Efficiency of minimizing compositions of convex functions and smooth maps *

D. Drusvyatskiy † C. Paquette ‡

Abstract

We consider the problem of minimizing a sum of a convex function and a composition of a convex function with a smooth map. Important examples include exact penalty formulations of nonlinear programs and nonlinear least squares problems with side constraints. The basic algorithm we rely on is the well-known prox-linear method, which in each iteration solves a regularized subproblem formed by linearizing the smooth map. When the subproblems are solved exactly, the method has the efficiency guarantee $\mathcal{O}(\varepsilon^{-2})$, akin to gradient descent for smooth minimization. Our contributions are threefold: we (1) derive an inertial prox-linear method that accelerates in presence of convexity, (2) quantify the impact of inexact subproblem solves, and (3) discuss efficiency of smoothing techniques. When the subproblems can only be solved by first-order methods, we show that surprisingly a simple combination of the prox-linear method, smoothing, and fast-gradient subproblem solves yields a scheme with overall efficiency $\widetilde{\mathcal{O}}(\varepsilon^{-3})$. This appears to be the best known complexity bound for the problem class among first-order methods.

Key words. Composite minimization, fast gradient methods, Gauss-Newton, prox-gradient, inexactness, complexity, smoothing

AMS Subject Classification. *Primary* 97N60, 90C25; *Secondary* 90C06, 90C30.

1 Introduction

In this work, we consider the class of *composite optimization problems*

$$\min_x \; F(x) := g(x) + h(c(x)), \tag{1.1}$$

where $g\colon \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$ and $h\colon \mathbf{R}^m \to \mathbf{R}$ are closed convex functions and $c\colon \mathbf{R}^n \to \mathbf{R}^m$ is a smooth map. Such problems are typically nonsmooth and nonconvex, but are highly structured. Classical examples include regularized Nonlinear Least Squares and exact penalty formulations of nonlinear programs; see Section 3 for more details. The setting where c maps to the real line and h is the identity function, namely

$$\min_x \; c(x) + g(x), \tag{1.2}$$

*University of Washington, Department of Mathematics, Seattle, WA 98195; Research of Drusvyatskiy and Paquette was partially supported by the AFOSR YIP award FA9550-15-1-0237.

†E-mail: ddrusv@uw.edu; `http://www.math.washington.edu/~ddrusv/`

‡E-mail: yumiko88@uw.edu;

is now commonplace in large-scale optimization and high-dimensional statistics. In this work, we use the term *additive composite minimization* for (1.2) to distinguish it from the more general composite problem class (1.1).

The standard prox-gradient algorithm for additive composite minimization (1.2) quickly extends to the entire problem class (1.1). The resulting scheme, called the *prox-linear method* was recently investigated in [11, 18, 26], though the ideas behind the algorithm and of its trust-region variants are decades-old [6, 11, 21, 40, 41, 47, 48]. The prox-linear method in each iteration linearizes the smooth map $c(\cdot)$ and solves the *proximal subproblem*:

$$x_{k+1} = \operatorname*{argmin}_x \Big\{ g(x) + h\Big(c(x_k) + \nabla c(x_k)(x - x_k)\Big) + \tfrac{1}{2t}\|x - x_k\|^2 \Big\}, \tag{1.3}$$

for an appropriately chosen parameter $t > 0$. The underlying assumption is that the strongly convex proximal subproblems (1.3) can be solved efficiently. The prox-linear method reduces to the popular prox-gradient algorithm [3, 37] for additive composite minimization, while for nonlinear least squares, the scheme is the Levenberg-Marquardt method – a damped variant of Gauss-Newton [38, Section 10].

To make precise global rates of convergence, we assume that h is L-Lipschitz and the Jacobian map ∇c is β-Lipschitz. As in the analysis of the prox-gradient method in Nesterov [33, 36], it is convenient to measure the progress of the prox-linear method in terms of the scaled steps

$$\mathcal{G}_t(x_k) := t^{-1}(x_k - x_{k+1}),$$

called the *prox-gradients*. Indeed, the quantities $\|\mathcal{G}_t(x_k)\|$ measure the approximate stationarity of the iterates.[1] A short argument shows that with the optimal choice $t = L\beta$, the prox-linear algorithm will find a point x satisfying $\|\mathcal{G}_{\frac{1}{L\beta}}(x)\| \leq \varepsilon$ after at most $\mathcal{O}(\frac{L\beta}{\varepsilon^2})$ iterations; see e.g. [11, 18]. We mention in passing that iterate convergence under the KŁ-inequality was recently shown by Pauwels [39], while local linear rates under quadratic growth conditions (and KŁ-inequality with exponent 1/2) were proved in [18]. Even faster local rates are possible under sharpness assumptions [9]. In contrast, our focus here is on global sublinear rates of convergence.

For additive composite problems, with c in addition convex, the prox-gradient method is sub-optimal from the viewpoint of computational complexity [32,33]. Accelerated gradient methods, beginning with Nesterov [35] and extended by Beck-Teboulle [3] and Nesterov [37], achieve the superior rate $\mathcal{O}((\frac{\beta}{\varepsilon})^{2/3})$, and even a faster rate is possible by first regularizing the problem [34]. Consequently, desirable would be an algorithm that *automatically* accelerates in presence of convexity, while performing no worse than the prox-gradient method on nonconvex instances. In the recent manuscript [22], Ghadimi and Lan described such a scheme for additive composite problems. Similar acceleration techniques have also been used for exact penalty formulations of nonlinear programs (1.1) with numerical success, but without formal justification; the paper of Burke, Curtis et al. [8] is a good example.

The contributions of this work are threefold.

1. **(Acceleration)** We extend the accelerated algorithm of Ghadimi and Lan [22] for additive composite problems to the entire problem class (1.1). Assuming that the domain of g is bounded, the algorithm has worst-case convergence guarantees analogous to those of the prox-linear method, while achieving an accelerated rate for composite functions satisfying a

[1] Theorem 5.3 in the recent manuscript [18] shows that a small prox-gradient $\|\mathcal{G}_t(x_k)\|$ guarantees that x_{k+1} is proportionally close to some point that is nearly stationary for F.

convexity condition. More precisely, setting $M := \operatorname{diam}(\operatorname{dom} g)$, the scheme comes equipped with the guarantee

$$\min_{j=1,\ldots,k} \left\| \mathcal{G}_{\frac{1}{2L\beta}}(x_j) \right\|^2 \leq (L\beta M)^2 \cdot \mathcal{O}\left(\frac{1}{k^3} + \frac{c_2}{k^2} + \frac{c_1}{k} \right),$$

where the constants $0 \leq c_1 \leq c_2 \leq 1$ quantify "convexity-like behavior" of the composition.

2. **(Inexactness and complexity of first-order methods)** For the general composite class (1.1), coping with inexactness in the proximal subproblem solves (1.3) is unavoidable. We perform an inexact analysis of both the prox-linear method and of its accelerated variant based on two natural models of inexactness: (i) near-optimality in function value and (ii) near-stationarity in the dual.

   Based on the inexact analysis, we derive overall efficiency estimates for the (accelerated) prox-linear method, where the proximal subproblems are themselves solved by first-order algorithms. Assuming that the proximal maps of h and g are computable, we show that an inexact prox-linear method with proximal subproblems solved by fast-gradient methods, has the efficiency estimate $\mathcal{O}(\frac{L^3\beta^{3/2}\operatorname{lip}_g(c)}{\varepsilon^4})$, where $\operatorname{lip}_g(c)$ denotes the Lipschitz constant of c on the domain of g. This naive strategy turns out to be suboptimal, however, and can be improved through smoothing.

3. **(Improved complexity of first-order methods through smoothing)** Smoothing is a common technique in nonsmooth optimization. The influential paper of Nesterov [36], in particular, derives convergence guarantees for algorithms based on infimal convolution smoothing in structured convex optimization. In contrast, for nonconvex problems, worst-case global efficiency estimates based on smoothing are less common. The reason is that one can not use progress in function values to decide on an appropriate smoothing parameter. In the context of the composite class (1.1), smoothing is indeed appealing. In the simplest case, one replaces the function h by a smooth approximation and solves the resulting smooth problem instead. For example, one can simply apply the prox-gradient method, though we will see that this leads to poor guarantees. Surprisingly, we show that running the prox-linear method on the smooth approximation, with the proximal subproblems solved by fast-gradient methods, yields an algorithm with overall efficiency $\widetilde{\mathcal{O}}(\frac{L^2\beta\operatorname{lip}_g(c)}{\varepsilon^3})$. Here $\widetilde{\mathcal{O}}$ hides logarithmic terms. To the best of our knowledge, this is the best known complexity bound for the problem class (1.1) among first-order methods.

The outline of the manuscript is as follows. Section 2 records basic notation that we use throughout the paper. In Section 3, we introduce the composite problem class and first-order stationarity, along with motivating examples. Section 4 introduces the accelerated prox-linear method and proves its convergence guarantees. Section 5 discusses the impact of inexactness in the proximal subproblems, while Section 6 derives overall efficiency estimates when the proximal subproblems are solved by first-order methods. We complete the paper with Section 7, discussing efficiency of smoothing.

2 Notation

The notation we follow is standard. Throughout, we consider a Euclidean space, denoted by $\mathbf{R}^n$, with an inner product $\langle \cdot, \cdot \rangle$ and the induced norm $\| \cdot \|$. Given a set Q in $\mathbf{R}^n$, the *distance*

and *projection* of a point x onto Q are given by

$$\operatorname{dist}(x;Q) := \inf_{y\in Q} \|y-x\|, \qquad \operatorname{proj}(x;Q) := \operatorname*{argmin}_{y\in Q} \|y-x\|,$$

respectively.

The functions we consider take values in the extended-real-line $\overline{\mathbf{R}} := \mathbf{R}\cup\{\pm\infty\}$. The *domain* and the *epigraph* of any function $f\colon \mathbf{R}^n \to \overline{\mathbf{R}}$ are the sets

$$\operatorname{dom} f := \{x \in \mathbf{R}^n : f(x) < +\infty\}, \qquad \operatorname{epi} f := \{(x,r) \in \mathbf{R}^n \times \mathbf{R} : f(x) \le r\},$$

respectively. We say that f is *closed* if its epigraph, $\operatorname{epi} f$, is a closed set. Throughout, we will assume that all functions that we encounter are *proper*, meaning they have nonempty domains are never take on the value $-\infty$. The indicator function of a set $Q \subseteq \mathbf{R}^n$, denoted by δ_Q, is defined to be zero on Q and $+\infty$ off it.

Given a convex function $f\colon \mathbf{R}^n \to \overline{\mathbf{R}}$, a vector v is called a *subgradient* of f at a point $x \in \operatorname{dom} f$ if the inequality

$$f(y) \ge f(x) + \langle v, y-x\rangle \qquad \text{holds for all } y \in \mathbf{R}^n.$$

The set of all subgradients of f at x is denoted by $\partial f(x)$, and is called the *subdifferential* of f at x. For any point $x \notin \operatorname{dom} f$, we set $\partial f(x)$ to be the empty set. With any convex function f, we associate its *Fenchel conjugate*

$$f^\star(y) := \sup_x \{\langle y, x\rangle - f(x)\}.$$

It is well-known that whenever f is closed and convex, equality $f = f^{\star\star}$ holds.

For any function f and parameter $\nu > 0$, the *Moreau envelope* and the *proximal mapping* are defined by

$$f_\nu(x) := \inf_z \left\{ f(z) + \frac{1}{2\nu}\|z-x\|^2 \right\},$$
$$\operatorname{prox}_{\nu f}(x) := \operatorname*{argmin}_z \left\{ f(z) + \frac{1}{2\nu}\|z-x\|^2 \right\}.$$

respectively. In particular, the Moreau envelope of an indicator function δ_Q is simply the map $x \mapsto \frac{1}{2\nu}\operatorname{dist}^2(x;Q)$ and the proximal mapping of δ_Q is the projection $x \mapsto \operatorname{proj}(x;Q)$. The following lemma lists well-known regularization properties of the Moreau envelope.

Lemma 2.1 (Regularization properties of the envelope)**.** *Suppose that f is convex and L-Lipschitz. Then the envelope $f_\nu(\cdot)$ is convex, L-Lipschitz, and satisfies*

$$0 \le f(x) - f_\nu(x) \le \frac{L^2\nu}{2} \qquad \text{for all } x \in \mathbf{R}^n. \tag{2.1}$$

Moreover, f_ν is C^1-smooth with $\nabla f_\nu(x) = \frac{1}{\nu}(x - \operatorname{prox}_{\nu f}(x))$ and $\operatorname{lip}(\nabla f_\nu) \le \frac{1}{\nu}$.

Proof. The expression $\nabla f_\nu(x) = \frac{1}{\nu}(x - \operatorname{prox}_{\nu f}(x))$, along with the inequality $\operatorname{lip}(\nabla f_\nu) \le \frac{1}{\nu}$, can be found for example in [42, Exercise 12.23]. The expression (2.1) follows from rewriting $f_\nu(x) = (f^\star + \frac{\nu}{2}\|\cdot\|^2)^\star(x) = \sup_z \{\langle x, z\rangle - f^\star(z) - \frac{\nu}{2}\|z\|^2\}$ (as in e.g. [42, Example 11.26]) and noting that the domain of $f^\star$ is bounded in norm by L. Finally, the claim that f_ν is L-Lipschitz is due to the elementary observation that for any two closed convex function g and h satisfying $g \le h$, the inequality $\operatorname{lip}(g) \le \operatorname{lip}(h)$ holds. □

3 The composite problem class

Our work centers around nonsmooth and nonconvex optimization problems of the form

$$\min_x F(x) := g(x) + h(c(x)). \tag{3.1}$$

Throughout, we make the following assumptions on the functional components of the problem:

1. $g\colon \mathbf{R}^n \to \overline{\mathbf{R}}$ is a closed convex function;
2. $h\colon \mathbf{R}^m \to \mathbf{R}$ is an L-Lipschitz continuous convex function;
3. $c\colon \mathbf{R}^n \to \mathbf{R}^m$ is a C^1-smooth mapping with the Jacobian $x \mapsto \nabla c(x)$ that is β-Lipschitz continuous.

For ease of reference in the latter parts of the paper, we also define the constants

$$\mu := L\beta, \qquad M := \sup_{x,y \in \operatorname{dom} g} \|x - y\|, \qquad \|\nabla c\| := \sup_{x \in \operatorname{dom} g} \|\nabla c(x)\|.$$

None of the algorithms we discuss require M nor $\|\nabla c\|$ to be finite (or known) for the implementation; these two quantities will only appear in the efficiency estimates. Similarly, if h and ∇c are only locally Lipschitz continuous, then one can use a backtracking line search instead of fixed step-sizes in the schemes we discuss.

Before discussing algorithms, it is instructive to consider some motivating examples fitting into the composite framework (3.1).

Example 3.1 (Additive composite minimization)**.** The most prevalent example of the composite class (3.1) is additive composite minimization. In this case, the map c maps to the real line and h is the identity function:

$$\min_x \; g(x) + c(x). \tag{3.2}$$

Such problems appear often in statistical learning and imaging, for example. Numerous algorithms are available, especially when c is convex, such as proximal gradient methods and their accelerated variants [3, 37].

Example 3.2 (Nonlinear least squares)**.** The composite problem class also captures nonlinear least squares problems with bound constraints:

$$\min_x \; \|c(x)\| \qquad \text{subject to} \qquad l_i \le x_i \le u_i \quad \text{for } i = 1, \ldots, m.$$

Gauss-Newton type algorithm [25, 28, 30] are often the methods of choice for such problems.

Example 3.3 (Exact penalty formulations)**.** Consider a nonlinear optimization problem:

$$\min_x \; \{f(x) : G(x) \in \mathcal{K}\},$$

where $f\colon \mathbf{R}^n \to \mathbf{R}$ and $G\colon \mathbf{R}^n \to \mathbf{R}^m$ are smooth mappings and $\mathcal{K} \subseteq \mathbf{R}^m$ is a closed convex cone. An accompanying *penalty formulation* – ubiquitous in nonlinear optimization [7, 10, 13, 16, 20] – takes the form

$$\min_x \; f(x) + \lambda \cdot \theta_{\mathcal{K}}(G(x)),$$

where $\theta_{\mathcal{K}}\colon \mathbf{R}^m \to \mathbf{R}$ is a nonnegative convex function that is zero only on $\mathcal{K}$ and $\lambda > 0$ is a penalty parameter. For example, $\theta_{\mathcal{K}}(y)$ is often the distance of y to the convex cone $\mathcal{K}$ in some norm. This is an example of (3.1) under the identification $c(x) = (f(x), G(x))$ and $h(f, G) = f + \lambda\theta_{\mathcal{K}}(G)$.

