the ACG method of Subsection 3.1 as a subroutine in its implementation (see step 1 of the AIPP method). For simplicity, we refer to all ACG iterations performed during calls to the ACG method as inner iterations.

We now make a few remarks about the QP-AIPP method. First, it follows from Corollary 14(b) that the pair  $(z_g, v_g)$  is a  $\hat{\rho}$ -approximate solution of (64) in the sense of (11) with  $g = g_c$  and  $g_c$  as in (66). As a consequence, Lemma 15 implies that the output  $(\hat{z}, \hat{v}, \hat{p})$  satisfies the inclusion and the first inequality in (2). Second, since  $(\lambda, z^-, z, w, \varepsilon)$  computed at step 1 is an iterate of the AIPP method, which in turn is a special instance of the GIPP framework, and  $\phi_c(z_g) \leq \phi_c(z)$  due to (91) of Lemma 19, we conclude from Lemma 17 that  $\hat{z} = z_g$  satisfies (72). Third, since every loop of the QP-AIPP method doubles  $c$ , the condition  $c > T_{\hat{\eta}}(\lambda_1)$  will be eventually satisfied. Hence, in view of the previous remark, the  $z_g$  corresponding to this  $c$  will satisfy the feasibility condition  $\|Az_g - b\| \leq \hat{\rho}$  and the QP-AIPP method will stop in view of its stopping criterion in step 3. Finally, in view of the first and third remarks, we conclude that the QP-AIPP method terminates in step 3 with a triple  $(\hat{z}, \hat{v}, \hat{p})$  satisfying (2).

The next result derives a bound on the overall number of inner iterations of the quadratic penalty AIPP method to obtain an approximate solution of (1) in the sense of (2).

THEOREM 18. *Let  $\hat{c}$  be as in (C1) and define*

$$(76) \quad \lambda := \frac{1}{2m_f}, \quad T_{\hat{\eta}} := \frac{2R_{\hat{c}}(\lambda)}{\hat{\eta}^2(1-\sigma)\lambda} + \hat{c}, \quad \Theta := \frac{L_f + T_{\hat{\eta}}\|A\|^2}{m_f}$$

where  $R_{\hat{c}}(\lambda)$  is as in (67). Then, the QP-AIPP method outputs a triple  $(\hat{z}, \hat{v}, \hat{p})$  satisfying (2) in a total number of inner iterations bounded by

$$(77) \quad \mathcal{O} \left( \sqrt{\Theta} \left[ \frac{m_f^2 R_{\hat{c}}(\lambda)}{\hat{\rho}^2} + \log_1^+(\Theta) \right] \right).$$

*Proof.* Let  $c_1 := \hat{c} + L_f/\|A\|^2$ . Noting the stopping criterion in step 3, using the second remark preceding the theorem and the fact that (74) implies  $c = c_l := 2^{l-1}c_1$

at the  $l$ -th loop of the QP-AIPP method, we conclude that the QP-AIPP method stops in at most  $\bar{l}$  loops where  $\bar{l}$  is the first index  $l \geq 1$  such that  $2^{l-1}c_1 > T_{\hat{\eta}}$ . We claim that

$$(78) \quad \sum_{l=1}^{\bar{l}} \sqrt{\frac{L_f + c_l \|A\|^2}{m_f}} \leq \mathcal{O} \left( \sqrt{\frac{L_f + T_{\hat{\eta}} \|A\|^2}{m_f}} \right) = \mathcal{O}(\sqrt{\Theta}).$$

Before establishing the above claim, we will use it to show that the conclusion of the theorem holds. Indeed, first note that the definition of  $c_1$  and the above definition of  $c_l$  imply that  $c_l \geq c_1 \geq \hat{c}$  for every  $l \geq 1$ . Hence, it follows from the second inequality in (69) with  $(c, \hat{\lambda}) = (c_l, \lambda)$  that  $R_{c_l}(\lambda) \leq R_{\hat{c}}(\lambda)$ . Since (24) and (67) easily imply that  $R(\varphi_{c_l}, \lambda) \leq R_{c_l}(\lambda)$ , we then conclude that  $R(\varphi_{c_l}, \lambda) \leq R_{\hat{c}}(\lambda)$ . The latter conclusion, (78) and Corollary 14(a) with  $\phi = \varphi_{c_l}$  and  $(m, M)$  as in (75) then imply that the number of inner iterations during the  $l$ -th loop of the QP-AIPP method is bounded by

$$(79) \quad \mathcal{O} \left( \sqrt{\frac{L_f + c_l \|A\|^2}{m_f}} \left[ \frac{m_f^2 R_{\hat{c}}(\lambda)}{\hat{\rho}^2} + \log_1^+(\Theta) \right] \right).$$

Hence, the total number of inner iterations performed by the QP-AIPP method is bounded by the sum of the previous bound over  $l = 1, \dots, \bar{l}$  which is equal to (77) in view of (78).

