Efficiency of minimizing compositions of convex functions and smooth maps *

D. Drusvyatskiy † C. Paquette ‡

Abstract

We consider global efficiency of algorithms for minimizing a sum of a convex function and a composition of a Lipschitz convex function with a smooth map. The basic algorithm we rely on is the 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 efficiency $\mathcal{O}(\varepsilon^{-2})$, akin to gradient descent for smooth minimization. We show that when the subproblems can only be solved by first-order methods, a simple combination of smoothing, the prox-linear method, and a fast-gradient scheme yields an algorithm with complexity $\widetilde{\mathcal{O}}(\varepsilon^{-3})$. The technique readily extends to minimizing an average of m composite functions, with complexity $\widetilde{\mathcal{O}}(m/\varepsilon^2 + \sqrt{m}/\varepsilon^3)$ in expectation. We round off the paper with an inertial prox-linear method that automatically accelerates in presence of convexity.

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

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 $g\colon \mathbf{R}^d \to \mathbf{R} \cup \{\infty\}$ and $h\colon \mathbf{R}^m \to \mathbf{R}$ are closed convex functions and $c\colon \mathbf{R}^d \to \mathbf{R}^m$ is a smooth map. Regularized nonlinear least squares [55, Section 10.3] and exact penalty formulations of nonlinear programs [55, Section 17.2] are classical examples, while notable contemporary instances include robust phase retrieval [24, 25] and matrix factorization problems such as NMF [31, 32]. The setting where c maps to the real line and h is the identity function,

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

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

*University of Washington, Department of Mathematics, Seattle, WA 98195; Research of Drusvyatskiy and Paquette was partially supported by the AFOSR award FA9550-15-1-0237 and NSF DMS 1651851.

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

‡E-mail: yumiko88@uw.edu;

The proximal gradient algorithm, investigated by Beck-Teboulle [4] and Nesterov [54, Section 3], is a popular first-order method for additive composite minimization. Much of the current paper will center around the *prox-linear method*, which is a direct extension of the prox-gradient algorithm to the entire problem class (1.1). In each iteration, the prox-linear method 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 here is that the strongly convex proximal subproblems (1.3) can be solved efficiently. This is indeed reasonable in some circumstances. For example, one may have available specialized methods for the proximal subproblems, or interior-point points methods may be available for moderate dimensions d and m, or it may be that case that computing an accurate estimate of $\nabla c(x)$ is already the bottleneck (see e.g. Example 3.5). The prox-linear method was recently investigated in [13, 23, 38, 53], though the ideas behind the algorithm and of its trust-region variants are much older [8, 13, 28, 58, 59, 70, 72]. The scheme (1.3) reduces to the popular prox-gradient algorithm for additive composite minimization, while for nonlinear least squares, the algorithm is closely related to the Gauss-Newton algorithm [55, Section 10].

Our work focuses on global efficiency estimates of numerical methods. Therefore, in line with standard assumptions in the literature, 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 [48, 52], it is convenient to measure the progress of the prox-linear method in terms of the scaled steps, called the *prox-gradients*:

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

A short argument shows that with the optimal choice $t = (L\beta)^{-1}$, the prox-linear algorithm will find a point x satisfying $\|\mathcal{G}_{\frac{1}{L\beta}}(x)\| \le \varepsilon$ after at most $\mathcal{O}(\frac{L\beta}{\varepsilon^2}(F(x_0) - \inf F))$ iterations; see e.g. [13,23]. We mention in passing that iterate convergence under the KŁ-inequality was recently shown in [5, 56], while local linear/quadratic rates under appropriate regularity conditions were proved in [11, 23, 53]. The contributions of our work are as follows.