Example 3.4 (Misfit measures)**.** Often, one is interested in minimizing an error between a nonlinear process model $G(x)$ and observed data b through a misfit measure h. The resulting problem takes the form

$$\min_x \; h\big(b - G(x)\big) + g(x),$$

where g may be a convex surrogate encouraging prior structural information on x, such as the l_1-norm, squared l_2-norm, or the indicator of the nonnegative orthant. The misfit $h = \|\cdot\|_2$, in particular, appears in nonlinear least squares. The l_1-norm $h = \|\cdot\|_1$ is commonly used in the Least Absolute Deviations (LAD) technique in regression [31, 44] and in Kalman smoothing with impulsive disturbances [1].

Another popular class of misfit measures h is a sum $h = \sum_i h_\kappa(y_i)$ of Huber functions

$$h_\kappa(\tau) = \begin{cases} \frac{1}{2\kappa}\tau^2 & , \tau \in [-\kappa, \kappa] \\ |\tau| - \frac{\kappa}{2} & , \text{otherwise} \end{cases}$$

The Huber function figures prominently in robust regression [12, 19, 24, 27], being much less sensitive to outliers than the least squares penalty due to its linear tail growth. The function h thus defined is smooth with $\operatorname{lip}(\nabla h) \sim 1/\kappa$. Hence in particular the term $h(b - G(x))$ can be treated as a smooth term reducing to the setting of additive composite minimization (Example 3.1). On the other hand, we will see that because of the poor conditioning of the gradient ∇h, methods that take into account the non-additive composite structure can have superior efficiency estimates.

Example 3.5 (Grey-box minimization)**.** In industrial applications, one is often interested in functions that are available only *implicitly*. For example, function and derivative evaluations may require execution of an expensive simulation. Such problems often exhibit an underlying composite structure $h(c(x))$. The penalty function h is known (and chosen) explicitly and is simple, whereas the mapping $c(x)$ and the Jacobian ∇c might only be available through a simulation. Problems of this type are sometimes called *grey-box minimization problems*, in contrast to black-box minimization. The explicit separation of the noisy mapping c and the user chosen penalty h can help in designing algorithms. See for example Conn-Scheinberg-Vicente [14] and Wild [46], and references therein.

Before discussing algorithms for the problem class (3.1), we must first explain the goal of all such methods. Since the optimization problem (3.1) is nonconvex, it is natural to seek points x that are only first-order stationary. One makes this notion precise through subdifferentials – workhorse of variational analysis. For more details, see for example the monographs of Mordukhovich [29] and Rockafellar-Wets [42].

Definition 3.1 (Subdifferential and stationary points)**.** The *subdifferential* of F at a point $x \in \operatorname{dom} F$ is the set

$$\partial F(x) := \partial g(x) + \nabla c(x)^* \partial h(c(x)).$$

We say that x is *stationary* for F if the inclusion $0 \in \partial F(x)$ holds.

The algorithms we consider aim to find stationary points of F. Stationarity is a meaningful concept: a point x is stationary for F if and only if the directional derivative of F at x is nonnegative in every direction [42, Theorem 8.3]. More precisely, the equality holds:

$$\operatorname{dist}(0; \partial f(x)) = -\inf_{v:\, \|v\| \le 1} F'(x; v),$$

where $F'(x; v)$ is the directional derivative of F at x in direction v [42, Definition 8.1].

4 Prox-linear and accelerated prox-linear methods

The most basic algorithm for additive composite minimization (3.2) is the prox-gradient method

$$x_{k+1} := \underset{x}{\operatorname{argmin}} \ \left\{ c(x_k) + \langle \nabla c(x_k), x - x_k \rangle + g(x) + \frac{1}{2t}\|x - x_k\|^2 \right\} \tag{4.1}$$

or equivalently

$$x_{k+1} = \operatorname{prox}_{tg}\left(x_k - t\nabla c(x_k)\right).$$

Notice that an underlying assumption here is that the proximal map prox_{tg} is computable.

Convergence analysis of the prox-gradient algorithm derives from the fact that the function minimized in (4.1) is an upper model of F whenever $t \leq \beta^{-1}$. This majorization viewpoint quickly yields an algorithm for the entire problem class (3.1). The so-called *prox-linear algorithm* iteratively linearizes the map c and solves a proximal subproblem. To formalize the scheme, we use the following notation.

For any points $x, y \in \mathbf{R}^n$ and a real $t > 0$, define

$$\begin{aligned} F(z;y) &:= g(z) + h\Big(c(y) + \nabla c(y)(z-y)\Big), \\ F_t(z;y) &:= F(z;y) + \frac{1}{2t}\|z-y\|^2, \\ S_t(y) &:= \underset{z}{\operatorname{argmin}} \ F_t(z;y). \end{aligned}$$

One can readily verify the inequalities

$$-\frac{\mu}{2}\|z-y\|^2 \leq F(z) - F(z;y) \leq \frac{\mu}{2}\|z-y\|^2. \tag{4.2}$$

In particular $F_t(\cdot;y)$ is an upper model for F for any $t < \mu^{-1}$, meaning $F_t(z;y) \geq F(z)$ for all points y, z. The *prox-linear method* is then simply the recurrence $x_{k+1} = S_t(x_k)$. Specializing to the the additive composite setting (3.2), equality $S_t(x) = \operatorname{prox}_{tg}(x - t\nabla c(x))$ holds and the prox-linear method reduces to the familiar prox-gradient iteration.

Algorithm 1: Prox-linear method

Initialize : A point $x_0 \in \operatorname{dom} g$ and a real $t > 0$.

Step k: ($k \geq 0$) Compute

$$x_{k+1} = \underset{x}{\operatorname{argmin}} \left\{ g(x) + h\Big(c(x_k) + \nabla c(x_k)(x - x_k)\Big) + \frac{1}{2t}\|x - x_k\|^2 \right\}. \tag{4.3}$$

Notice that we are implicitly assuming here that the proximal subproblem (4.3) is solvable. We will discuss the impact of inexact evaluation of $S_t(\cdot)$ in Section 5.

Convergence guarantees for the prox-linear method are best stated in terms of the *prox-gradient* mapping

$$\mathcal{G}_t(x) := t^{-1}(x - S_t(x)).$$

Observe that the optimality conditions for the proximal subproblem $\min_z F(z;x)$ read

$$\mathcal{G}_t(x) \in \partial g(S_t(x)) + \nabla c(x)^T \partial h(c(x) + \nabla c(x)(S_t(x) - x))$$

In particular, for any $t > 0$, a point x is first-order stationary for F if and only if equality $\mathcal{G}_t(x) = 0$ holds. Hence the norm $\|\mathcal{G}_t(x)\|$ serves as a measure of proximity to stationarity; a more precise relaceship between the prox-gradient and near-stationarity will be explained in Section 7.1, and will be crucially used for smoothing techniques. Let us review here the rudimentary convergence guarantees of the method, as presented for example in [18, Section 5]. We provide a quick proof for completeness.

Proposition 4.1 (Efficiency of the prox-linear method)**.** *Supposing $t \leq \mu^{-1}$, the iterates generated by Algorithm 1 satisfy*

$$\min_{j=0,\dots,N-1} \|\mathcal{G}_t(x_j)\|^2 \leq \frac{2t^{-1}\big(F(x_0) - F^*\big)}{N},$$

where we set $F^* := \lim_{N\to\infty} F(x_N)$.

Proof. Taking into account that $F_t(\cdot; x_k)$ is strongly convex with modulus $1/t$, we obtain

$$F(x_k) = F_t(x_k; x_k) \geq F_t(x_{k+1}; x_k) + \tfrac{t}{2}\|\mathcal{G}_t(x_k)\|^2 \geq F(x_{k+1}) + \tfrac{t}{2}\|\mathcal{G}_t(x_k)\|^2.$$

Summing the inequalities yields

$$\min_{j=0,\dots,N-1} \|\mathcal{G}_t(x_j)\|^2 \leq \frac{1}{N}\sum_{j=0}^{N-1} \|\mathcal{G}_t(x_j)\|^2 \leq \frac{2t^{-1}\big(F(x_0) - F^*\big)}{N},$$

as claimed. □

4.1 An accelerated prox-linear algorithm

In this section, we describe an acceleration strategy for the prox-linear method, in a sense to be made precise shortly. To motivate the algorithm, let us consider again the additive composite setting (3.2) with $c(\cdot)$ in addition convex. Algorithms in the style of Nesterov's second accelerated method incorporate steps of the form $v_{k+1} = \mathrm{prox}_{tg}\,(v_k - t\nabla c(y_k))$. That is, one moves from a point v_k in the direction of the negative gradient $-\nabla c(y_k)$ from a different point v_k, followed by a proximal operation. Equivalently, after completing a square one can write

$$v_{k+1} := \operatorname*{argmin}_z \ \left\{ c(y_k) + \langle \nabla c(y_k), z - v_k\rangle + \frac{1}{2t}\|z - v_k\|^2 + g(z) \right\}.$$

This is also the construction used by Ghadimi and Lan [22, Equation 2.37] for nonconvex additive composite problems. The algorithm we consider emulates this operation. There is a slight complication, however, in that the composite structure requires us to incorporate an additional scaling parameter α in the construction. We use the following notation:

$$\begin{aligned} F_\alpha(z; y, v) &:= g(z) + \frac{1}{\alpha}\cdot h\big(c(y) + \alpha\nabla c(y)(z - v)\big), \\ F_{t,\alpha}(z; y, v) &:= F_\alpha(z; y, v) + \frac{1}{2t}\|z - v\|^2, \\ S_{t,\alpha}(y, v) &:= \operatorname*{argmin}_z \ F_{t,\alpha}(z; y, v). \end{aligned}$$

Observe the equality $S_{t,1}(x,x) = S_t(x)$. In the additive composite setting, the mapping $S_{t,\alpha}(y,v)$ does not depend on α and the definition reduces to

$$S_{t,\alpha}(y,v) = \operatorname*{argmin}_z \left\{ c(y) + \langle \nabla c(y), z - v \rangle + \frac{1}{2t} \|z - v\|^2 + g(z) \right\} = \operatorname{prox}_{tg}\left(v - t\nabla c(y)\right).$$

The scheme we propose is summarized in Algorithm 2.

Algorithm 2: Accelerated prox-linear method

Initialize : Fix two points $x_0, v_0 \in \operatorname{dom} g$ and a real number $\tilde{\mu} > \mu$.

Step k: ($k \geq 1$) Compute

$$a_k = \tfrac{2}{k+1} \tag{4.4}$$
$$y_k = a_k v_{k-1} + (1 - a_k) x_{k-1} \tag{4.5}$$
$$x_k = S_{1/\tilde{\mu}}(y_k) \tag{4.6}$$
$$v_k = S_{\frac{1}{\tilde{\mu} a_k}, a_k}(y_k, v_{k-1}) \tag{4.7}$$

Remark 4.2 (Backtracking line-search)**.** We remark that when the constants L and β are unknown, one can instead equip Algorithm 2 with a backtracking line search. More precisely, one can insert Algorithm 3 after line (4.5) in Algorithm 2. The line search is entirely analogous to the one used in FISTA [3]. Moreover, one may use the interpolation weights used in FISTA [3]; namely, the sequence a_k may be chosen to satisfy the polynomial $\frac{1-a_k}{a_k^2} = \frac{1}{a_{k-1}^2}$, with similar convergence guarantees.

Algorithm 3: Backtracking line search in iteration k

Initialize : Real numbers $\eta, c \in (0,1)$ and $t_{k-1} > 0$.

$t \leftarrow t_{k-1}$

while $F(S_t(y_k)) > F_t(S_t(y_k))$ **do**

  $t \leftarrow \eta t$

end

$t_k \leftarrow t$

Choose $\tilde{\mu} = c^{-1} t_k$

return $\tilde{\mu}$, t_k;

4.2 Convergence guarantees and convexity moduli

We will see momentarily that convergence guarantees of Algorithm 2 are adaptive to convexity (or lack thereof) of the composition $h \circ c$. To simply notation, henceforth set

$$\Phi := h \circ c.$$

4.2.1 Weak convexity and convexity of the pair

It appears that there are two different convexity-like properties of the composite problem that govern convergence of Algorithm 2. The first is classical.

Definition 4.3 (Weak convexity of the composition)**.** We say that the composite function $\Phi = h \circ c$ is *ρ-weakly convex on a set U* if for any points $x, y \in U$ and $a \in [0,1]$, the approximate secant inequality holds:

$$\Phi(ax + (1-a)y) \leq a\Phi(x) + (1-a)\Phi(y) + \rho a(1-a)\|x-y\|^2.$$

It is well-known that ρ-weak convexity of Φ on $\mathbf{R}^n$ is equivalent to either of the following properties; see e.g. [15, Theorem 3.1].

1. **(Perturbed convexity)** The function $\Phi + \frac{\rho}{2}\|\cdot\|^2$ is convex.
2. **(Quadratic lower-estimators)** For any $x, y \in \mathbf{R}^n$ and $v \in \partial\Phi(x)$, the inequality

$$\Phi(y) \geq \Phi(x) + \langle v, y-x\rangle - \frac{\rho}{2}\|y-x\|^2 \qquad \text{holds.}$$

In particular, the following in true.

Lemma 4.4. *The function Φ is ρ-weakly convex for some $\rho \in [0, \mu]$.*

Proof. Fix two point $x, y \in \mathbf{R}^n$ and a vector $v \in \partial\Phi(x)$. We can write $v = \nabla c(x)^* w$ for some vector $w \in \partial h(c(x))$. Taking into account convexity of h, we then deduce

$$\begin{aligned}\Phi(y) = h(c(y)) \geq h(c(x)) + \langle w, c(y) - c(x)\rangle &\geq \Phi(x) + \langle w, \nabla c(x)(y-x)\rangle - \frac{\beta\|w\|}{2}\|y-x\|^2 \\ &\geq \Phi(x) + \langle v, y-x\rangle - \frac{\mu}{2}\|y-x\|^2.\end{aligned}$$

The result follows. □

Weak convexity is a property of the composite function $h \circ c$ and is not directly related to h nor c individually. In contrast, the algorithm we consider uses explicitly the composite structure. In particular, it seems that the extent to which the "linearization" $z \mapsto h(c(y) + \nabla c(y)(z-y))$ lower bounds $h(c(z))$ should play a role.

Definition 4.5 (Convexity of the pair)**.** A real number $r > 0$ is called a *convexity constant of the pair (h, c) on a set U* if the inequality

$$h\big(c(y) + \nabla c(y)(z-y)\big) \leq h(c(z)) + \frac{r}{2}\|z-y\|^2 \qquad \text{holds for all } z, y \in U.$$

Inequalities (4.2) show that the pair (h, c) has a convexity constant $r \in [0, \mu]$ on $\mathbf{R}^n$. The convexity constant r of the pair (h, c) always upper bounds the weak convexity constant ρ of the composition $h \circ c$.

Lemma 4.6 (Convexity of the pair implies convexity of the composition)**.**
If r is a convexity constant of (h, c) on a convex set U, then Φ is r-weakly convex on U.

Proof. Suppose r is a convexity constant of (h, c) on U. Observe that the subdifferential of the convex function Φ and that of the linearization $h\big(c(y) + \nabla c(y)(\cdot - y)\big)$ coincide at $y = x$. Therefore a quick argument shows that for any $x, y \in U$ and $v \in \partial\Phi(y)$ we have

$$\Phi(x) \geq h(c(y) + \nabla c(y)(x - y)) - \frac{r}{2}\|x - y\|^2 \geq \Phi(y) + \langle v, x - y\rangle - \frac{r}{2}\|x - y\|^2.$$

The rest of the proof follows along the same lines as [15, Theorem 3.1]. We omit the details. □

Remark 4.7. The converse of the lemma is false. Consider for example setting $c(x) = (x, x^2)$ and $h(x, z) = x^2 - z$. Then the composition $h \circ c$ is identically zero and hence convex. On the other hand, one can easily check that the pair (h, c) has a nonzero convexity constant.

4.2.2 Convergence guarantees

Henceforth, let ρ be the weak convexity constant of $h \circ c$ on dom g and let r be the convexity constant of (h, c) on dom g. From the above discussion, we can always assume $0 \leq \rho \leq r \leq \mu$. We are now ready to state and prove convergence guarantees of Algorithm 2.