We will now show that (78) holds. If  $\bar{l} = 1$  then it follows from the definitions of  $\Theta$  in (76) and  $c_1$  in the beginning of the proof that

$$(80) \quad \frac{c_1 \|A\|^2}{m_f} = \frac{L_f + \hat{c} \|A\|^2}{m_f} \leq \frac{L_f + T_{\hat{\eta}} \|A\|^2}{m_f} = \Theta$$

and hence (78) holds. Consider now the case in which  $\bar{l} > 1$ . Using the fact that  $c_l = 2^{l-1}c_1$  together with the first equality in (80), we have that  $c_l \|A\|^2 / m_f \geq c_1 \|A\|^2 / m_f \geq L_f / m_f$  and hence

$$(81) \quad \sum_{l=1}^{\bar{l}} \sqrt{\frac{L_f + c_l \|A\|^2}{m_f}} \leq \sqrt{\frac{2c_1 \|A\|^2}{m_f}} \sum_{l=1}^{\bar{l}} (\sqrt{2})^{l-1} \leq \mathcal{O} \left( \sqrt{\frac{c_1 \|A\|^2}{m_f}} \sqrt{2}^{\bar{l}} \right).$$

Using the definition of  $\bar{l}$  and the fact that  $\bar{l} > 1$ , we easily see that  $2^{\bar{l}-1}c_1 \leq 2T_{\hat{\eta}}$  and hence that  $(\sqrt{2})^{\bar{l}} \leq 2(T_{\hat{\eta}}/c_1)^{1/2}$ . The last inequality together with (81) and the definition of  $\Theta$  in (76) then imply (78).  $\square$

Before ending this section, we make three remarks about Theorem 18. First, (70) implies the quantity  $R_{\hat{c}}(\lambda)$  admits the upper bound

$$R_{\hat{c}}(\lambda) \leq \frac{1}{2} \hat{d}_0^2 + (1 - \sigma) \lambda [\hat{\varphi}_* - \hat{\varphi}_{\hat{c}}]$$

where  $\hat{d}_0 := \{\|z_0 - z_*\| : z_* \text{ is an optimal solution of (1)}\}$ . Second, in terms of the tolerance pair  $(\hat{\rho}, \hat{\eta})$  only, the iteration-complexity bound (77) reduces to  $\mathcal{O}(1/(\hat{\rho}^2 \hat{\eta}))$  for an arbitrary initial point  $z_0 \in \text{dom } h$ . Third, the iteration-complexity bound (77) is almost the same as the one corresponding to the case in which  $T_{\hat{\eta}} = T_{\hat{\eta}}(\lambda)$  as in (71) is known and the penalty parameter is set to  $c = T_{\hat{\eta}}$ , namely,

$$\mathcal{O} \left( \sqrt{\Theta} \left[ \frac{m_f^2 R(\varphi_c, \lambda)}{\hat{\rho}^2} + \log_1^+(\Theta) \right] \right) = \mathcal{O} \left( \sqrt{\Theta} \left[ \frac{m_f^2 R_c(\lambda)}{\hat{\rho}^2} + \log_1^+(\Theta) \right] \right)$$

which follows as a consequence of Corollary 14 with  $M$  as in (75). Note that the two bounds differ only in that the quantity  $R_c(\lambda)$ , which appears in the above bound, may be strictly smaller than the quantity  $R_{\hat{c}}(\lambda)$  in (77) (see (69)).

**5. Computational Results.** The goal of this section is to present a few computational results that show the performance of the AIPP method, which would consequently assess the performance of the QP-AIPP method. In particular, while the experiments are limited in the sense that they do not directly test the actual behavior of the QP-AIPP method, they can be considered as examples of subproblems in the execution of the penalty-based method. The AIPP method is benchmarked against two other nonconvex optimization methods, namely the projected gradient (PG) method and the accelerated gradient (AG) method recently proposed and analyzed in [9].

Several instances of the quadratic programming (QP) problem

$$(82) \quad \min \left\{ g(z) := -\frac{\xi}{2} \|DBz\|^2 + \frac{\tau}{2} \|Az - b\|^2 : z \in \Delta_n \right\}$$

were considered where  $A \in \mathbb{R}^{l \times n}$ ,  $B \in \mathbb{R}^{n \times n}$ ,  $D \in \mathbb{R}^{n \times n}$  is a diagonal matrix,  $b \in \mathbb{R}^{l \times 1}$ ,  $(\xi, \tau) \in \mathbb{R}_{++}^2$ , and  $\Delta_n := \{z \in \mathbb{R}^n : \sum_{i=1}^n z_i = 1, z_i \geq 0\}$ . More specifically, we set the dimensions to be  $(l, n) = (20, 300)$ . We also generated the entries of  $A, B$  and  $b$  by sampling from the uniform distribution  $\mathcal{U}[0, 1]$  and the diagonal entries of  $D$  by sampling from the discrete uniform distribution  $\mathcal{U}\{1, 1000\}$ . By appropriately choosing the scalars  $\xi$  and  $\tau$ , the instance corresponding to a pair of parameters  $(M, m) \in \mathbb{R}_{++}^2$  was generated so that  $M = \lambda_{\max}(\nabla^2 g)$  and  $-m = \lambda_{\min}(\nabla^2 g)$  where  $\lambda_{\max}(\nabla^2 g)$  and  $\lambda_{\min}(\nabla^2 g)$  denote the largest and smallest eigenvalues of the Hessian of  $g$  respectively. The parameters  $(\lambda, \sigma)$  were set to be  $(0.9/m, 0.3)$ . The AIPP, PG, and AG methods were implemented in MATLAB 2016a scripts and were run on Linux 64-bit machines each containing Xeon E5520 processors and at least 8 GB of memory.

All three methods use the centroid of the set  $\Delta_n$  as the initial starting point  $z_0$  and were run until a pair  $(z, v)$  was generated satisfying the condition

$$(83) \quad v \in \nabla g(z) + N_{\Delta_n}(z), \quad \frac{\|v\|}{\|\nabla g(z_0)\| + 1} \leq \bar{\rho}$$

for a given tolerance  $\bar{\rho} > 0$ . Here,  $N_X(z)$  denotes the normal cone of  $X$  at  $z$ , i.e.  $N_X(z) = \{u \in \mathbb{R}^n : \langle u, \tilde{z} - z \rangle \leq 0, \forall \tilde{z} \in X\}$ . The results of the two tables below were obtained with  $\bar{\rho} = 10^{-7}$ . They present results for different choices of the curvature pair  $(M, m)$ . Each entry in the  $\bar{g}$ -column is the value of the objective function of (82) at the last iterate generated by each method. Since they are approximately the same for all three methods, only one value is reported. The bold numbers in each table highlight the algorithm that performed the most efficiently in terms of total number of iterations. It should be noted that both AIPP and PG methods perform a single projection step per iteration while the AG method performs two.

Size		$\bar{g}$	Iteration Count		
$M$	$m$		PG	AG	AIPP
16777216	16777216	-2.24E+05	5445	<b>374</b>	14822
16777216	1048576	-3.83E+04	7988	<b>4429</b>	6711
16777216	65536	-4.46E+02	91295	<b>22087</b>	24129
16777216	4096	4.07E+03	80963	26053	<b>5706</b>
16777216	256	4.38E+03	82029	20371	<b>1625</b>
16777216	16	4.40E+03	81883	20761	<b>2308</b>

**Table 1:** Numerical results with  $\sigma = 0.3$  and  $\lambda = 0.9/m$ 

Size		$\bar{g}$	Iteration Count		
$M$	$m$		PG	AG	AIPP
4000	1	9.68E-01	80560	24813	<b>5752</b>
16000	1	4.11E+00	77813	24861	<b>2830</b>
64000	1	1.67E+01	82000	20373	<b>1621</b>
256000	1	6.71E+01	81929	20767	<b>1942</b>
1024000	1	2.68E+02	81882	20761	<b>2297</b>
4096000	1	1.07E+03	81871	20759	<b>2083</b>

**Table 2:** Numerical results with  $\sigma = 0.3$  and  $\lambda = 0.9/m$ 

From the tables, we can conclude that if the curvature ratio  $M/m$  is sufficiently large then the AIPP method performs fewer iterations than the PG and the AG methods. This indicates that the QP-AIPP, which is based on the AIPP method, might be a promising approach towards solving linearly constrained nonconvex optimization problems. This is due to the fact that the curvature ratios of the penalty subproblems grow substantially as  $c$  increases and, as a result, AIPP can efficiently solve them. On the other hand, AIPP does not do well on instances whose associated curvature ratio is small. However, preliminary computational experiments seem to indicate that a variant of AIPP can also efficiently solve instances with small curvature ratios by significantly choosing  $\lambda$  much larger than  $0.9/m$ . Since this situation is not covered by the theory presented in this paper, we are not reporting these results in this paper, opting instead to leave this preliminary investigation for a future work.