Theorem 4.8 (Convergence guarantees)**.** *Fix a real number $\tilde{\mu} > \mu$ and let x^* be any point satisfying $F(x^*) < F(x_k)$ for all iterates x_k generated by Algorithm 2. Then the efficiency estimate holds:*

$$\min_{j=1,\ldots,N} \left\|\mathcal{G}_{1/\tilde{\mu}}(y_j)\right\|^2 \leq \frac{24\tilde{\mu}^2}{\tilde{\mu} - \mu}\left(\frac{\tilde{\mu}\,\|x^* - v_0\|^2}{N(N+1)(2N+1)} + \frac{M^2(r + \frac{\rho}{2}(N+3))}{(N+1)(2N+1)}\right).$$

In the case $r = 0$, the inequality above holds with the second summand on the right-hand-side replaced by zero (even if $M = \infty$), and moreover the efficiency bound on function values holds:

$$F(x_N) - F(x^*) \leq \frac{2\tilde{\mu}\,\|x^* - v_0\|^2}{(N+1)^2}.$$

Succinctly, setting $\tilde{\mu} := 2\mu$, Theorem 4.8 guarantees the bound

$$\min_{j=1,\ldots,N} \left\|\mathcal{G}_{1/\tilde{\mu}}(y_j)\right\|^2 \leq \mathcal{O}\left(\frac{\mu^2\|x^* - v_0\|^2}{N^3}\right) + \frac{r}{\mu}\cdot\mathcal{O}\left(\frac{\mu^2 M^2}{N^2}\right) + \frac{\rho}{\mu}\cdot\mathcal{O}\left(\frac{\mu^2 M^2}{N}\right).$$

The fractions $0 \leq \frac{\rho}{\mu} \leq \frac{r}{\mu} \leq 1$ balance the three terms, corresponding to different levels of "convexity".

Our proof of Theorem 4.8 is based on two basic lemmas, as is common for accelerated methods [45].

Lemma 4.9 (Three-point comparison)**.** *Consider the point $z := S_{t,\alpha}(y, v)$ for some points $y, v \in \mathbf{R}^n$ and real numbers $t, \alpha > 0$. Then for all $w \in \mathbf{R}^n$ the inequality holds:*

$$F_\alpha(z; y, v) \leq F_\alpha(w; y, v) + \frac{1}{2t}\left(\|w - v\|^2 - \|w - z\|^2 - \|z - v\|^2\right).$$

Proof. This follows immediately by noting that the function $F_{t,\alpha}(\cdot; y, v)$ is strongly convex with constant $1/t$ and z is its minimizer by definition. □

Lemma 4.10 (Telescoping)**.** *Let a_k, y_k, x_k, and v_k be the iterates generated by Algorithm 2. Then for any point $x \in \mathbf{R}^n$ and any index k, the inequality holds:*

$$\begin{aligned} F(x_k) \leq a_k F(x) + (1-a_k)F(x_{k-1}) + \frac{\tilde{\mu} a_k^2}{2}(\|x - v_{k-1}\|^2 - \|x - v_k\|^2) \\ - \frac{\tilde{\mu} - \mu}{2}\|y_k - x_k\|^2 + \rho a_k \|x - x_{k-1}\|^2 + \frac{r a_k^2}{2}\|x - v_{k-1}\|^2. \end{aligned} \tag{4.8}$$

Proof. Notice that all the points x_k, y_k, and v_k lie in dom g. From inequality (4.2), we have

$$F(x_k) \leq h\big(c(y_k) + \nabla c(y_k)(x_k - y_k)\big) + g(x_k) + \frac{\mu}{2}\|x_k - y_k\|^2. \tag{4.9}$$

Define the point $w_k := a_k v_k + (1-a_k)x_{k-1}$. Applying Lemma 4.9 to $x_k = S_{1/\tilde{\mu},1}(y_k, y_k)$ with $w = w_k$ yields the inequality

$$\begin{aligned} h(c(y_k) + \nabla c(y_k)(x_k - y_k)) + g(x_k) \leq h(c(y_k) + \nabla c(y_k)(w_k - y_k)) \\ + \frac{\tilde{\mu}}{2}(\|w_k - y_k\|^2 - \|w_k - x_k\|^2 - \|x_k - y_k\|^2) \\ + a_k g(v_k) + (1-a_k)g(x_{k-1}). \end{aligned}$$

Note the equality $w_k - y_k = a_k(v_k - v_{k-1})$. Applying Lemma 4.9 again with $v_k = S_{\frac{1}{\tilde{\mu} a_k}, a_k}(y_k, v_{k-1})$ and $w = x$ yields

$$\begin{aligned} h\big(c(y_k) + a_k \nabla c(y_k)(v_k - v_{k-1})\big) + a_k g(v_k) \leq h\big(c(y_k) + a_k \nabla c(y_k)(x - v_{k-1})\big) + a_k g(x) \\ + \frac{\tilde{\mu} a_k^2}{2}\left(\|x - v_{k-1}\|^2 - \|x - v_k\|^2 - \|v_k - v_{k-1}\|^2\right). \end{aligned} \tag{4.10}$$

Define the point $\hat{x} := a_k x + (1-a_k)x_{k-1}$. Taking into account $a_k(x - v_{k-1}) = \hat{x} - y_k$, we conclude

$$\begin{aligned} h(c(y_k) + \nabla c(y_k)(\hat{x} - y_k)) &\leq (h \circ c)(\hat{x}) + \frac{r}{2}\|\hat{x} - y_k\|^2 \\ &\leq a_k h(c(x)) + (1-a_k)h(c(x_{k-1})) \\ &\quad + \rho a_k (1-a_k)\|x - x_{k-1}\|^2 + \frac{r a_k^2}{2}\|x - v_{k-1}\|^2. \end{aligned} \tag{4.11}$$

Thus combining inequalities (4.9), (4.10), and (4.11), and upper bounding $1 - a_k \leq 1$ and $-\|w_k - x_k\|^2 \leq 0$, we obtain

$$\begin{aligned} F(x_k) \leq a_k F(x) + (1-a_k)F(x_{k-1}) + \frac{\tilde{\mu} a_k^2}{2}(\|x - v_{k-1}\|^2 - \|x - v_k\|^2) \\ - \frac{\tilde{\mu} - \mu}{2}\|y_k - x_k\|^2 + \rho a_k \|x - x_{k-1}\|^2 + \frac{r a_k^2}{2}\|x - v_{k-1}\|^2. \end{aligned}$$

The proof is complete. □

The proof of Theorem 4.8 now quickly follows.

Proof of Theorem 4.8. Set $x = x^*$ in inequality (4.8). Rewriting (4.8) by subtracting F^* from both sides, we obtain

$$\begin{aligned} \frac{F(x_k) - F(x^*)}{a_k^2} + \frac{\tilde{\mu}}{2}\|x^* - v_k\|^2 \leq \frac{1-a_k}{a_k^2}\big(F(x_{k-1}) - F(x^*)\big) + \frac{\tilde{\mu}}{2}\|x^* - v_{k-1}\|^2 \\ + \frac{\rho M^2}{a_k} + \frac{r M^2}{2} - \frac{\tilde{\mu} - \mu}{2a_k^2}\|x_k - y_k\|^2. \end{aligned} \tag{4.12}$$

Using the inequality $\frac{1-a_k}{a_k^2} \leq \frac{1}{a_{k-1}^2}$ and recursively applying the inequality above N times, we get

$$\frac{F(x_N) - F(x^*)}{a_N^2} + \frac{\tilde{\mu}}{2}\|x^* - v_N\|^2 \leq \frac{1-a_1}{a_1^2}\big(F(x_0) - F(x^*)\big) + \frac{\tilde{\mu}}{2}\|x^* - v_0\|^2 + \rho M^2\left(\sum_{j=1}^{N}\frac{1}{a_j}\right) + \frac{NrM^2}{2} - \frac{\tilde{\mu}-\mu}{2}\sum_{j=1}^{N}\frac{\|x_j - y_j\|^2}{a_j^2}. \tag{4.13}$$

Noting $F(x_N) - F(x^*) > 0$ and $a_1 = 1$, we obtain

$$\frac{\tilde{\mu}-\mu}{2}\sum_{j=1}^{N}\frac{\|x_j - y_j\|^2}{a_j^2} \leq \frac{\tilde{\mu}}{2}\|x^* - v_0\|^2 + \rho M^2\left(\sum_{j=1}^{N}\frac{1}{a_j}\right) + \frac{NrM^2}{2} \tag{4.14}$$

and hence

$$\frac{\tilde{\mu}-\mu}{2}\left(\sum_{j=1}^{N}\frac{1}{a_j^2}\right)\min_{j=1,\dots,N}\|x_j - y_j\|^2 \leq \frac{\tilde{\mu}}{2}\|x^* - v_0\|^2 + \rho M^2\left(\sum_{j=1}^{N}\frac{1}{a_j}\right) + \frac{NrM^2}{2}.$$

Using the definition $a_k = \frac{2}{k+1}$, we conclude

$$\sum_{j=1}^{N}\frac{1}{a_j^2} = \frac{1}{4}\sum_{j=1}^{N}(j+1)^2 \geq \frac{1}{4}\sum_{j=1}^{N}j^2 = \frac{N(N+1)(2N+1)}{24}$$

and

$$\sum_{j=1}^{N}\frac{1}{a_j} = \sum_{j=1}^{N}\frac{j+1}{2} = \frac{N(N+3)}{4}.$$

With these bounds, we finally deduce

$$\min_{j=1,\dots N}\|x_j - y_j\|^2 \leq \frac{24}{\tilde{\mu}-\mu}\left(\frac{\tilde{\mu}\|x^* - v_0\|^2}{N(N+1)(2N+1)} + \frac{M^2(r + \frac{\rho}{2}(N+3))}{(N+1)(2N+1)}\right),$$

thereby establishing the first claimed efficiency estimate in Theorem 4.8.

Finally suppose $r = 0$, and hence we can assume $\rho = 0$ by Lemma 4.6. Inequality (4.13) then becomes

$$\frac{F(x_N) - F(x^*)}{a_N^2} + \frac{\tilde{\mu}}{2}\|x^* - v_N\|^2 \leq \frac{\tilde{\mu}}{2}\|x^* - v_0\|^2 - \frac{\tilde{\mu}-\mu}{2}\sum_{j=1}^{N}\frac{\|x_j - y_j\|^2}{a_j^2}.$$

Dropping terms, we deduce $\frac{F(x_N)-F(x^*)}{a_N^2} \leq \frac{\tilde{\mu}}{2}\|x^* - v_0\|^2$, and the claimed efficiency estimate follows. □

5 Inexact computation

In practice, it is often impossible to solve the proximal subproblems $\min_z F_{t,\alpha}(z; y, v)$ exactly. In this section, we explain the effect of inexactness in the proximal subproblems (4.3), (4.6), and (4.7) on the overall performance of the prox-linear algorithm and its accelerated variant. By "inexactness", one can mean a variety of concepts. Two most natural ones are that of (i) terminating the subproblems based on near-optimality in function value and (ii) terminating based on "near-stationarity".

Which of the two criteria is used depends on the algorithms that are available for solving the proximal subproblems. If primal-dual interior-point methods are applicable, then termination based on near-optimality in function value is most appropriate. When the subproblems themselves can only be solved by first-order methods, the situation is less clear. In particular, if near-optimality in function value is the goal, then one must use saddle-point methods. Efficiency estimates of saddle-point algorithms, on the other hand, depend on the diameter of the feasible region, rather than on the quality of the initial iterate. Thus saddle-point methods cannot be warm-started, that is one cannot easily use iterates from previous prox-linear subproblems to speed up the algorithm for the current subproblem. Moreover, there is a conceptual incompatibility of the (accelerated) prox-linear method with termination based in functional near-optimality. Indeed, the (accelerated) prox-linear method seeks to make the stationarity measure $\|\mathcal{G}_t(x)\|$ small, and so it seems more fitting that the proximal subproblems are solved based on near-stationarity themselves.

In this section, we consider both termination criteria. We will see, however, that termination criteria based on near-stationarity lead to more transparent efficiency estimates, simpler proofs, and superior overall efficiency guarantees. Proofs of the main theorems 5.2, 5.3, 5.5, and 5.6 are somewhat lengthy, building on the proofs of the analogous results in the exact regime. Consequently, we have placed the arguments in the appendix.

5.1 Near-stationarity in the subproblems

In this section, we consider a model of inexactness for the proximal subproblems based on near-stationarity. To motivate the discussion, consider the proximal subproblem

$$\min_z \; F_t(z; y) = h\Big(c(y) + \nabla c(y)(z - y)\Big) + g(z) + \frac{1}{2t}\|z - y\|^2. \tag{5.1}$$

A first naive attempt would be to consider a point ε-stationary for the subproblem (5.1) if it satisfies

$$\operatorname{dist}(0; \partial_z F_t(z; y)) \le \varepsilon.$$

This assumption, however, is not reasonable since first-order methods for (5.1) do not produce such points z, unless h is smooth. Instead, let us look at the Fenchel dual problem. To simply notation, write the target subproblem (5.1) as

$$\min_z \; h(b - Az) + G(z) \tag{5.2}$$

under the identification $G(z) = g(z) + \frac{1}{2t}\|z - y\|^2$, $A = -\nabla c(y)$, and $b = c(y) - \nabla c(y)y$. Notice that G is t^{-1}-strongly convex. The Fenchel dual problem, after negation, takes the form

$$\min_w \; \varphi(w) := G^\star(A^* w) - \langle b, w\rangle + h^\star(w).$$

Thus the dual objective function φ is a sum of a smooth convex function $G^\star(A^*w) - \langle b, w\rangle$ and the simple nonsmooth term $h^\star$. Typical first-order methods, such as prox-gradient and its accelerated variants can generate a point w for this problem satisfying

$$\operatorname{dist}(0; \partial\varphi(w)) \leq \varepsilon$$

up to any specified tolerance $\varepsilon > 0$. One can then use such a point w to generate an "approximate" minimizer of the original prox-linear subproblem. Indeed, given a vector $\zeta \in \partial\varphi(w)$ one can quickly check from primal-dual optimality conditions that the point $x := \nabla G^\star(A^*w)$ is a true minimizer of the slight perturbation of the target problem (5.2), namely

$$\min_z \; h(\zeta + b - Az) + G(z).$$

This motivates the following inexact extension of the prox-linear algorithm (Algorithm 4).

Algorithm 4: Inexact prox-linear method: near-stationarity

Initialize : A point $x_0 \in \operatorname{dom} g$, a real $t > 0$, and a sequence $\{\varepsilon_i\}_{i=1}^\infty \subset [0, +\infty)$.

Step k: $(k \geq 0)$ Let x_{k+1} be a minimizer of the function

$$z \mapsto g(z) + h\Big(\zeta_k + c(x_k) + \nabla c(x_k)(z - x_k)\Big) + \frac{1}{2t}\|z - x_k\|^2 \tag{5.3}$$

for some vector ζ_{k+1} satisfying $\|\zeta_{k+1}\| \leq \varepsilon_{k+1}$.

Before, stating convergence guarantees of the method, we record the following observation stating that the step-size $\|x_{k+1} - x_k\|$ and the error ε_{k+1} jointly control the stationarity measure $\|\mathcal{G}_t(x_k)\|$. In other words, one can use the step-size $\|x_{k+1} - x_k\|$, generated throughout the algorithm, as a surrogate for the true stationarity measure $\|\mathcal{G}_t(x_k)\|$.

Lemma 5.1. *Suppose z^+ is a minimizer of the function*

$$z \mapsto g(z) + h\Big(\zeta + c(y) + \nabla c(y)(z - y)\Big) + \frac{1}{2t}\|z - y\|^2$$

for some vector ζ. Then for any real $t > 0$, the inequality holds:

$$\|\mathcal{G}_t(y)\|^2 \leq 8Lt^{-1} \cdot \|\zeta\| + 2\left\|t^{-1}(z^+ - y)\right\|^2. \tag{5.4}$$

Proof. Define the function

$$l(z) = g(z) + h\Big(\zeta + c(y) + \nabla c(y)(z - y)\Big) + \frac{1}{2t}\|z - y\|^2.$$

Let z^* be the true minimizer of $F_t(\cdot; y)$. We successively deduce

$$\begin{aligned}
\|\mathcal{G}_t(y)\|^2 &\leq \frac{4}{t} \cdot \frac{1}{2t}\left\|z^+ - z^*\right\|^2 + 2\left\|t^{-1}(z^+ - y)\right\|^2 \\
&\leq \frac{4}{t} \cdot \big(F_t(z^+; y) - F_t(z^*; y)\big) + 2\left\|t^{-1}(z^+ - y)\right\|^2 \\
&\leq \frac{4}{t}(l(z^+) - l(z^*) + 2L\|\zeta\|) + 2\left\|t^{-1}(z^+ - y)\right\|^2 \\
&\leq 8t^{-1}L\|\zeta\| + 2\left\|t^{-1}(z^+ - y)\right\|^2,
\end{aligned} \tag{5.5}$$

where the first inequality follows from the triangle inequality and the estimate $(a+b)^2 \leq 2(a^2+b^2)$ for any reals a, b, the second inequality is an immediate consequence of strong convexity of the function $F_t(\cdot; y)$, and the third follows from Lipschitz continuity of h. □

Theorem 5.2 explains the convergence guarantees of the method; c.f. Proposition 4.1.