**6. Concluding remarks.** Paper [14] proposed a linearized version of the augmented Lagrangian method to solve (1) but assumes the strong condition (among a few others) that  $h = 0$ , which most important problems arising in applications do not satisfy. To circumvent this technical issue, [15] proposed a penalty ADDM approach which introduces an artificial variable  $y$  in (1) and then penalizes  $y$  to obtain the penalized problem

$$(84) \quad \min \left\{ f(z) + h(z) + \frac{c}{2} \|y\|^2 : Ax + y = b \right\},$$

which is then solved by a two-block ADMM. Since (84) satisfies the assumption that its  $y$ -block objective function component has Lipschitz continuous gradient everywhere and its  $y$ -block coefficient matrix is the identity, an iteration-complexity of the two-block ADDM for solving (84), and hence (1), can be established. More specifically, it has been shown in Remark 4.3 of [15] that the overall number of composite gradient steps performed by the aforementioned two-block ADMM penalty scheme to obtain

a triple  $(\hat{z}, \hat{v}, \hat{p})$  satisfying (2) is bounded by  $\mathcal{O}(\hat{\rho}^{-6})$  under the assumptions that  $\hat{\eta} = \hat{\rho}$ , the level sets of  $f + h$  are bounded and the initial triple  $(z_0, y_0, p_0)$  satisfies  $(y_0, p_0) = (0, 0)$ ,  $Az_0 = b$  and  $z_0 \in \text{dom } h$ .

Note that the last complexity bound is derived under a boundedness assumption and is worse than the one obtained in this paper for the QP-AIPP method, namely  $\mathcal{O}(\hat{\rho}^{-2}\hat{\eta}^{-1})$ , without any boundedness assumption. Moreover, in contrast to the complexity of the QP-AIPP which is established for an arbitrary infeasible point  $z_0 \in \text{dom } h$ , the complexity bound of the aforementioned two-block ADMM penalty scheme assumes that  $z_0$  is feasible for (1). In fact, as far as we know, QP-AIPP is the first method for solving (1) from an infeasible starting point with a guaranteed complexity bound under the general assumptions considered in this paper.

### Appendix A. Proof of Proposition 2.

We first state two technical lemmas before giving the proof of Proposition 2.

LEMMA 19. Assume that  $h \in \overline{\text{Conv}}(\mathbb{R}^n)$ ,  $z \in \text{dom } h$  and  $f$  is a differentiable function on  $\text{dom } h$  which, for some  $L > 0$ , satisfies

$$(85) \quad f(u) - \ell_f(u; z) \leq \frac{L}{2} \|u - z\|^2, \quad \forall u \in \text{dom } h,$$

and define

$$(86) \quad z_f = z(z; f) := \operatorname{argmin}_u \left\{ \ell_f(u; z) + h(u) + \frac{L}{2} \|u - z\|^2 \right\},$$

$$(87) \quad q_f = q(z; f) := L(z - z_f),$$

$$(88) \quad \delta_f = \delta(z; f) := h(z) - h(z_f) - \langle q_f - \nabla f(z), z - z_f \rangle.$$

Then, there hold

$$(89) \quad q_f \in \nabla f(z) + \partial h(z_f), \quad q_f \in \nabla f(z) + \partial_{\delta_f} h(z), \quad \delta_f \geq 0,$$

$$(90) \quad (q_f, \delta_f) = \operatorname{argmin}_{(r, \varepsilon)} \left\{ \frac{1}{2L} \|r\|^2 + \varepsilon : r \in \nabla f(z) + \partial_\varepsilon h(z) \right\},$$

$$(91) \quad \delta_f + \frac{1}{2L} \|q_f\|^2 \leq (f + h)(z) - (f + h)(z_f).$$

*Proof.* We first show that (89) holds. The optimality condition for (86) and the definition of  $q_f$  in (87) immediately yield the first inclusion in (89). Hence, it follows from Proposition 1 and the definition of  $\delta_f$  in (88) that the second inclusion and the inequality in (89) also hold.

We now show that (90) holds. Clearly, the second inclusion in (89) implies that  $(q_f, \delta_f)$  is feasible to (90). Assume now that  $(r, \varepsilon)$  satisfies  $r \in \nabla f(z) + \partial_\varepsilon h(z)$ , or equivalently,

$$h(u) \geq h(z) + \langle r - \nabla f(z), u - z \rangle - \varepsilon \quad \forall u \in \mathbb{R}^n.$$

Using the above inequality with  $u = z_f$  and the definitions of  $q_f$  and  $\delta_f$  given in (87) and (88), respectively, we then conclude that

$$\begin{aligned} \delta_f + \frac{\|q_f\|^2}{2L} &= h(z) - h(z_f) - \langle \nabla f(z), z_f - z \rangle - \frac{L}{2} \|z - z_f\|^2 \\ &\leq -\langle r, z_f - z \rangle + \varepsilon - \frac{\|q_f\|^2}{2L} = \frac{1}{L} \langle r, q_f \rangle + \varepsilon - \frac{\|q_f\|^2}{2L} \\ &\leq \frac{1}{2L} \|r\|^2 + \frac{1}{2L} \|q_f\|^2 + \varepsilon - \frac{\|q_f\|^2}{2L} = \frac{1}{2L} \|r\|^2 + \varepsilon \end{aligned}$$



where the last inequality is due to Cauchy Schwarz inequality and  $2ab \leq a^2 + b^2$ , for every  $a, b \in \mathbb{R}$ . Hence (90) holds. Finally, to see that (91) holds, note that the last relation, the definition of  $\ell_f(\cdot; z)$  in (8) and inequality (85) with  $u = z_f$  imply that

$$\begin{aligned} \delta_f + \frac{\|q_f\|^2}{2L} &= (f+h)(z) - (f+h)(z_f) + [f(z_f) - \ell_f(z_f; z)] - \frac{L}{2}\|z - z_f\|^2 \\ &\leq (f+h)(z) - (f+h)(z_f). \end{aligned} \quad \square$$

LEMMA 20. Assume that  $h \in \overline{\text{Conv}}(\mathbb{R}^n)$ ,  $z \in \text{dom } h$  and  $g$  is a differentiable function on  $\text{dom } h$  which, for some  $M > 0$ , satisfies (85) with  $(f, L)$  replaced by  $(g, M)$ . Let  $\lambda > 0$ ,  $(z^-, z, w, \varepsilon) \in \mathbb{R}^n \times \text{dom } h \times \mathbb{R}^n \times \mathbb{R}_+$  and  $\rho > 0$  be such that

$$(92) \quad w \in \partial_\varepsilon \left( g + h + \frac{1}{2\lambda} \|\cdot - z^-\|^2 \right) (z), \quad \left\| \frac{1}{\lambda} (z^- - z) + w \right\| \leq \rho$$

and set

$$(93) \quad \begin{aligned} L &= M + \lambda^{-1}, \quad f(\cdot) = g(\cdot) + \frac{1}{2\lambda} \|\cdot - z^-\|^2 - \langle w, \cdot \rangle, \\ (z_f, q_f, \delta_f) &= (z(z; f), q(z; f), \delta(z; f)), \end{aligned}$$

where the quantities  $z(z; f)$ ,  $q(z; f)$  and  $\delta(z; f)$  are defined in (86), (87) and (88), respectively. Then, the following statements hold:

(a) the pair  $(v, \delta_f)$  where

$$(94) \quad v := q_f + \frac{z^- - z}{\lambda} + w,$$

satisfies

$$(95) \quad v \in \nabla g(z) + \partial_{\delta_f} h(z), \quad 0 \leq \delta_f \leq \varepsilon;$$

$$(96) \quad \|v\|^2 + 2(M + \lambda^{-1})\delta_f \leq \left[ \rho + \sqrt{2(M + \lambda^{-1})\varepsilon} \right]^2;$$

(b) if  $\nabla g$  is  $M$ -Lipschitz continuous, then the pair  $(z_f, v_f)$  where

$$(97) \quad v_f := v + \nabla g(z_f) - \nabla g(z)$$

satisfies

$$(98) \quad v_f \in \nabla g(z_f) + \partial h(z_f), \quad \|v_f\| \leq \rho + 2\sqrt{2(M + \lambda^{-1})\varepsilon}.$$