Theorem 5.2 (Convergence of the inexact prox-linear method: near-stationarity)**.** *Supposing $t \leq \mu^{-1}$, the iterates generated by Algorithm 4 satisfy*

$$\min_{j=0,\ldots,N-1} \|\mathcal{G}_t(x_j)\|^2 \leq \frac{4t^{-1}\big(F(x_0) - F^* + 4L \cdot \sum_{j=1}^N \varepsilon_j\big)}{N},$$

where we set $F^* := \lim_{k\to\infty} F(x_k)$.

In particular, to maintain the same rate in N as the exact prox-linear method in Proposition 4.1, we must be sure that the sequence ε_k is summable. Hence, we can set $\varepsilon_k \sim \frac{1}{k^{1+q}}$ for any $q > 0$.

Completely analogously, we can consider an inexact accelerated prox-linear method (Algorithm 5).

Algorithm 5: Inexact accelerated prox-linear method: near-stationarity

Initialize : Fix two points $x_0, v_0 \in \operatorname{dom} g$ and a real number $\tilde{\mu} > \mu$.

Step k: ($k \geq 1$) Compute

$$a_k = \tfrac{2}{k+1}$$
$$y_k = a_k v_{k-1} + (1 - a_k)x_{k-1}$$

- Let x_k be a minimizer of the function

$$z \mapsto g(z) + h\Big(\zeta_k + c(y_k) + \nabla c(y_k)(z - y_k)\Big) + \frac{\tilde{\mu}}{2}\|z - y_k\|^2 \tag{5.6}$$

for some vector ζ_k satisfying $\|\zeta_k\| \leq \varepsilon_k$.

- Let v_k be a minimizer of the function

$$v \mapsto g(v) + \frac{1}{a_k} h\Big(\xi_k + c(y_k) + a_k \nabla c(y_k)(v - v_{k-1})\Big) + \frac{\tilde{\mu} a_k}{2}\|v - v_{k-1}\|^2 \tag{5.7}$$

for some vector ξ_k satisfying $\|\xi_k\| \leq \delta_k$.

Theorem 5.3 (Convergence of inexact accelerated prox-linear method: near- stationarity)**.** *Fix a real number $\tilde{\mu} \geq \mu$ and let x^* be any point satisfying $F(x^*) < F(x_k)$ for all x_k generated by Algorithm 5. Then for any $N \geq 1$, the iterates generated by Algorithm 5 satisfy the inequality:*

$$\min_{i=1,\ldots,N} \|\mathcal{G}_{1/\tilde{\mu}}(y_j)\|^2 \leq \frac{48\tilde{\mu}^2}{\tilde{\mu} - \mu}\left(\frac{\|x^* - v_0\|^2}{N(N+1)(2N+1)} + \frac{M^2(r + \frac{\rho}{2}(N+3))}{(N+1)(2N+1)} + \frac{4L\sum_{j=1}^N \frac{2\varepsilon_j + \delta_j}{a_j^2}}{N(N+1)(2N+1)}\right).$$

Moreover, in the case $r = 0$, the inequality above holds with the second summand on the right-hand-side replaced by zero (even if $M = \infty$) and the following complexity bound on function values holds:

$$F(x_N) - F^* \leq \frac{2\tilde{\mu}\|v_0 - x^*\|^2 + 8L\sum_{j=1}^{N} \frac{\varepsilon_j + \delta_j}{a_j^2}}{(N+1)^2}.$$

Thus to preserve the rate in N of the exact accelerated prox-linear method in Theorem 4.8, we must require the sequences $\frac{\varepsilon_j}{a_j^2}, \frac{\delta_j}{a_j^2}$ to be summable. Hence we can set $\varepsilon_j, \delta_j \sim \frac{1}{j^{3+q}}$ for some $q > 0$.

5.2 Near-optimality in the subproblems

Next, we consider the effect of solving the proximal subproblems up to a tolerance on function values. Given a tolerance $\varepsilon > 0$, we say that a point x is an *ε-approximate minimizer* of a function $f \colon \mathbf{R}^n \to \overline{\mathbf{R}}$ whenever the inequality holds:

$$f(x) \leq \inf f + \varepsilon.$$

Consider now a sequence of tolerances $\varepsilon_k > 0$ for $k = 1, 2 \ldots, \infty$. Then given a current iterate x_k, an *inexact prox-linear algorithm* for minimizing F can simply declare x_{k+1} to be an ε_{k+1}-approximate minimizer of $F_t(\cdot; x_k)$.

Algorithm 6: Inexact prox-linear method: near-optimality
Initialize : A point $x_0 \in \operatorname{dom} g$, a real $t > 0$, and a sequence $\{\varepsilon_i \geq 0\}_{i=1}^{\infty} \subset [0, +\infty)$.
Step k: ($k \geq 0$) Set x_{k+1} to be a ε_{k+1}-approximate minimizer of $F_t(\cdot; x_k)$.

Before, stating convergence guarantees of the method, we record the following observation stating that the step-size of the inexact prox-linear method $\|x_{k+1} - x_k\|$ and the accuracy ε_k jointly control the size of the true prox-gradient $\|\mathcal{G}_t(x_k)\|$. As a consequence, the step-sizes $\|x_{k+1} - x_k\|$ generated throughout the algorithm can be use as surrogates for the true stationarity measure $\|\mathcal{G}_t(x_k)\|$.

Lemma 5.4. *Suppose x^+ is an ε-approximate minimizer of $F_t(\cdot; y)$. Then the inequality holds:*

$$\|\mathcal{G}_t(y)\|^2 \leq 4t^{-1}\varepsilon + 2\left\|t^{-1}(x^+ - y)\right\|^2.$$

Proof. Let x^* be the true minimizer of $F_t(\cdot, y)$. We successively deduce

$$\begin{aligned} \|\mathcal{G}_t(y)\|^2 &\leq \frac{4}{t} \cdot \frac{1}{2t}\left\|x^+ - x^*\right\|^2 + 2\left\|t^{-1}(x^+ - y)\right\|^2 \\ &\leq \frac{4}{t} \cdot \big(F_t(x^+; y) - F_t(x^*; y)\big) + 2\left\|t^{-1}(x^+ - y)\right\|^2 \\ &\leq \frac{4}{t} \cdot \varepsilon + 2\left\|t^{-1}(x^+ - y)\right\|^2, \end{aligned} \tag{5.8}$$

where the first inequality follows from the triangle inequality and the estimate $(a+b)^2 \leq 2(a^2+b^2)$ for any reals a, b, and the second inequality is an immediate consequence of strong convexity of the function $F_t(\cdot; y)$. $\square$

The inexact prox-linear algorithm comes equipped with the following guarantee.

Theorem 5.5 (Convergence of the inexact prox-linear algorithm: near-optimality)**.** *Supposing $t \leq \mu^{-1}$, the iterates generated by Algorithm 6 satisfy*

$$\min_{j=0,\dots,N-1} \|\mathcal{G}_t(x_j)\|^2 \leq \frac{4t^{-1}\big(F(x_0) - F^* + 4L\sum_{j=1}^N \varepsilon_j\big)}{N},$$

where we set $F^* := \lim_{k\to\infty} F(x_k)$.

Thus in order to maintain the rate afforded by the exact prox-linear method, it suffices for the errors $\{\varepsilon_k\}_{k=1}^\infty$ to be summable; e.g. set $\varepsilon_k \sim \frac{1}{k^{1+q}}$ with $q > 0$.

Similarly, we can consider an inexact version of the accelerated prox-linear method. The scheme is recorded in Algorithm 7.

Algorithm 7: Accelerated prox-linear method: near-optimality

Initialize : Fix two points $x_0, v_0 \in \operatorname{dom} g$, a real number $\tilde{\mu} > L\beta$, and two sequences $\varepsilon_i, \delta_i \geq 0$ for $i = 1, 2, \dots, \infty$.

Step k: ($k \geq 1$) Compute

$$a_k = \tfrac{2}{k+1} \tag{5.9}$$

$$y_k = a_k v_{k-1} + (1 - a_k)x_{k-1} \tag{5.10}$$

$$\text{Set } x_k \text{ to be a } \varepsilon_k\text{-approximate minimizer of } F_{1/\tilde{\mu}}(\cdot; y_k) \tag{5.11}$$

$$\text{Set } v_k \text{ to be a } \delta_k\text{-approximate minimizer of } F_{\frac{1}{\tilde{\mu}a_k}, a_k}(\cdot; y_k, v_{k-1}) \tag{5.12}$$

Theorem 5.6 presents convergence guarantees of Algorithm 7. The statement is much more cumbersome than the analogues Theorem 5.3. The only take-away message for the reader is that to preserve the rate of the exact accelerated prox-linear method in Theorem 4.8 in terms of N, the sequences $i^2\varepsilon_i$ and $i^2\delta_i$ must be summable. Thus it suffices to take $\varepsilon_i, \delta_i \sim \frac{1}{i^{3+q}}$ for some $q > 0$.

Theorem 5.6 (Convergence of the accelerated prox-linear algorithm: near-optimality)**.** *Fix a real number $\tilde{\mu} > \mu$, and let x^* be any point satisfying $F(x^*) < F(x_k)$ for all x_k generated by Algorithm 7. Then the iterates generated by Algorithm 7 satisfy the inequality:*

$$\begin{aligned}\min_{i=1,\dots,N} \|\mathcal{G}_{1/\tilde{\mu}}(y_i)\|^2 \leq & \frac{96\tilde{\mu}^2}{\tilde{\mu}-\mu}\Big(\frac{\tilde{\mu}\|x^* - v_0\|^2}{2N(N+1)(2N+1)} + + \frac{M^2(r + \frac{\rho}{2}(N+3))}{2(N+1)(2N+1)} \\ & + \frac{\sum_{i=1}^N (\frac{\delta_i a_i + 3\varepsilon_i}{a_i^2}) + A_N\sqrt{2\tilde{\mu}}\sum_{i=1}^N \sqrt{\frac{\delta_i}{a_i}}}{N(N+1)(2N+1)}\Big)\end{aligned}$$

with

$$A_N := \sqrt{\tfrac{2}{\tilde{\mu}}} \sum_{i=1}^{N} \sqrt{\tfrac{\delta_i}{a_i}} \quad + \left(\|x^* - v_0\|^2 + \tfrac{M^2 N (r + \frac{\rho}{2}(N+3))}{\tilde{\mu}} + \tfrac{2}{\tilde{\mu}} \sum_{i=1}^{N} \tfrac{\delta_i a_i + 2\varepsilon_i}{a_i^2} + \tfrac{2}{\tilde{\mu}} \left(\sum_{i=1}^{N} \sqrt{\tfrac{\delta_i}{a_i}} \right)^2 \right)^{1/2}.$$

Moreover, in the case $r = 0$, the inequality above holds with the second summand on the right-hand-side replaced by zero (even if $M = \infty$), and the following complexity bound on function values holds:

$$F(x_N) - F(x^*) \leq \frac{2\tilde{\mu}\,\|x^* - v_0\|^2 + 4\sum_{i=1}^{N} \frac{\delta_i a_i + 2\varepsilon_i}{a_i^2} + 4A_N \sqrt{2\tilde{\mu}} \sum_{i=1}^{N} \sqrt{\frac{\delta_i}{a_i}}}{(N+1)^2}.$$

6 Overall efficiency estimates of first-order methods

In light of the results in the previous sections, we can now derive efficiency estimates for the (accelerated) prox-linear method, where the proximal subproblems are themselves solved by first-order methods. As is standard, we will assume that the functions h and g are *prox-friendly*, meaning that prox_{th} and prox_{tg} can be evaluated. Given a target accuracy $\varepsilon > 0$, we aim to determine the number of *basic operations* – matrix-vector multiplications, evaluations of prox_{th}, prox_{tg} – needed to find a point x satisfying $\|\mathcal{G}_t(x)\|^2 \leq \epsilon$.

Duality will play a central role in this section. To ease notation, let us rewrite the target proximal subproblem

$$\min_z \; F_{t,\alpha}(z; y, v) = g(z) + \frac{1}{\alpha} h\Big(c(y) + \alpha \nabla c(y)(z - v)\Big) + \frac{1}{2t}\|z - v\|^2$$

in the form

$$\min_z \; G(z) + H\big(b - Az\big) \tag{6.1}$$

under the identification $G(z) = g(z) + \frac{1}{2t}\|z - v\|^2$, $H = \alpha^{-1} h$, $A = -\alpha \nabla c(y)$, and $b = c(y) - \alpha \nabla c(y) v$. The Fenchel dual of (6.1), after negation, is then given by

$$\min_w \; \varphi(w) := G^\star(A^* w) - \langle b, w \rangle + H^\star(w) \tag{6.2}$$

6.1 When h is smooth

We begin with the setting when h is smooth with Lipschitz continuous gradient.

Cost of finding approximate minimizers of the subproblems

Let us first look at the cost of solving the proximal subproblems (6.1) based on tolerance on functional error (Section 5.2). Notice that the objective function in (6.1) is a sum of the strongly convex and prox-friendly term G and the smooth convex function $z \mapsto H(b - Az)$. The function G is t^{-1}-strongly convex, while the gradient of the smooth term $z \mapsto H(b - Az)$ is Lipschitz continuous with constant $\alpha \|\nabla c(y)\|^2 \,\mathrm{lip}\,(\nabla h)$. Consequently, we can apply any linearly convergent fast gradient method to the formulation (6.1). For example, the method in [37, Section 4] will find an ε-approximate minimizer z of $F_{t,\alpha}(\cdot; y, v)$ after at most

$$\mathcal{O}\left(\sqrt{t\alpha \|\nabla c(y)\|^2 \,\mathrm{lip}\,(\nabla h)} \cdot \ln\left(\frac{\alpha \|\nabla c(y)\| \,\mathrm{lip}\,(\nabla h) \|z_0 - z^*\|}{\varepsilon} \right) \right) \qquad \text{iterations.} \tag{6.3}$$

Cost of finding near-stationary points in the subproblems

Next let us look at the near-stationarity model of inexactness as in Section 5.1. Following the recipe presented there, consider the (negated) Fenchel dual (6.2). The function $w \mapsto G^\star(A^*w) - \langle b, w\rangle$ is C^1-smooth with gradient having Lipschitz constant $\alpha^2\|\nabla c(y)\|^2 t$. The nonsmooth term $(\alpha^{-1}h)^\star$ is strongly convex with constant $\alpha/\operatorname{lip}(\nabla h)$. Thus we can apply a linearly convergent fast-gradient method, which is guaranteed to find a point w with $\operatorname{dist}(0; \partial\varphi(w)) \leq \varepsilon$ after at most

$$\mathcal{O}\left(\sqrt{t\alpha\|\nabla c(y)\|^2 \operatorname{lip}(\nabla h)} \cdot \ln\left(\frac{t\alpha\|\nabla c(y)\| \operatorname{lip}(\nabla h)\|w_0 - w^*\|}{\varepsilon}\right)\right) \qquad \text{iterations.} \tag{6.4}$$

Letting $\xi \in \partial\varphi(w)$ satisfy $\|\xi\| \leq \varepsilon$, the point $x := \nabla G^\star(A^*w)$ is the true minimizer of the function

$$z \mapsto g(z) + \frac{1}{\alpha}h\Big(\xi + c(y) + \alpha\nabla c(y)(z - v)\Big) + \frac{1}{2t}\|z - v\|^2.$$

Total cost of the (accelerated) prox-linear method

We can now interpret the overall costs of the inexact prox-linear methods (Algorithms 4 and 6). For both algorithms, we set $\alpha = 1$ and $t = \mu^{-1}$. Theorems 5.2 and 5.5 show that if we set $\varepsilon_i = \frac{1}{Li^2}$ in each iteration i of either Algorithm 4 or 6, then after N iterations, both algorithms enjoy the efficiency estimate

$$\min_{j=0,\ldots,N-1} \left\|\mathcal{G}_{\frac{1}{\mu}}(x_j)\right\|^2 \leq \frac{4\mu\big(F(x_0) - F^* + 8\big)}{N}. \tag{6.5}$$

According to (6.4), each iteration of Algorithm 4 requires at most

$$\mathcal{O}\left(\sqrt{\frac{\|\nabla c\|^2 \cdot \operatorname{lip}(\nabla h)}{\mu}} \cdot \ln\left(\frac{\|\nabla c\| \operatorname{lip}(\nabla h) LN}{\mu}\right)\right) \tag{6.6}$$

basic operations, while according to (6.3) each iteration of Algorithm 6 requires at most

$$\mathcal{O}\left(\sqrt{\frac{\|\nabla c\|^2 \cdot \operatorname{lip}(\nabla h)}{\mu}} \cdot \ln\left(\frac{\|\nabla c\| \operatorname{lip}(\nabla h) LMN}{\mu}\right)\right) \tag{6.7}$$

iterations. Notice that M, the diameter of dom g, does not appear in the bound (6.6). Thus we have established the following.

Theorem 6.1 (Overall cost of the prox-linear method)**.** *There exists an algorithm for the problem* $\min_x F(x)$*, which given any* $\varepsilon > 0$ *is guaranteed to find a point* x *satisfying* $\|\mathcal{G}_{\frac{1}{\mu}}(x)\|^2 \leq \varepsilon$ *after at most*

$$\widetilde{\mathcal{O}}\left(\frac{\sqrt{\|\nabla c\|^2 \cdot \operatorname{lip}(\nabla h)\mu} \cdot \big(F(x_0) - F^*\big)}{\varepsilon}\right) \qquad \textit{basic operations,}$$

where $\widetilde{\mathcal{O}}$ *hides universal constants and logarithmic terms in* $\varepsilon, \|\nabla c\|, \operatorname{lip}(\nabla h), \mu$*, and* $F(x_0) - F^*$*.*

An entirely analogous analysis applies for the accelerated prox-linear methods (Algorithms 5 and 7). In this case, we set $t = \frac{1}{2\mu a_k}$ and $\alpha = a_k$. Theorems 5.3 and 5.6 then show that if we set $\varepsilon_i, \delta_i = \frac{1}{LN^4}$ then the inexact methods will enjoy the same efficiency estimates as the exact accelerated prox-linear method in Theorem 4.8. Plugging in these values into (6.3) and (6.4) we deduce that obtaining a ε_N-approximate stationary point and a ε_N-approximate minimizer of the proximal subproblem $\min F_{t,\alpha}(\cdot; y, v)$ requires no more basic operations than the estimates in (6.6) and (6.7), respectively.

6.2 When h is nonsmooth

Suppose now that h is nonsmooth. In this setting, the proximal subproblems (6.1) can not be solved by linearly convergent algorithms.

Cost of finding approximate minimizers of the subproblems

In order to find approximate minimizers of the subproblems (6.1), we can rely on (fast) primal-dual methods with sublinear guarantees. We follow the exposition in [45]. To explain the type of methods that are available, consider a general problem of the form

$$\min_x \; f^p(x) = f(Bx) + p(x) \tag{6.8}$$

with f and p closed, convex functions, and B a linear map. Suppose moreover that f is C^1-smooth and the inequality

$$\max_{x,y \in \operatorname{dom} p} \|y - x\| \leq R$$

holds for some finite number R. Shortly, we will identify this formulation with the target problem (6.2). The Fenchel dual of (6.8) is given by

$$\max_v \; q^p(v) := -f^\star(v) - p^\star(-B^T v).$$

We will shortly identify this dual with the proximal subproblem (6.1). Consider the following algorithm:

Algorithm 8: A primal-dual optimal method (Nesterov [33], Auslender-Teboule [2])

Initialize : Fix two points $x_0, z_0 \in \operatorname{dom} p$, set $l := \|B\|^2 \cdot \operatorname{lip}(\nabla f)$ and $a_0 = 1$.
Step k: $(k \geq 0)$ Compute

$$\begin{aligned}
y_k &= (1 - a_k)x_k + a_k z_k \\
z_{k+1} &= \operatorname{prox}_{\frac{p}{a_k l}}\left(z_k - \tfrac{1}{a_k l}\nabla f(y_k)\right) \\
x_{k+1} &= (1 - a_k)x_k + a_k z_{k+1} \\
v_k &= (1 - a_k)v_{k-1} + a_k \nabla f(By_k) \\
a_{k+1} &= \tfrac{\sqrt{a_k^4 + 4a_k^2} - a_k^2}{2}
\end{aligned}$$

The scheme comes equipped with the following guarantee; see e.g. Auslender-Teboule [2] or Tseng [45].

Theorem 6.2. *For every index k, the inequality holds:*

$$0 \leq f^p(x_{k+1}) - q^p(v_k) \leq \frac{2\|B\|^2 \cdot \operatorname{lip}(\nabla f) \cdot R^2}{(k+2)^2}.$$

Returning to the primal dual pair, (6.1) and (6.2), we can set $B = A^* = -\alpha \nabla c(y)^*$, $f = G^\star$, and $p = H^\star - \langle b, \cdot \rangle$. We then have $\operatorname{lip}(\nabla f) = \operatorname{lip}(\nabla G^*) = t$ and $R \leq L/\alpha$. We thus deduce that we can find an ε-approximate minimizer z of $F_{t,\alpha}(\cdot; y, v)$ after at most on the order of

$$\sqrt{\frac{\|\nabla c(y)\|^2 L^2 t}{\varepsilon}} \tag{6.9}$$

iterations of Algorithm 8.

Cost of finding near-stationary points in the subproblems

Let us return again to the formulation (6.8). We seek a fast algorithm that generates a point x satisfying $\operatorname{dist}(0; \partial f^p(x)) \leq \varepsilon$. To this end, consider minimizing the strongly convex objective $\widehat{f^p}(x) := f^p(x) + \frac{\varepsilon}{4R}\|x - x_0\|^2$ for some point $x_0 \in \operatorname{dom} p$. Suppose we could find a point x satisfying $\operatorname{dist}(0; \partial \widehat{f^p}(x)) \leq \frac{\varepsilon}{2}$. Then we would deduce $\operatorname{dist}(0; \partial f^p(x)) \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2R}\|x - x_0\| \leq \varepsilon$, as desired. Since $\widehat{f^p}$ is strongly convex with parameter $\frac{\varepsilon}{2R}$, there are fast gradient methods that can find such a point x in

$$\mathcal{O}\left(\sqrt{\frac{R\|B\|^2 \operatorname{lip}(\nabla f)}{\varepsilon}} \cdot \ln\left(\frac{R\|B\| \operatorname{lip}(\nabla f)}{\varepsilon}\right)\right)$$

iterations. See the discussion in [34]. Specializing to the problem (6.2), we deduce that we can find a point w satisfying $\operatorname{dist}(0; \partial\varphi(w)) \leq \varepsilon$ in at most

$$\mathcal{O}\left(\sqrt{\frac{\alpha t L \|\nabla c(y)\|^2}{\varepsilon}} \cdot \ln\left(\frac{\alpha t L \|\nabla c(y)\|}{\varepsilon}\right)\right) \tag{6.10}$$

iterations.

Total cost of the (accelerated) prox-linear method

We can now interpret the overall costs of the inexact prox-linear methods (Algorithms 4 and 6). Fix a real $q > 0$ and set $\varepsilon_i := \frac{1}{Li^{1+q}}$ in each iteration i of either Algorithm 4 or 6. Then after N iterations, both algorithms enjoy the efficiency estimate

$$\min_{j=0,\ldots,N-1} \left\|\mathcal{G}_{\frac{1}{\mu}}(x_j)\right\|^2 \leq \frac{4\mu\big(F(x_0) - F^* + 4(1+q^{-1})\big)}{N}. \tag{6.11}$$

Setting $t = \mu$, each iteration of Algorithm 6 requires at most

$$\sqrt{\frac{\|\nabla c\|^2 L^2 \cdot i^{1+q}}{\beta}}$$

iterations, while each iteration of Algorithm 4 requires at most

$$\mathcal{O}\left(\sqrt{\frac{\|\nabla c\|^2 L^2 \cdot i^{1+q}}{\beta}} \cdot \ln\left(\frac{\|\nabla c\| L \cdot i^{1+q}}{\beta}\right)\right)$$

iterations. Thus, we have established the following.

Theorem 6.3 (Overall cost of the prox-linear method)**.** *There exists an algorithm for the problem* $\min_x F(x)$*, which given any* $\varepsilon > 0$ *and* $q > 0$ *is guaranteed to find a point* x *satisfying* $\|\mathcal{G}_{\frac{1}{\mu}}(x)\|^2 \leq \varepsilon$ *after at most*

$$\widetilde{\mathcal{O}}\left(\sqrt{\frac{\|\nabla c\|^2 L^2}{\beta}} \cdot \left(\frac{\mu(F(x_0) - F^* + 4(1+q^{-1}))}{\varepsilon}\right)^{\frac{3+q}{2}}\right) \qquad \textit{basic operations,}$$

where $\widetilde{\mathcal{O}}$ *hides universal constants and logarithmic terms in* $\varepsilon, \|\nabla c\|, \mu$*, and* $F(x_0) - F^*$*.*

Let us look at the accelerated prox-linear methods (Algorithms 5 and 7). In this case, we set $t = \frac{1}{2\mu a_k}$ and $\alpha = a_k$. Theorems 5.3 and 5.6 then show that if we set $\varepsilon_i, \delta_i = \frac{1}{Li^{3+q}}$ then the inexact methods will enjoy the same efficiency estimates as the exact accelerated prox-linear method in Theorem 4.8. Plugging in these values into (6.9) and (6.10) we deduce that each subproblem of Algorithms 7 and Algorithm 5 requires at most

$$\mathcal{O}\left(\sqrt{\frac{\|\nabla c\|^2 L^2 \cdot i^{4+q}}{\beta}}\right) \qquad \text{and} \qquad \widetilde{\mathcal{O}}\left(\sqrt{\frac{\|\nabla c\|^2 L \cdot i^{3+q}}{\beta}}\right)$$

basic operations, respectively. We deduce that Algorithm 5 is superior to Algorithm 7.

7 Efficiency of smoothing

It is appealing to now ask how the convergence guarantees of the inexact (accelerated) prox-linear method compare with other algorithms for the composite problem class. For example, one such competitor is based on smoothing h and then simply applying the prox-gradient method to the smoothed problem. There is, however, an important and necessary ingredient for a well-grounded comparison, which we have yet to discuss.

To illustrate, consider the setting when h is already smooth and $g = 0$. Then as an alternative to the prox-linear method, we can apply gradient descent directly to the smooth function $F(x) = h(c(x))$, yielding the recurrence

$$x_{k+1} = x_k - \gamma \cdot \nabla F(x_k)$$

for an appropriate parameter $\gamma > 0$. The gradient descent method drives the gradient norm $\|\nabla F\|$ to zero at a certain rate, while the prox-linear method drives the prox-gradient $\|\mathcal{G}_t\|$ to zero. Therefore a comparison of gradient descent and the prox-linear method requires a comparison of the two different measures of stationarity, $\|\mathcal{G}_t\|$ and $\|\nabla F\|$. The following section explains this relationship, and sets forward a strategy for comparing the prox-linear method to competitors. Sections 7.2 and 7.3 then implement the strategy on simple smoothing algorithms. The final Section 7.4 describes a strategy combining smoothing and the prox-linear method, yielding best known efficiency estimates.

7.1 Step-size and stationarity

It is instructive to consider first the additive composite setting (3.2), where the relationship between $\|\mathcal{G}_t\|$ and ∂F is immediate. Here, the prox-linear method reduces to the prox-gradient recurrence

$$x_{k+1} = \operatorname{prox}_{g/\beta}\left(x_k - \frac{1}{\beta}\cdot\nabla c(x_k)\right).$$

First-order optimality conditions for the proximal subproblems amount to the inclusion

$$\mathcal{G}_{\frac{1}{\beta}}(x_k) \in \nabla c(x_k) + \partial g(x_{k+1}),$$

or equivalently

$$\mathcal{G}_{\frac{1}{\beta}}(x_k) + (\nabla c(x_{k+1}) - \nabla c(x_k)) \in \nabla c(x_{k+1}) + \partial g(x_{k+1}),$$

Taking into account that ∇c is β-Lipschitz continuous, we deduce

$$\operatorname{dist}(0; \partial F(x_{k+1})) \leq 2\|G_{\frac{1}{\beta}}(x_k)\|. \tag{7.1}$$

Thus a small prox-gradient $\|G_{\frac{1}{\beta}}(x_k)\|$ certifies that x_{k+1} is nearly stationary for F. For the general composite class (1.1), i.e. when h is not the identity, the relationship between the prox-gradient $\|G_{\frac{1}{\beta}}(x_k)\|$ and near-stationarity is less immediate. Indeed, when h is nonsmooth, the quantity $\operatorname{dist}(0; \partial F(x_{k+1}))$ will typically not even tend to zero in the limit, while $\|G_{\frac{1}{\beta}}(x_k)\|$ will tend to zero. For example, the prox-linear algorithm, when applied to the univariate function $f(x) = |\frac{1}{2}x^2 + x|$ and initiated to the right of the origin, will generate a sequence $x_k \to 0$ with $|f'(x_k)| \to 1$.

An appropriate comparison between the prox-gradient and stationarity relies on the observation that the prox-linear algorithm is an approximation to the true *proximal point algorithm* on the function F:

$$z_{k+1} = \operatorname{prox}_{tF}(z_k).$$

The precise statement stems from the following basic result, called the smooth variational principle in [5, Theorem 2.4.1]; we provide a quick proof for completeness.

Theorem 7.1 (Smooth variational principle)**.** *Consider a closed function $f\colon \mathbf{R}^n \to \overline{\mathbf{R}}$ and suppose the inequality $f(x) - f^* \leq \varepsilon$ holds for some point x and real $\varepsilon > 0$. Then for any $\rho > 0$, the inequality holds:*

$$\|\rho^{-1}(x - \operatorname{prox}_{\rho f}(x))\| \leq \sqrt{\frac{2\varepsilon}{\rho}}$$

If f is convex, then the estimate slightly improves to

$$\|\rho^{-1}(x - \operatorname{prox}_{\rho f}(x))\| \leq \sqrt{\frac{\varepsilon}{\rho}}$$

Proof. Fix a point $y \in \operatorname*{argmin}_z \left\{ f(z) + \frac{1}{2\rho}\|z - x\|^2 \right\}$. We deduce

$$f(y) + \frac{1}{2\rho}\|y - x\|^2 \leq f(x) \leq f^* + \varepsilon \leq f(y) + \varepsilon.$$

Hence we deduce $\rho^{-1}\|y-x\| \leq \sqrt{\frac{2\varepsilon}{\rho}}$, as claimed. If f is convex, then strong convexity of the function $z \mapsto f(z) + \frac{1}{2\rho}\|z-x\|^2$ implies instead

$$\left(f(y) + \frac{1}{2\rho}\|y-x\|^2\right) + \frac{1}{2\rho}\|y-x\|^2 \leq f(x) \leq f^* + \varepsilon \leq f(y) + \varepsilon.$$

The claimed inequality follows along the same lines. □

We can now quantify the precise relationship between the step-size taken by the prox-linear method and the true proximal point algorithm.