*Proof.* (a) First note that the pair  $(f, L)$  defined in (93) satisfies (85) and that  $\lambda$  and  $(z^-, z, w, \varepsilon)$  satisfy the inclusion in (92) if and only if  $0 \in \partial_\varepsilon (f+h)(z)$ , or equivalently,  $(f+h)(u) \geq (f+h)(z) - \varepsilon$  for every  $u$ . In particular, the latter inequality with  $u = z_f$  implies that  $(f+h)(z) - (f+h)(z_f) \leq \varepsilon$ . Hence, combining the last inequality, Lemma 19 with  $(f, L)$  as in (93), and the definition of  $v$  given in (94), we conclude that the relations in (95) hold and

$$(99) \quad \|q_f\|^2 + 2(M + \lambda^{-1})\delta_f \leq 2(M + \lambda^{-1})\varepsilon.$$

Now, the inequality in (92), definition of  $v$  in (94) and the triangle inequality for norms imply that  $\|v\| \leq \rho + \|q_f\|$ , and hence, in view of (99), that

$$\begin{aligned} \|v\|^2 + 2(M + \lambda^{-1})\delta_f &\leq [\|q_f\|^2 + 2(M + \lambda^{-1})\delta_f] + 2\rho\|q_f\| + \rho^2 \\ &\leq 2(M + \lambda^{-1})\varepsilon + 2\rho\sqrt{2(M + \lambda^{-1})\varepsilon} + \rho^2 \\ &= \left[\rho + \sqrt{2(M + \lambda^{-1})\varepsilon}\right]^2, \end{aligned}$$

showing that (96) holds.

(b) The inclusion in (98) follows immediately from the first inclusion in (89) and definition of  $v_f$  in (97). Finally, using the assumption that  $\nabla g$  is  $M$ -Lipschitz continuous, the triangle inequality for norms, definition of  $v_f$  and  $q_f$  in (97) and (87), respectively, we conclude that

$$\|v_f\| - \|v\| \leq \|v_f - v\| = \|\nabla g(z_f) - \nabla g(z)\| \leq M\|z_f - z\| = \frac{M}{M + \lambda^{-1}}\|q_f\| \leq \|q_f\|$$

and hence, in view of (96) and the inequality in (92), the inequality in (98) holds.  $\square$

Lemma 20 assumes that the inclusion in (12) holds, or equivalently, that the function  $f$  defined in (93) satisfies  $(f + h)(u) \geq (f + h)(z) - \varepsilon$  for every  $u$ . However, a close examination of its proof shows that the latter inequality is used only for  $u = z_f$ .

**Proof of Proposition 2.** First note that, since  $\nabla g$  satisfies the second inequality in (10), we see that (85) is satisfied with  $f = g$  and  $L = M$  (in particular  $L = M + \lambda^{-1}$ ). Moreover, the elements defined in (13), (14), and (15) correspond to (86), (87), and (88), respectively, with  $f$  replaced by  $g$  and  $L$  replaced by  $M + \lambda^{-1}$ . Hence, the inclusions in (a) and (b) as well as the first inequality in (b) follow immediately from (89). Now note that the equality in (a) follows immediately from the definition of  $q_g$  in (14). Moreover, the inequality in (b) implies the inequality in (a). Hence, let us proceed to prove the inequality in (b). It follows from Lemma 20 that the pair  $(v, \delta_f)$  as in (94) and (88) satisfies the inclusion in (95) and hence due to (90) with  $(f, L)$  replaced by  $(g, M + \lambda^{-1})$  and (96), we have

$$\|q_g\|^2 + 2(M + \lambda^{-1})\delta_g \leq \|v\|^2 + 2(M + \lambda^{-1})\delta_f \leq \left[\bar{\rho} + \sqrt{2(M + \lambda^{-1})\varepsilon}\right]^2,$$

proving the second inequality in (b), and consequently concluding the proof of (a) and (b). Now to prove (c) first note that the inclusion follows immediately from the inclusion in (a) and the definition of  $v_g$  in (16). On the other hand, the  $M$ -Lipschitz continuity of  $\nabla g$  together with the definitions of  $q_g$  and  $v_g$  in (14) and (16), respectively, and the triangle inequality for norms imply that

$$\|v_g\| \leq M\|z - z_g\| + \|q_g\| = \frac{M}{M + \lambda^{-1}}\|q_g\| + \|q_g\| \leq 2\|q_g\|$$

which combined with the inequality in (a) proves the inequality in (c).  $\square$

## Appendix B. Proof of Proposition 7.

From the optimality condition for (30), we obtain

$$(100) \quad 0 \in \nabla g(z_{k-1}) + \frac{z_k - z_{k-1}}{\lambda} + \partial h(z_k).$$