Theorem 7.2 (Prox-gradient and near-stationarity)**.** *For any point x, the inequality holds:*

$$4 \cdot \left\| \mathcal{G}_{\frac{1}{\mu}}(x) \right\| \geq \left\| 2\mu(x - \operatorname{prox}_{\frac{F}{2\mu}}(x)) \right\|.$$

Proof. For any point z, we successively deduce

$$\begin{aligned} F(z) &\geq F_{\frac{1}{\mu}}(z;x) - \mu\|z-x\|^2 \geq F_{\frac{1}{\mu}}(x^t;x) + \tfrac{\mu}{2}\|x^t - z\|^2 - \mu\|z-x\|^2 \\ &\geq F(x^t) + \frac{\mu}{2}\|x^t - z\|^2 - \mu\|z-x\|^2. \end{aligned} \tag{7.2}$$

Define the function $\zeta(z) := F(z) + \mu\|z-x\|^2 - \frac{\mu}{2}\|x^t - z\|^2$ and notice that ζ is convex. Inequality (7.2) directly implies

$$\zeta(x^t) - \inf \zeta \leq (F(x^t) + \mu\|x^t - x\|^2) - F(x^t) \leq \mu\|x^t - x\|^2.$$

Notice the equality, $\operatorname{prox}_{\frac{\zeta}{\mu}}(x^t) = \operatorname{prox}_{\frac{F}{2\mu}}(x)$. Setting $\rho := \mu^{-1}$ and $\varepsilon := \mu\|x^t - x\|^2$ and using Theorem 7.1, we conclude

$$\mu\|x^t - x\| \geq \|\mu(x^t - \operatorname{prox}_{\frac{F}{2\mu}}(x))\| \geq \|\mu(x - \operatorname{prox}_{\frac{F}{2\mu}}(x))\| - \|\mu(x^t - x)\|.$$

Rearranging, the result follows. □

A closely related result has recently appeared in [18, Theorem 5.3], and has been extended to general Taylor-like approximations in [17]. An immediate consequence of Theorem 7.2 is that for any point x, there exists a point $\hat{x}$ (namely $\hat{x} = \operatorname{prox}_{F/2\mu}(x)$) satisfying

$$\begin{cases} \|\hat{x} - x\| & \leq \frac{2}{\mu}\|\mathcal{G}_{1/\mu}(x)\|, \\ \operatorname{dist}(0; \partial F(\hat{x})) & \leq 4\|\mathcal{G}_{1/\mu}(x)\|. \end{cases} \tag{7.3}$$

Thus if $\|\mathcal{G}_{1/\mu}(x)\|$ is small, the point x is "near" some point $\hat{x}$ that is "nearly-stationary" for F. Notice that $\hat{x}$ is not computable, since it requires evaluation of $\operatorname{prox}_{F/2\mu}$. Computing $\hat{x}$ is not the point, however; the sole purpose of $\hat{x}$ is to certify that x is nearly stationary in the sense above.

The outlined viewpoint naturally motivates an appropriate criterion on which to base the comparison of the prox-linear method and competitors. Define the set of *ε-stationary points*:

$$\mathcal{C}_\varepsilon = \{z \in \mathbf{R}^n : \operatorname{dist}(0; \partial F(z)) \leq \varepsilon\}. \tag{7.4}$$

We will be interested in computing the number of iterations required by an algorithm to find a point x satisfying

$$\mu \cdot \operatorname{dist}(x; \mathcal{C}_\varepsilon) \leq \varepsilon. \tag{7.5}$$

7.2 When h is smooth: prox-gradient vs. prox-linear algorithms

Supposing that h is smooth, let us now compare the inexact prox-linear algorithm with the prox-gradient method. Classically, the iterates generated by the prox-gradient algorithm satisfy the efficiency estimate

$$\min_{j=0,\dots,k-1} \|t^{-1}(x_k - x_{k+1})\|^2 \leq \frac{2\operatorname{lip}(\nabla(h\circ c))\cdot\big(F(x_0)-F^*\big)}{k},$$

where we set $t = \operatorname{lip}(\nabla(h\circ c))$. In particular, by the discussion above, the prox-gradient method will find a point x satisfying $\operatorname{dist}(0;\partial F(x)) \leq \varepsilon$ after at most

$$\mathcal{O}\left(\frac{\operatorname{lip}(\nabla(h\circ c))\cdot\big(F(x_0)-F^*\big)}{\varepsilon^2}\right)$$

iterations. Clearly then x trivially satisfies the desired condition (7.5).

Lemma 7.3. *The inequality,* $\operatorname{lip}(\nabla(h\circ c)) \leq \|\nabla c\|^2 \operatorname{lip}(\nabla h) + \mu$, *holds.*

Proof. Observe

$$\begin{aligned}\|\nabla(h\circ c)(y) - \nabla(h\circ c)(x)\| &= \|\nabla c(y)^*\nabla h(c(y)) - \nabla c(x)^*\nabla h(c(x))\| \\ &= \|\nabla c(y)^*(\nabla h(c(y)) - \nabla h(c(x))) + (\nabla c(y) - \nabla c(x))^*\nabla h(c(x))\| \\ &\leq \big(\|\nabla c\|^2 \operatorname{lip}(\nabla h) + \mu\big)\,\|y-x\|,\end{aligned}$$

as claimed. □

Thus, the efficiency estimate of the prox-gradient method becomes

$$\mathcal{O}\left(\frac{(\|\nabla c\|^2\cdot\operatorname{lip}(\nabla h)+\mu)\cdot(F(x_0)-F^*))}{\varepsilon^2}\right) \qquad \text{iterations.}$$

In contrast, taking into account (7.3), we deduce that the inexact prox-linear method described in Section 6.1 will find a point x satisfying (7.5) after at most

$$\widetilde{\mathcal{O}}\left(\frac{\sqrt{\|\nabla c\|^2\cdot\operatorname{lip}(\nabla h)\mu}\cdot\big(F(x_0)-F^*\big)}{\varepsilon^2}\right) \qquad \text{basic operations.}$$

Since the quantities $\sqrt{\|\nabla c\|^2\cdot\operatorname{lip}(\nabla h)\mu}$ and $(\|\nabla c\|^2\operatorname{lip}(\nabla h)+\mu)$ are in general comparable, we deduce that the worst-case efficiencies of the prox-linear and prox-gradient schemes are comparable as well. On the other hand, in the case $\|\nabla c\|^2\operatorname{lip}(\nabla h) \gg \mu$, e.g. when ∇h is poorly conditioned, the prox-linear method can be orders of magnitude more efficient than the prox-gradient method. Such circumstances are common. For example, the Huber penalty h in Example 3.4 has a small Lipschitz constant, while the Lipschitz constant of the gradient $\operatorname{lip}(\nabla h)$ can be huge.

7.3 When h is nonsmooth: choosing an optimal smoothing parameter

When h is nonsmooth, it is appealing to replace h by a smooth approximation and then minimize the resulting composite function by the basic prox-gradient method. It is already clear from the previous section that this is a poor strategy, since the Lipschitz constant of the gradient will be huge. Instead, one should use an inexact prox-linear method on the smoothed problem. Nonetheless, let us explain the corresponding efficiency estimate, before moving on to the best strategy. In particular, we will see how to choose an "optimal" smoothing parameter.

For the sake of simplicity and to ease the notation, we consider here only replacing h by its Moreau envelope h_ν. Analysis for different types of smoothing, such as those in [36], or more generally those in [4], are completely analogues. Define the smoothed composite function

$$F^\nu(x) := g(x) + h_\nu(c(x)).$$

We next record elementary properties of the smoothed function stemming from Lemma 2.1.

Corollary 7.4 (Properties of the smoothing)**.** *The following are true.*

1. *The inequality*
$$F^\nu(y) \geq F^\nu(x) + \langle w, y - x\rangle - \frac{\mu}{2}\|y - x\|^2$$
holds for all $x, y \in \mathbf{R}^n$ and $w \in \partial F^\nu(x)$.

2. *The function $h_\nu \circ c$ is C^1-smooth and the gradient $\nabla(h_\nu \circ c)$ is Lipschitz continuous on* dom g *with constant $\mu + \|\nabla c\|^2/\nu$.*

Proof. For any points $x, y \in \mathbf{R}^n$ observe

$$\begin{aligned} h_\nu(c(y)) &\geq h_\nu(c(x)) + \langle \nabla h_\nu(c(x)), c(y) - c(x)\rangle \\ &\geq h_\nu(c(x)) + \langle \nabla h_\nu(c(x)), \nabla c(x)(y - x)\rangle - \frac{\mu}{2}\|y - x\|^2, \end{aligned}$$

where we used the inequality $\|\nabla h_\nu(c(x))\| \leq L$ (see Lemma 2.1). Taking into account the inequality $g(y) \geq g(x) + \langle \partial g(x), y - x\rangle$, the first claim follows. The estimate on the Lipschitz constant of $\nabla(h_\nu \circ c)$ follows by the same reasoning as in Lemma 7.3, while taking into account Lemma 2.1. □

Next, to choose appropriately the smoothing parameter $\nu > 0$, we must understand the relationship between nearly stationary points of F^ν and those of F. The following result is in the same spirit as Theorem 7.2. More general perturbation results of this type appear in [17].

Lemma 7.5 (Stationarity of the smoothing)**.** *For any point $x \in$ dom F and $\rho > 0$, there exists a point $\hat{x}$ satisfying*

$$\begin{cases} \|\hat{x} - x\| & \leq \sqrt{\frac{L^2\rho\nu}{2}}, \\ \operatorname{dist}(0; \partial F(\hat{x})) & \leq \sqrt{\frac{L^2\nu}{2\rho}} + \sqrt{\frac{L^4\beta^2\rho\nu}{2}} + \operatorname{dist}(0; \partial F^\nu(x). \end{cases}$$

Proof. If the set $\partial F^\nu(x)$ is empty, simply set $\hat{x} := x$. Else, fix a vector $w \in \partial F^\nu(x)$ of minimal norm. Lemma 2.1 and Corollary 7.4 imply

$$F(y) \geq F^\nu(y) \geq F^\nu(x) + \langle w, y - x\rangle - \frac{\mu}{2}\|y - x\|^2$$

for any points $x, y \in \mathbf{R}^n$. Define the function

$$\zeta(y) := F(y) - \langle w, y - x\rangle + \frac{\mu}{2}\|y - x\|^2.$$

Thus we deduce

$$\zeta(x) - \inf \zeta \leq F(x) - F^\nu(x) \leq \frac{L^2\nu}{2}.$$

Setting $\varepsilon := \frac{L^2\nu}{2}$ and applying Theorem 7.1, we obtain for any $\rho > 0$ the inequality

$$\|\rho^{-1}(x - \operatorname{prox}_{\rho\zeta}(x))\| \leq \sqrt{\frac{L^2\nu}{2\rho}}.$$

Thus the point $\hat{x} := \operatorname{prox}_{\rho\zeta}(x)$ satisfies

$$\|x - \hat{x}\| \leq \sqrt{\frac{L^2\rho\nu}{2}}$$

and

$$\begin{aligned}\operatorname{dist}(0; \partial F(\hat{x})) &\leq \sqrt{\frac{L^2\nu}{2\rho}} + \operatorname{dist}(0; \partial F^\nu(x)) + \mu\|\hat{x} - x\| \\ &\leq \sqrt{\frac{L^2\nu}{2\rho}} + \mu\sqrt{\frac{L^2\rho\nu}{2}} + \operatorname{dist}(0; \partial F^\nu(x))\end{aligned}$$

as claimed. □

Seeking to satisfy (7.5) while keeping ν as large as possible, let us set

$$\sqrt{\frac{L^2\rho\nu}{2}} = \frac{\varepsilon}{3\mu}. \tag{7.6}$$

Then we deduce the bound

$$\operatorname{dist}(0; \partial F(\hat{x})) \leq \frac{\varepsilon}{3\mu\rho} + \frac{\varepsilon}{3} + \operatorname{dist}(0; \partial F^\nu(x)).$$

Hence if we set $\rho := 1/\mu$, any point x with $\operatorname{dist}(0; \partial F^\nu(x)) \leq \varepsilon/3$ would satisfy the desired condition (7.5). Plugging this value of ρ into (7.6), we deduce the value of the smoothing parameter

$$\nu = \frac{2\varepsilon^2}{9L^2\mu}.$$

Thus the cost of obtaining such a point x by running a prox-gradient method on F^ν is

$$\mathcal{O}\left(\frac{\operatorname{lip}(\nabla(h_\nu \circ c)) \cdot (F^\nu(x_0) - \inf F^\nu))}{\varepsilon^2}\right) \leq \mathcal{O}\left(\frac{(L^2\mu\|\nabla c\|^2 + \varepsilon^2\mu)(F(x_0) - F^* + \varepsilon^2/\mu)}{\varepsilon^4}\right).$$

Assuming the term $\varepsilon^2\mu$ is negligible, the estimate simplifies to

$$\mathcal{O}\left(\frac{\|\nabla c\|^2 L^2\mu(F(x_0) - F^*)}{\varepsilon^4}\right). \tag{7.7}$$

In contrast, the analogous estimate in Theorem 6.3 for the prox-linear method (with $q = 1$) is

$$\widetilde{\mathcal{O}}\left(\sqrt{\|\nabla c\|^2\beta}\cdot\left(\frac{L^2\mu(F(x_0)-F^*+8)^2}{\varepsilon^4}\right)\right),$$

Comparing with (7.7), we see that the prox-linear method is superior when $\sqrt{\beta}(F(x_0)-F^*) \ll \|\nabla c\|$. In any case, we will see that combining the smoothing technique with the prox-linear method yields a much better algorithm.

7.4 Combining smoothing with the prox-linear method

Let us consider solving the smoothed problem

$$\min_x \; F^\nu(x) = g(x) + h_\nu(c(x))$$

by the inexact prox-linear method. Namely, set

$$x^+ := \operatorname*{argmin}_z \left\{ g(z) + h_\nu(c(x) + \nabla c(x)(z-x)) + \frac{\mu}{2}\|z-x\|^2 \right\}$$

and

$$\mathcal{G}^\nu_{1/\mu}(x) = \mu(x - x^+).$$

Let us see how to set ν by understanding what a small value of $\|\mathcal{G}^\nu_{1/\mu}(x)\|$ entails about the quality of x. By (7.3), there exists a point z satisfying

$$\begin{cases} \|z-x\| & \leq \frac{2}{\mu}\|\mathcal{G}^\nu_{1/\mu}(x)\|, \\ \operatorname{dist}(0;\partial F^\nu(z)) & \leq 4\|\mathcal{G}^\nu_{1/\mu}(x)\|. \end{cases} \tag{7.8}$$

Now applying Lemma 7.5 with z in place of x and F^ν in place of F, we deduce that for any $\rho > 0$ there exists a point $\hat{x}$ satisfying

$$\begin{cases} \|\hat{x}-z\| & \leq \sqrt{\frac{L^2\rho\nu}{2}}, \\ \operatorname{dist}(0;\partial F(\hat{x})) & \leq \sqrt{\frac{L^2\nu}{2\rho}} + \sqrt{\frac{L^4\beta^2\rho\nu}{2}} + \operatorname{dist}(0;\partial F^\nu(z). \end{cases} \tag{7.9}$$

Combining inequalities (7.8) and (7.9) we deduce

$$\begin{cases} \|\hat{x}-x\| & \leq \sqrt{\frac{L^2\rho\nu}{2}} + \frac{2}{\mu}\|\mathcal{G}^\nu_{1/\mu}(x)\|, \\ \operatorname{dist}(0;\partial F(\hat{x})) & \leq \sqrt{\frac{L^2\nu}{2\rho}} + \mu\sqrt{\frac{L^2\rho\nu}{2}} + 4\|\mathcal{G}^\nu_{1/\mu}(x)\|. \end{cases}$$

Plugging in $\rho := \frac{1}{\mu}$ and $\nu := \frac{2\varepsilon^2}{9L^2\mu}$, we deduce

$$\|\hat{x}-x\| \leq \frac{\varepsilon}{3\mu} + \frac{2}{\mu}\|\mathcal{G}^\nu_{1/\mu}(x)\| \qquad \text{and} \qquad \operatorname{dist}(0;\partial F(\hat{x})) \leq \frac{2\varepsilon}{3} + 4\|\mathcal{G}^\nu_{1/\mu}(x)\|$$

Thus if the inequality $\|\mathcal{G}^\nu_{1/\mu}(x)\| \leq \frac{\varepsilon}{12}$ were to hold, then the point x would satisfy the desired termination condition (7.5). Taking into account $\operatorname{lip}(\nabla h_\nu) = \frac{1}{\nu} = \frac{9L^2\mu}{2\varepsilon^2}$ and appealing to Theorem 6.1 with h_ν replacing h, we deduce that the cost of obtaining such a point x is

$$\widetilde{\mathcal{O}}\left(\frac{L\mu\|\nabla c\|\cdot\left(F(x_0)-F^*+\frac{\varepsilon^2}{9\mu}\right)}{\varepsilon^3}\right) \qquad \text{basic operations,}$$

where $\widetilde{\mathcal{O}}$ hides universal constants and logarithmic terms in $1/\varepsilon, \|\nabla c\|, \mu$, and $F(x_0)-F^*$. To the best of our knowledge, this is the best-known efficiency estimate of any first-order method for the composite problem class (3.1).

References

[1] A. Aravkin, J.V. Burke, L. Ljung, A. Lozano, and G. Pilonetto. Generalized Kalman smoothing:modeling and algorithm. *Preprint arXiv:1609.06369*, 2016.

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

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

[4] A. Beck and M. Teboulle. Smoothing and first order methods: a unified framework. *SIAM J. Optim.*, 22(2):557–580, 2012.

[5] J.M. Borwein and Q.J. Zhu. *Techniques of Variational Analysis*. Springer Verlag, New York, 2005.

[6] J.V. Burke. Descent methods for composite nondifferentiable optimization problems. *Math. Programming*, 33(3):260–279, 1985.

[7] J.V. Burke. An exact penalization viewpoint of constrained optimization. *SIAM J. Control Optim.*, 29(4):968–998, 1991.

[8] J.V. Burke, F.E. Curtis, H. Wang, and J. Wang. Iterative reweighted linear least squares for exact penalty subproblems on product sets. *SIAM J. Optim.*, 25(1):261–294, 2015.

[9] J.V. Burke and M.C. Ferris. A Gauss-Newton method for convex composite optimization. *Math. Programming*, 71(2, Ser. A):179–194, 1995.

[10] R.H. Byrd, J Nocedal, and R.A. Waltz. KNITRO: An integrated package for nonlinear optimization. In *Large-scale nonlinear optimization*, volume 83 of *Nonconvex Optim. Appl.*, pages 35–59. Springer, New York, 2006.

[11] C. Cartis, N.I.M. Gould, and P.L. Toint. On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming. *SIAM J. Optim.*, 21(4):1721–1739, 2011.

[12] D.I. Clark. The mathematical structure of Huber's M-estimator. *SIAM journal on scientific and statistical computing*, 6(1):209–219, 1985.

[13] T.F. Coleman and A.R. Conn. Nonlinear programming via an exact penalty function: global analysis. *Math. Programming*, 24(2):137–161, 1982.

[14] A.R. Conn, K. Scheinberg, and L.N. Vicente. *Introduction to derivative-free optimization*, volume 8 of *MPS/SIAM Series on Optimization*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Programming Society (MPS), Philadelphia, PA, 2009.

[15] A. Daniilidis and J. Malick. Filling the gap between lower-C^1 and lower-C^2 functions. *J. Convex Anal.*, 12(2):315–329, 2005.

[16] G. Di Pillo and L. Grippo. Exact penalty functions in constrained optimization. *SIAM J. Control Optim.*, 27(6):1333–1360, 1989.

[17] D. Drusvyatskiy, A.D. Ioffe, and A.S. Lewis. Nonsmooth optimization using taylor-like models: error bounds, convergence, and termination criteria. *Preprint arXiv:1610.03446*, 2016.

[18] D. Drusvyatskiy and A.S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Preprint arXiv:1602.06661*, 2016.

[19] R. Dutter and P.J. Huber. Numerical methods for the nonlinear robust regression problem. *J. Statist. Comput. Simulation*, 13(2):79–113, 1981.

[20] I.I. Eremin. The penalty method in convex programming. *Cybernetics*, 3(4):53–56 (1971), 1967.

[21] R. Fletcher. A model algorithm for composite nondifferentiable optimization problems. *Math. Programming Stud.*, (17):67–76, 1982. Nondifferential and variational techniques in optimization (Lexington, Ky., 1980).

[22] S. Ghadimi and G. Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Math. Program.*, 156(1-2, Ser. A):59–99, 2016.

[23] J.-B. Hiriart-Urruty. ε-subdifferential calculus. In *Convex analysis and optimization (London, 1980)*, volume 57 of *Res. Notes in Math.*, pages 43–92. Pitman, Boston, Mass.-London, 1982.

[24] P. J. Huber. *Robust Statistics*. John Wiley and Sons, 2 edition, 2004.

[25] K. Levenberg. A method for the solution of certain non-linear problems in least squares. *Quart. Appl. Math.*, 2:164–168, 1944.

[26] A.S. Lewis and S.J. Wright. A proximal method for composite minimization. *Math. Program.*, pages 1–46, 2015.

[27] W. Li and J. Swetits. The linear l1 estimator and the huber m-estimator. *SIAM Journal on Optimization*, 8(2):457–475, 1998.

[28] D.W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *J. Soc. Indust. Appl. Math.*, 11:431–441, 1963.

[29] B.S. Mordukhovich. *Variational analysis and generalized differentiation. I*, volume 330 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2006. Basic theory.

[30] J.J. Moré. The Levenberg-Marquardt algorithm: implementation and theory. In *Numerical analysis (Proc. 7th Biennial Conf., Univ. Dundee, Dundee, 1977)*, pages 105–116. Lecture Notes in Math., Vol. 630. Springer, Berlin, 1978.

[31] S.C. Narula and J.F. Wellington. The minimum sum of absolute errors regression: a state of the art survey. *Internat. Statist. Rev.*, 50(3):317–326, 1982.

[32] A.S. Nemirovsky and D.B. Yudin. *Problem complexity and method efficiency in optimization*. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1983. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics.

[33] Y. Nesterov. *Introductory lectures on convex optimization*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004. A basic course.

[34] Y. Nesterov. How to make the gradients small. *OPTIMA, MPS Newsletter*, (88):10–11, 2012.

[35] Yu. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk SSSR*, 269(3):543–547, 1983.

[36] Yu. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1, Ser. A):127–152, 2005.

[37] Yu. Nesterov. Gradient methods for minimizing composite functions. *Math. Program.*, 140(1, Ser. B):125–161, 2013.

[38] J. Nocedal and S.J. Wright. *Numerical optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition, 2006.

[39] E. Pauwels. The value function approach to convergence analysis in composite optimization. *Oper. Res. Lett.*, 44(6):790–795, 2016.

[40] M.J.D. Powell. General algorithms for discrete nonlinear approximation calculations. In *Approximation theory, IV (College Station, Tex., 1983)*, pages 187–218. Academic Press, New York, 1983.

[41] M.J.D. Powell. On the global convergence of trust region algorithms for unconstrained minimization. *Math. Programming*, 29(3):297–303, 1984.

[42] R.T. Rockafellar and R.J-B. Wets. *Variational Analysis*. Grundlehren der mathematischen Wissenschaften, Vol 317, Springer, Berlin, 1998.

[43] M. Schmidt, Nicolas L.R., and Francis R.B. Convergence rates of inexact proximal-gradient methods for convex optimization. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1458–1466. Curran Associates, Inc., 2011.

[44] E. Siemsen and K.A. Bollen. Least absolute deviation estimation in structural equation modeling. *Sociol. Methods Res.*, 36(2):227–265, 2007.

[45] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. Technical report, University of Washington, 2008.

[46] S.M. Wild. *Solving Derivative-Free Nonlinear Least Squares Problems with POUNDERS*. 2014. Argonne National Lab.

[47] S.J. Wright. Convergence of an inexact algorithm for composite nonsmooth optimization. *IMA J. Numer. Anal.*, 10(3):299–321, 1990.

[48] Y. Yuan. On the superlinear convergence of a trust region algorithm for nonsmooth optimization. *Math. Programming*, 31(3):269–285, 1985.

A Proofs of Theorems 5.2, 5.3, 5.5, and 5.6

In this section, we prove Theorems 5.2, 5.3, 5.5, and 5.6 in order.

Proof of Theorem 5.2. Observe the inequalities:

$$\begin{aligned} F(x_{k+1}) &\leq F_t(x_{k+1}; x_k) \\ &\leq h\big(\xi_{k+1} + c(x_k) + \nabla c(x_k)(x_{k+1} - x_k)\big) + g(x_{k+1}) + \tfrac{1}{2t}\|x_{k+1} - x_k\|^2 + L \cdot \varepsilon_{k+1}. \end{aligned}$$

Since the point x_{k+1} minimizes the $\frac{1}{t}$-strongly convex function in (5.3), plugging in $z = x_k$ we deduce

$$\begin{aligned} F(x_{k+1}) &\leq h\big(\xi_{k+1} + c(x_k)\big) + g(x_k) + L \cdot \varepsilon_{k+1} - \tfrac{1}{2t}\|x_{k+1} - x_k\|^2 \\ &\leq F(x_k) + 2L \cdot \varepsilon_{k+1} - \tfrac{1}{2t}\|x_{k+1} - x_k\|^2. \end{aligned}$$

Summing along the indices $j = 0, \dots, N-1$ yields

$$\sum_{j=0}^{N-1} \|t^{-1}(x_{j+1} - x_j)\|^2 \leq \frac{2}{t}\left(F(x_0) - F^* + 2L\sum_{j=0}^{N-1}\varepsilon_{j+1}\right).$$

Taking into account Lemma 5.1, we deduce

$$\min_{j=0,1,\dots,N-1} \|\mathcal{G}_t(x_j)\|^2 \leq \frac{1}{N}\sum_{i=0}^{N-1}\|\mathcal{G}_t(x_j)\|^2 \leq \frac{4t^{-1}(F(x_0) - F^* + 4L\sum_{j=1}^{N}\varepsilon_j)}{N},$$

as claimed. □

Proof of Theorem 5.3. The proof is a modification of the proof Theorem 4.8; as such, we skip some details. For any point w, we successively deduce

$$\begin{aligned} F(x_k) &\leq h\big(\zeta_k + c(y_k) + \nabla c(y_k)(x_k - y_k)\big) + g(x_k) + \frac{\mu}{2}\|x_k - y_k\|^2 + L \cdot \varepsilon_k \\ &\leq \Big(h\big(\zeta_k + c(y_k) + \nabla c(y_k)(x_k - y_k)\big) + g(x_k) + \frac{\tilde{\mu}}{2}\|x_k - y_k\|^2\Big) - \frac{\tilde{\mu} - \mu}{2}\|x_k - y_k\|^2 + L \cdot \varepsilon_k \\ &\leq h\big(\zeta_k + c(y_k) + \nabla c(y_k)(w - y_k)\big) + g(w) \\ &\quad + \frac{\tilde{\mu}}{2}\Big(\|w - y_k\|^2 - \|w - x_k\|^2\Big) - \frac{\tilde{\mu} - \mu}{2}\|x_k - y_k\|^2 + L \cdot \varepsilon_k \\ &\leq h\big(c(y_k) + \nabla c(y_k)(w - y_k)\big) + g(w) \\ &\quad + \frac{\tilde{\mu}}{2}\Big(\|w - y_k\|^2 - \|w - x_k\|^2\Big) - \frac{\tilde{\mu} - \mu}{2}\|x_k - y_k\|^2 + 2L \cdot \varepsilon_k. \end{aligned}$$

Setting $w := a_k v_k + (1 - a_k)x_{k-1}$ and noting the equality $w - y_k = a_k(v_k - v_{k-1})$ then yields

$$\begin{aligned} F(x_k) \leq & h(c(y_k) + a_k \nabla c(y_k)(v_k - v_{k-1})) + a_k g(v_k) + (1 - a_k)g(x_{k-1}) \\ & + \frac{\tilde{\mu}}{2}\Big(\|a_k(v_k - v_{k-1})\|^2 - \|w - x_k\|^2\Big) - \frac{\tilde{\mu} - \mu}{2}\|x_k - y_k\|^2 + 2L \cdot \varepsilon_k. \end{aligned}$$

Upper bounding $-\|w-x_k\|^2$ by zero and using Lipschitz continuity of h we obtain for any point x the inequalities

$$\begin{aligned}
F(x_k) &\leq a_k\Big(\frac{1}{a_k}h(\xi_k + c(y_k) + a_k\nabla c(y_k)(v_k - v_{k-1})) + g(v_k)\Big) + (1-a_k)g(x_{k-1}) \\
&+ \frac{\tilde{\mu}a_k^2}{2}\|v_k - v_{k-1}\|^2 - \frac{\tilde{\mu}-\mu}{2}\|x_k - y_k\|^2 + L\cdot\delta_k + 2L\cdot\varepsilon_k. \\
&\leq a_k\Big(\frac{1}{a_k}h(\xi_k + c(y_k) + a_k\nabla c(y_k)(x - v_{k-1})) + g(x) + \frac{\tilde{\mu}a_k}{2}(\|x - v_{k-1}\|^2 - \|v_k - v_{k-1}\|^2 \\
&- \|v_k - x\|^2)\Big) + (1-a_k)g(x_{k-1}) + \frac{\tilde{\mu}a_k^2}{2}\|v_k - v_{k-1}\|^2 - \frac{\tilde{\mu}-\mu}{2}\|x_k - y_k\|^2 + L\delta_k + 2L\varepsilon_k. \\
&\leq h(c(y_k) + a_k\nabla c(y_k)(x - v_{k-1})) + a_k g(x) + \frac{\tilde{\mu}a_k^2}{2}(\|x - v_{k-1}\|^2 - \|v_k - x\|^2) \\
&+ (1-a_k)g(x_{k-1}) - \frac{\tilde{\mu}-\mu}{2}\|x_k - y_k\|^2 + 2L\delta_k + 2L\varepsilon_k.
\end{aligned}$$

Define $\hat{x} := a_k x + (1-a_k)x_{k-1}$ and note $a_k(x - v_{k-1}) = \hat{x} - y_k$. The same argument as that of (4.11) yields

$$\begin{aligned}
h(c(y_k) + \nabla c(y_k)(\hat{x} - y_k)) \leq& a_k h(c(x)) + (1-a_k)h(c(x_{k-1})) + \\
&\rho a_k(1-a_k)\|x - x_{k-1}\|^2 + \frac{ra_k^2}{2}\|x - v_{k-1}\|^2.
\end{aligned}$$

Hence upper bounding $1 - a_k \leq 1$ we deduce

$$\begin{aligned}
F(x_k) \leq& a_k F(x) + (1-a_k)F(x_{k-1}) + \frac{\tilde{\mu}a_k^2}{2}(\|x - v_{k-1}\|^2 - \|x - v_k\|^2) \\
&- \frac{\tilde{\mu}-\mu}{2}\|y_k - x_k\|^2 + \rho a_k\|x - x_{k-1}\|^2 + \frac{ra_k^2}{2}\|x - v_{k-1}\|^2 + 2L(\delta_k + \varepsilon_k).
\end{aligned}$$

This expression is identical to that of (4.8) except for the error term $2L(\delta_k + \varepsilon_k)$. The same argument as in the proof of Theorem 4.8 then shows

$$\begin{aligned}
\frac{F(x_N) - F(x^*)}{a_N^2} + \frac{\tilde{\mu}}{2}\|x^* - v_N\|^2 \leq& \frac{\tilde{\mu}}{2}\|x^* - v_0\|^2 + \rho M^2\left(\sum_{j=1}^N \frac{1}{a_j}\right) \\
&+ \frac{NrM^2}{2} - \frac{\tilde{\mu}-\mu}{2}\sum_{j=1}^N \frac{\|x_j - y_j\|^2}{a_j^2} + 2L\sum_{j=1}^N \frac{\varepsilon_j + \delta_j}{a_j^2}.
\end{aligned}$$

Hence appealing to Lemma 5.1, we deduce

$$\begin{aligned}
\sum_{j=1}^N \frac{\|\mathcal{G}_{1/\tilde{\mu}}(y_j)\|^2}{a_j^2} &\leq 8L\tilde{\mu}\sum_{j=1}^N \frac{\varepsilon_j}{a_j^2} + 2\sum_{j=1}^N \frac{\|\tilde{\mu}(x_j - y_j)\|^2}{a_j^2} \\
&\leq 8L\tilde{\mu}\sum_{j=1}^N \frac{\varepsilon_j}{a_j^2} + \frac{4\tilde{\mu}^2}{\tilde{\mu}-\mu}\left(\frac{\tilde{\mu}}{2}\|x^* - v_0\|^2 + \frac{NM^2(r + \frac{\rho}{2}(N+3))}{2} + 2L\sum_{j=1}^N \frac{\varepsilon_j + \delta_j}{a_j^2}\right).
\end{aligned}$$

Therefore

$$\min_{i=1,\dots,N}\|\mathcal{G}_{1/\tilde{\mu}}(y_j)\|^2 \leq \frac{8\cdot 24L\tilde{\mu}\sum_{j=1}^N \frac{\varepsilon_j}{a_j^2}}{N(N+1)(2N+1)} + \frac{48\tilde{\mu}^2}{\tilde{\mu}-\mu}\left(\frac{\|x^*-v_0\|^2}{N(N+1)(2N+1)} + \frac{M^2(r+\frac{\rho}{2}(N+3))}{(N+1)(2N+1)} + \frac{4L\sum_{j=1}^N \frac{\varepsilon_j+\delta_j}{a_j^2}}{N(N+1)(2N+1)}\right)$$

Combining the first and fourth terms and using the inequality $\tilde{\mu}\geq\mu$ yields the claimed efficiency estimate on $\|\mathcal{G}_{1/\tilde{\mu}}(y_j)\|^2$. Finally, the claimed efficiency estimate on the functional error $F(x_N)-F^*$ in the setting $r=0$ follows by the same reasoning as in Theorem 4.8. □

Proof of Theorem 5.5. Let x_{k+1}^* be the exact minimizer of $F_t(\cdot;x_k)$. Note then the equality $\mathcal{G}_t(x_k)=t^{-1}(x_{k+1}^*-x_k)$. Taking into account that $F_t(\cdot;x_k)$ is strongly convex with modulus $1/t$, we deduce

$$F(x_k)=F_t(x_k;x_k)\geq F_t(x_{k+1}^*;x_k)+\tfrac{t}{2}\|\mathcal{G}_t(x_k)\|^2\geq F_t(x_{k+1};x_k)-\varepsilon_{k+1}+\tfrac{t}{2}\|\mathcal{G}_t(x_k)\|^2.$$

Then the inequality $t\leq\mu^{-1}$ along with (4.2) implies that $F_t(\cdot;x_k)$ is an upper model of $F(\cdot)$ and therefore

$$F(x_k)\geq F(x_{k+1})-\varepsilon_{k+1}+\tfrac{t}{2}\|\mathcal{G}_t(x_k)\|^2.$$

We conclude

$$\min_{j=0,\dots,N-1}\|\mathcal{G}_t(x_j)\|^2\leq\frac{1}{N}\sum_{j=0}^{N-1}\|G_t(x_j)\|^2\leq\frac{2t^{-1}\left(\sum_{j=0}^{N-1}F(x_j)-F(x_{j+1})+\sum_{j=0}^{N-1}\varepsilon_{j+1}\right)}{N}$$
$$\leq\frac{2t^{-1}\big(F(x_0)-F^*+\sum_{j=0}^{N-1}\varepsilon_{j+1}\big)}{N}.$$

The proof is complete. □

We next prove Theorem 5.6. To this end, we will need the following lemma.