Now let

$$(101) \quad \Psi_\lambda = \Psi_{\lambda,k} := g + \frac{1}{2\lambda} \|\cdot - z_{k-1}\|^2, \quad r_k := \frac{z_{k-1} - z_k}{\lambda},$$

and note that  $\nabla \Psi_\lambda(z_{k-1}) = \nabla g(z_{k-1})$ , and that  $\Psi_\lambda$  is convex due to (10) and the assumption  $\lambda < 1/m$ . Hence Proposition 1 yields  $\nabla g(z_{k-1}) = \nabla \Psi_\lambda(z_{k-1}) \in \partial_{\varepsilon_k} \Psi_\lambda(z_k)$  where  $\varepsilon_k = \Psi_\lambda(z_k) - \Psi_\lambda(z_{k-1}) - \langle \nabla \Psi_\lambda(z_{k-1}), z_k - z_{k-1} \rangle \geq 0$ . The above inclusion combined with (100) and definition of  $r_k$  imply that  $r_k \in \partial h(z_k) + \partial_{\varepsilon_k} \Psi_\lambda(z_k) \subset \partial_{\varepsilon_k} (h + \Psi_\lambda)(z_k)$  where the last inclusion follows immediately from the definition of the operator  $\partial_{\varepsilon_k}$  and convexity of  $h$ . Hence, since  $(\tilde{\varepsilon}_k, \tilde{v}_k) = \lambda(\varepsilon_k, r_k)$  (see (31) and (101)), it follows from the above inclusion and the definition of  $\Psi_\lambda$  that the triple  $(z_k, \tilde{v}_k, \tilde{\varepsilon}_k)$  satisfies the inclusion in (18) with  $\phi = g + h$  and  $\lambda_k = \lambda$ .

Now, to prove that the inequality in (19) holds, first note that the definitions of  $\varepsilon_k$  and  $\Psi_\lambda$  together with property (10), imply that  $\varepsilon_k \leq (\lambda M + 1) \|z_{k-1} - z_k\|^2 / (2\lambda)$ . Combining the latter inequality with the relations  $\tilde{v}_k = z_{k-1} - z_k$  and  $\tilde{\varepsilon}_k = \lambda \varepsilon_k$ , we obtain

$$\begin{aligned} \|\tilde{v}_k\|^2 + 2\tilde{\varepsilon}_k &= \|z_{k-1} - z_k\|^2 + 2\lambda \varepsilon_k \leq \|z_{k-1} - z_k\|^2 + (\lambda M + 1) \|z_{k-1} - z_k\|^2 \\ &= (\lambda M + 2) \|z_{k-1} - z_k\|^2 = \frac{\lambda M + 2}{4} \|z_{k-1} - z_k + \tilde{v}_k\|^2. \end{aligned}$$

Hence, since  $\lambda M < 2$ , we conclude that  $\sigma = (\lambda M + 2)/4 < 1$  and that (19) holds.  $\square$

## REFERENCES

- [1] H. ATTOUCH AND J. BOLTE, *On the convergence of the proximal algorithm for nonsmooth functions involving analytic features*, Math. Programming, 116 (2009), pp. 5–16.
- [2] H. ATTOUCH, J. BOLTE, AND B. F. SVAITER, *Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss–Seidel methods*, Math. Programming, 137 (2011), pp. 91–129.
- [3] H. ATTOUCH AND J. PEYPOUQUET, *The rate of convergence of Nesterov’s accelerated forward-backward method is actually faster than  $1/k^{-2}$* , SIAM J. Optim., 26 (2016), pp. 1824–1834.
- [4] N. AYBAT AND G. IYENGAR, *A first-order smoothed penalty method for compressed sensing*, SIAM J. Optim., 21 (2011), pp. 287–313.
- [5] Y. CARMON, J. C. DUCHI, O. HINDER, AND A. SIDFORD, *Accelerated methods for non-convex optimization*, Available on arXiv:1611.00756, (2017).
- [6] E. CHOUZENOUX, J. PESQUET, AND A. REPETTI, *A block coordinate variable metric forward-backward algorithm*, J. Global Optim., 66 (2016), pp. 457–485.
- [7] D. DRUSVYATSKIY AND C. PAQUETTE, *Efficiency of minimizing compositions of convex functions and smooth maps*, Available on arXiv:1605.00125, (2016).
- [8] P. FRANKEL, G. GARRIGOS, AND J. PEYPOUQUET, *Splitting methods with variable metric for kurdyka–łojasiewicz functions and general convergence rates*, J. Optim. Theory Appl., 165 (2015), pp. 874–900.
- [9] S. GHADIMI AND G. LAN, *Accelerated gradient methods for nonconvex nonlinear and stochastic programming*, Math. Programming, 156 (2016), pp. 59–99.
- [10] S. GHADIMI, G. LAN, AND H. ZHANG, *Generalized uniformly optimal methods for nonlinear programming*, Available on arXiv:1508.07384, (2015).
- [11] Y. HE AND R. D. C. MONTEIRO, *Accelerating block-decomposition first-order methods for solving composite saddle-point and two-player Nash equilibrium problems*, SIAM J. Optim., 25 (2015), pp. 2182–2211.
- [12] Y. HE AND R. D. C. MONTEIRO, *An accelerated HPE-type algorithm for a class of composite convex-concave saddle-point problems*, SIAM J. Optim., 26 (2016), pp. 29–56.
- [13] J. HIRIART-URRUTY AND C. LEMARECHAL, *Convex Analysis and Minimization Algorithms II*, Springer, Berlin, 1993.
- [14] M. HONG, *Decomposing linearly constrained nonconvex problems by a proximal primal dual approach: algorithms, convergence, and applications*, Available on arXiv:1604.00543, (2016).