Lemma A.1 (Lemma 1 in [43])**.** *Suppose the following recurrence relation is satisfied*

$$d_k^2\leq d_0^2+c_k+\sum_{i=1}^k\beta_i d_i$$

for some sequences $d_i,\beta_i\geq 0$ and an increasing sequence $c_i\geq 0$. Then the inequality holds:

$$d_k\leq A_k:=\frac{1}{2}\sum_{i=1}^k\beta_i+\left(d_0^2+c_k+\left(\frac{1}{2}\sum_{i=1}^k\beta_i\right)^2\right)^{1/2}.$$

Moreover since the terms on the right-hand side increase in k, we also conclude for any $k\leq N$ the inequality $d_k\leq A_N$.

The *ε-subdifferential* of a function $f\colon \mathbf{R}^n \to \overline{\mathbf{R}}$ at a point $\bar{x}$ is the set

$$\partial_\varepsilon f(\bar{x}) := \{v \in \mathbf{R}^n : f(x) - f(\bar{x}) \geq \langle v, x - \bar{x}\rangle - \varepsilon \quad \text{for all } x \in \mathbf{R}^n\}.$$

In particular, notice that $\bar{x}$ is an ε-approximate minimizer of f if and only if the inclusion $0 \in \partial_\varepsilon f(\bar{x})$ holds. For the purpose of analysis, it is useful to decompose the function $F_{t,\alpha}(z, y, v)$ into a sum

$$F_{t,\alpha}(z; y, v) = F_\alpha(z; y, v) + \frac{1}{2t}\|z - v\|^2$$

The sum rule for ε-subdifferentials [23, Theorem 2.1] guarantees

$$\partial_\varepsilon F_{t,\alpha}(\cdot; y, v) \subseteq \partial_\varepsilon F_\alpha(\cdot; y, v) + \partial_\varepsilon \left(\frac{1}{2t}\|\cdot - v\|^2\right).$$

Lemma A.2. *The ε-subdifferential $\partial_\varepsilon \left(\frac{1}{2t}\|\cdot - v\|^2\right)$ at a point $\bar{z}$ is the set*

$$\left\{t^{-1}(z - v + \gamma) \,:\, \frac{1}{2t}\,\|\gamma\|^2 \leq \varepsilon\right\}.$$

Proof. This follows by completing the square in the definition of the ε-subdifferential. □

In particular, suppose that z^+ is an ε-approximate minimizer of $F_{t,\alpha}(\cdot; y, v)$. Then Lemma A.2 shows that there is a vector γ satisfying $\|\gamma\|^2 \leq 2t\varepsilon$ and

$$t^{-1}(v - z^+ - \gamma) \in \partial_\varepsilon F_\alpha(z^+; y, v). \tag{A.1}$$

We are now ready to prove Theorem 5.6.

Proof of Theorem 5.6. Let x_k, y_k, and v_k be the iterates generated by Algorithm 7. We imitate the proof of Theorem 4.8, while taking into account inexactness. First, inequality (4.9) is still valid:

$$F(x_k) \leq F(x_k; y_k) + \tfrac{\mu}{2}\,\|x_k - y_k\|^2.$$

Since x_k is an ε_k-approximate minimizer of the function $F(\cdot; y_k) = F_{1/\tilde{\mu},1}(\cdot; y_k, y_k)$, from (A.1), we obtain a vector γ_k satisfying $\|\gamma_k\|^2 \leq 2\varepsilon_k\tilde{\mu}^{-1}$ and $\tilde{\mu}(y_k - x_k - \gamma_k) \in \partial_{\varepsilon_k} F(x_k; y_k)$. Consequently for all points w we deduce the inequality

$$F(x_k) \leq F(w; y_k) + \tfrac{\mu}{2}\,\|x_k - y_k\|^2 + \langle\tilde{\mu}(y_k - x_k - \gamma_k), x_k - w\rangle + \varepsilon_k. \tag{A.2}$$

Set $w_k := a_k v_k + (1 - a_k)x_{k-1}$ and define $c_k := x_k - w_k$. Taking into account $w_k - y_k = a_k(v_k - v_{k-1})$, the previous inequality with $w = w_k$ becomes

$$\begin{aligned} F(x_k) \leq\; & h(c(y_k) + a_k\nabla c(y_k)(v_k - v_{k-1})) + a_k g(v_k) + (1 - a_k)g(x_{k-1}) + \tfrac{\mu}{2}\,\|x_k - y_k\|^2 \\ & + \tilde{\mu}\langle y_k - x_k, c_k\rangle - \tilde{\mu}\langle\gamma_k, c_k\rangle + \varepsilon_k. \end{aligned} \tag{A.3}$$

By completing the square, one can check

$$\tilde{\mu}\langle y_k - x_k, c_k\rangle = \tfrac{\tilde{\mu}}{2}\big(\,\|a_k v_k - a_k v_{k-1}\|^2 - \|x_k - y_k\|^2 - \|c_k\|^2\,\big).$$

Observe in addition

$$-\tilde{\mu}\langle\gamma_k, c_k\rangle - \tfrac{\tilde{\mu}}{2}\|c_k\|^2 = -\tfrac{\tilde{\mu}}{2}\|\gamma_k + c_k\|^2 + \tfrac{\tilde{\mu}}{2}\|\gamma_k\|^2.$$

By combining the two equalities with (A.3) and dropping the term $\frac{\tilde{\mu}}{2}\|\gamma_k - c_k\|^2$, we deduce

$$\begin{aligned} F(x_k) \leq\, & h(c(y_k) + a_k\nabla c(y_k)(v_k - v_{k-1})) + a_k g(v_k) + (1-a_k)g(x_{k-1}) \\ & + \tfrac{\tilde{\mu}a_k^2}{2}\|v_k - v_{k-1}\|^2 - \tfrac{\tilde{\mu}-\mu}{2}\|x_k - y_k\|^2 + \varepsilon_k + \tfrac{\tilde{\mu}}{2}\|\gamma_k\|^2. \end{aligned} \tag{A.4}$$

Next recall that v_k is a δ_k-approximate minimizer of $F_{(\tilde{\mu}a_k)^{-1},a_k}(\cdot; y_k, v_{k-1})$. Using (A.1), we obtain a vector η_k satisfying $\|\eta_k\|^2 \leq \frac{2\delta_k}{a_k\tilde{\mu}}$ and $a_k\tilde{\mu}(v_{k-1} - v_k - \eta_k) \in \partial_{\delta_k} F_{a_k}(v_k; y_k, v_{k-1})$. Hence, we conclude for all the points x the inequality

$$\begin{aligned} F_{a_k}(v_k; y_k, v_{k-1}) \leq\, & \frac{1}{a_k}h(c(y_k) + a_k\nabla c(y_k)(x - v_{k-1}) + g(x) \\ & + \tilde{\mu}a_k\langle v_{k-1} - v_k - \eta_k, v_k - x\rangle + \delta_k. \end{aligned} \tag{A.5}$$

Completing the square, one can verify

$$\langle v_{k-1} - v_k, v_k - x\rangle = \frac{1}{2}(\|x - v_{k-1}\|^2 - \|x - v_k\|^2 - \|v_k - v_{k-1}\|^2).$$

Hence combining this with (A.4) and (A.5), while taking into account the inequalities $\|\gamma_k\|^2 \leq 2\varepsilon_k\tilde{\mu}^{-1}$ and $\|\eta_k\|^2 \leq \frac{2\delta_k}{a_k\tilde{\mu}}$, we deduce

$$\begin{aligned} F(x_k) \leq\, & h(c(y_k) + a_k\nabla c(y_k)(x - v_{k-1}) + a_k g(x) + (1-a_k)g(x_{k-1}) \\ & + \tfrac{\tilde{\mu}a_k^2}{2}(\|x - v_{k-1}\|^2 - \|x - v_k\|^2) + a_k\delta_k - \tfrac{\tilde{\mu}-\mu}{2}\|x_k - y_k\|^2 + 2\varepsilon_k \\ & + a_k^{3/2}\sqrt{2\tilde{\mu}\delta_k}\cdot\|v_k - x\|. \end{aligned}$$

Following an analogous part of the proof of Theorem 4.8, define now the point $\hat{x} = a_k x + (1-a_k)x_{k-1}$. Taking into account $a_k(x - v_{k-1}) = \hat{x} - y_k$, we conclude

$$\begin{aligned} h(c(y_k) + \nabla c(y_k)(\hat{x} - y_k)) &\leq (h\circ c)(\hat{x}) + \frac{r}{2}\|\hat{x} - y_k\|^2 \\ &\leq a_k h(c(x)) + (1-a_k)h(c(x_{k-1})) \\ &\quad + \rho a_k(1-a_k)\|x - x_{k-1}\|^2 + \frac{ra_k^2}{2}\|x - v_{k-1}\|^2. \end{aligned}$$

Thus we obtain

$$\begin{aligned} F(x_k) \leq\, & a_k F(x) + (1-a_k)F(x_{k-1}) + \rho a_k\|x - x_{k-1}\|^2 + \frac{ra_k^2}{2}\|x - v_{k-1}\|^2 \\ & + \tfrac{\tilde{\mu}a_k^2}{2}(\|x - v_{k-1}\|^2 - \|x - v_k\|^2) + a_k\delta_k - \tfrac{\tilde{\mu}-\mu}{2}\|x_k - y_k\|^2 + 2\varepsilon_k \\ & + a_k^{3/2}\sqrt{2\tilde{\mu}\delta_k}\cdot\|v_k - x\|. \end{aligned}$$

As in the proof of Theorem 4.8, setting $x = x^*$, we deduce

$$\frac{F(x_N)-F^*}{a_N^2}+\frac{\tilde{\mu}}{2}\|x^*-v_N\|^2 \leq \frac{\tilde{\mu}}{2}\|x^*-v_0\|^2+\rho M^2\sum_{i=1}^N\frac{1}{a_i}+\frac{NrM^2}{2}+\sum_{i=1}^N\frac{\delta_i}{a_i}$$
$$-\frac{\tilde{\mu}-\mu}{2}\sum_{i=1}^N\frac{\|x_i-y_i\|^2}{a_i^2}+2\sum_{i=1}^N\frac{\varepsilon_i}{a_i^2}+\sqrt{2\tilde{\mu}}\sum_{i=1}^N\|x^*-v_i\|\cdot\sqrt{\frac{\delta_i}{a_i}}.$$

In particular, we have

$$\begin{aligned}\frac{\tilde{\mu}-\mu}{2}\sum_{i=1}^N\frac{\|x_i-y_i\|^2}{a_i^2} \leq &\frac{\tilde{\mu}}{2}\|x^*-v_0\|^2+\frac{\rho M^2N(N+3)}{4}+\frac{NrM^2}{2}+\sum_{i=1}^N\frac{\delta_i}{a_i}\\ &+2\sum_{i=1}^N\frac{\varepsilon_i}{a_i^2}+\sqrt{2\tilde{\mu}}\sum_{i=1}^N\|x^*-v_i\|\cdot\sqrt{\frac{\delta_i}{a_i}}.\end{aligned}\tag{A.6}$$

and

$$\begin{aligned}\frac{\tilde{\mu}}{2}\|x^*-v_N\|^2 \leq &\frac{\tilde{\mu}}{2}\|x^*-v_0\|^2+\frac{\rho M^2N(N+3)}{4}+\frac{NrM^2}{2}+\sum_{i=1}^N\frac{\delta_i}{a_i}\\ &+2\sum_{i=1}^N\frac{\varepsilon_i}{a_i^2}+\sqrt{2\tilde{\mu}}\sum_{i=1}^N\|x^*-v_i\|\cdot\sqrt{\frac{\delta_i}{a_i}}.\end{aligned}$$

Appealing to Lemma A.1 with $d_k=\|x^*-v_k\|$, we conclude $\|x^*-v_N\|\leq A_N$ for the constant

$$\begin{aligned}A_N :=&\sqrt{\frac{2}{\tilde{\mu}}}\sum_{i=1}^N\sqrt{\frac{\delta_i}{a_i}}+\\ &+\left(\|x^*-v_0\|^2+\frac{M^2N(r+\frac{\rho}{2}(N+3))}{\tilde{\mu}}+\frac{2}{\tilde{\mu}}\sum_{i=1}^N\frac{\delta_i}{a_i}+\frac{4}{\tilde{\mu}}\sum_{i=1}^N\frac{\varepsilon_i}{a_i^2}+\frac{2}{\mu}\left(\sum_{i=1}^N\sqrt{\frac{\delta_i}{a_i}}\right)^2\right)^{1/2}.\end{aligned}$$

Finally, combining inequality (A.6) with Lemma 5.4 we deduce

$$\begin{aligned}\frac{\tilde{\mu}-\mu}{2}\sum_{i=1}^n\frac{\|\mathcal{G}_{1/\tilde{\mu}}(y_i)\|^2}{a_i^2} \leq 2\tilde{\mu}(\tilde{\mu}-\mu)\sum_{i=1}^N\frac{\varepsilon_i}{a_i^2}+2\tilde{\mu}^2\Big(&\frac{\tilde{\mu}}{2}\|x^*-v_0\|^2+\frac{\rho M^2N(N+3)}{4}+\frac{NrM^2}{2}+\\ &+\sum_{i=1}^N\frac{\delta_i}{a_i}+2\sum_{i=1}^N\frac{\varepsilon_i}{a_i^2}+A_N\sqrt{2\tilde{\mu}}\sum_{i=1}^N\sqrt{\frac{\delta_i}{a_i}}\Big).\end{aligned}$$

Hence

$$\begin{aligned}\min_{i=1,\ldots,N}\|\mathcal{G}_{1/\tilde{\mu}}(y_i)\|^2 \leq \frac{96\tilde{\mu}\sum_{i=1}^N\frac{\varepsilon_i}{a_i^2}}{N(N+1)(2N+1)}+\frac{96\tilde{\mu}^2}{\tilde{\mu}-\mu}\Big(&\frac{\tilde{\mu}\|x^*-v_0\|^2}{2N(N+1)(2N+1)}+\frac{M^2(r+\frac{\rho}{2}(N+3))}{2(N+1)(2N+1)}+\\ &+\frac{\sum_{i=1}^N(\frac{\delta_ia_i+2\varepsilon_i}{a_i^2})+A_N\sqrt{2\tilde{\mu}}\sum_{i=1}^N\sqrt{\frac{\delta_i}{a_i}}}{N(N+1)(2N+1)}\Big).\end{aligned}$$

Combining the first and the fourth terms, the result follows. The efficiency estimate on $F(x_N)-F^*$ in the setting $r=0$ follows by the same argument as in the proof of Theorem 4.8. □