- [15] B. JIANG, T. LIN, S. MA, AND S. ZHANG, *Structured nonconvex and nonsmooth optimization: algorithms and iteration complexity analysis*, Comput. Optim. Appl., (2018).
- [16] O. KOLOSSOSKI AND R. D. C. MONTEIRO, *An accelerated non-euclidean hybrid proximal extragradient-type algorithm for convex-concave saddle-point problems*, Optim. Methods Softw., 32 (2017), pp. 1244–1272.
- [17] G. LAN AND R. D. C. MONTEIRO, *Iteration-complexity of first-order penalty methods for convex programming*, Math. Programming, 138 (2013), pp. 115–139.
- [18] G. LAN AND R. D. C. MONTEIRO, *Iteration-complexity of first-order augmented Lagrangian methods for convex programming*, Math. Programming, 155 (2016), pp. 511–547.
- [19] H. LI AND Z. LIN, *Accelerated proximal gradient methods for nonconvex programming*, Adv. Neural Inf. Process. Syst., 28 (2015), pp. 379–387.
- [20] C. MOLINARI, J. PEYPOUQUET, AND F. ROLDAN, *Alternating forward-backward splitting for linearly constrained optimization problems*, Optim. Lett., (2019), pp. 1–18.
- [21] R. D. C. MONTEIRO AND B. F. SVAITER, *On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean*, SIAM J. Optim., 20 (2010), pp. 2755–2787.
- [22] R. D. C. MONTEIRO AND B. F. SVAITER, *An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods*, SIAM J. Optim., 23 (2013), pp. 1092–1125.
- [23] I. NECOARA, A. PATRASCU, AND F. GLINEUR, *Complexity of first-order inexact Lagrangian and penalty methods for conic convex programming*, Optim. Methods Softw., (2017), pp. 1–31.
- [24] M. O. NEILL AND S. J. WRIGHT, *Behavior of accelerated gradient methods near critical points of nonconvex functions*, Available on arXiv:1706.07993, (2017).
- [25] Y. NESTEROV, *Gradient methods for minimizing composite functions*, Math. Programming, (2012), pp. 1–37.
- [26] Y. E. NESTEROV, *Introductory lectures on convex optimization : a basic course*, Kluwer Academic Publ., Boston, 2004.
- [27] Y. OUYANG, Y. CHEN, G. LAN, AND E. PASILIAO JR., *An accelerated linearized alternating direction method of multipliers*, SIAM J. Imaging Sci., 8 (2015), pp. 644–681.
- [28] C. PAQUETTE, H. LIN, D. DRUSVYATSKIY, J. MAIRAL, AND Z. HARCHAOUI, *Catalyst acceleration for gradient-based non-convex optimization*, Available on arXiv:1703.10993, (2017).
- [29] A. PATRASCU, I. NECOARA, AND Q. TRAN-DINH, *Adaptive inexact fast augmented Lagrangian methods for constrained convex optimization*, Optim. Lett., 11 (2017), pp. 609–626.
- [30] M. V. SOLODOV AND B. F. SVAITER, *A hybrid approximate extragradient-proximal point algorithm using the enlargement of a maximal monotone operator*, Set-Valued Anal., 7 (1999), pp. 323–345.
- [31] P. TSENG, *On accelerated proximal gradient methods for convex-concave optimization*, <http://www.mit.edu/~dimitrib/PTseng/papers.html>, (2008).