1. **(Prox-gradient and the Moreau envelope)** The size of the prox-gradient $\|\mathcal{G}_t(x_k)\|$ plays a basic role in this work. In particular, all convergence rates are stated in terms of this quantity. Consequently, it is important to understand precisely what this quantity entails about the quality of the point x_k (or x_{k+1}). For additive composite problems (1.2), the situation is clear. Indeed, the proximal gradient method generates iterates satisfying $F'(x_{k+1}; u) \ge -2\|\mathcal{G}_{\frac{1}{\beta}}(x_k)\|$ for all unit vectors u, where $F'(x; u)$ is the directional derivative of F at x in direction u [54, Corollary 1]. Therefore, a small prox-gradient $\|\mathcal{G}_{\frac{1}{\beta}}(x_k)\|$ guarantees that x_{k+1} is nearly stationary for the problem, since the derivative of F at x_{k+1} in any unit direction is nearly nonnegative. For the general composite class (1.1), such a conclusion is decisively false: the prox-linear method will typically generate an iterate sequence along which F is differentiable with gradient norms $\|\nabla F(x_k)\|$ uniformly bounded away from zero, in spite of the norms $\|\mathcal{G}_{\frac{1}{L\beta}}(x_k)\|$ tending to zero.[1] Therefore, one must justify the focus on the norm $\|\mathcal{G}_{\frac{1}{L\beta}}(x_k)\|$ by other means. To this end, our first contribution is Theorem 4.5, where we prove that $\|\mathcal{G}_{\frac{1}{L\beta}}(x)\|$ is proportional to the norm of the true gradient of the Moreau envelope of F — a well studied

[1] See the beginning of Section 4 for a simple example of this type of behavior.

smooth approximation of F having identical stationary points. An immediate consequence is that even though x might not be nearly stationary for F, a small prox-gradient $\|\mathcal{G}_{\frac{1}{L\beta}}(x)\|$ guarantees that x is near some point $\hat{x}$ (the proximal point), which is nearly stationary for F. In this sense, a small prox-gradient $\|\mathcal{G}_{\frac{1}{L\beta}}(x)\|$ is informative about the quality of x. We note that an earlier version of this conclusion based on a more indirect argument, appeared in [23, Theorem 5.3], and was used to derive linear/quadratic rates of convergence for the prox-linear method under suitable regularity conditions.

2. **(Inexactness and 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 the prox-linear method 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, it is routine to derive overall efficiency estimates for the prox-linear method, where the proximal subproblems are themselves solved by first-order algorithms. Unfortunately, the efficiency estimates we can prove for such direct methods appear to either be unsatisfactory or the algorithms themselves appear not to be very practical (Appendix C). Instead, we present algorithms based on a smoothing technique.

3. **(Complexity of first-order methods through smoothing)** Smoothing is a common technique in nonsmooth optimization. The seminal paper of Nesterov [52], in particular, derives convergence guarantees for algorithms based on infimal convolution smoothing in structured convex optimization. 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.

   We advocate running an inexact prox-linear method on the smooth approximation, with the proximal subproblems approximately solved by fast-gradient methods. To state the resulting complexity bounds, let us suppose that there is a finite upper bound on the operator norms $\|\nabla c(x)\|_{\mathrm{op}}$ over all x in the domain of g, and denote it by $\|\nabla c\|$.[2] We prove that the outlined scheme requires at most

$$\widetilde{\mathcal{O}}\left(\frac{L^2\beta\|\nabla c\|}{\varepsilon^3}(F(x_0)-\inf F)\right) \tag{1.4}$$

   evaluations of $c(x)$, matrix vector products $\nabla c(x)v$, $\nabla c(x)^T w$, and proximal operations of g and h to find a point x satisfying $\|\mathcal{G}_{\frac{1}{L\beta}}(x)\| \le \varepsilon$. To the best of our knowledge, this is the best known complexity bound for the problem class (1.1) among first-order methods. Here, the symbol $\widetilde{\mathcal{O}}$ hides logarithmic terms.[3]

4. (**Complexity of finite-sum problems)** Common large-scale problems in machine learning and high dimensional statistics lead to minimizing an average of a large number of functions. Consequently, we consider the finite-sum extension of the composite problem class,

$$\min_x \; F(x) := \frac{1}{m}\sum_{i=1}^{m} h_i(c_i(x)) + g(x),$$

   where now each h_i is L-Lipschitz and each c_i is C^1-smooth with β-Lipschitz gradient. Clearly,

[2] It is sufficient for the inequality $\|\nabla c\| \ge \|\nabla c(x_k)\|_{\mathrm{op}}$ to hold just along the iterate sequence x_k generated by the method; in particular, $\|\nabla c\|$ does not need to be specified when initializing the algorithm.

[3] If a good estimate on the gap $F(x_0) - \inf F$ is known, the logarithmic terms can be eliminated by a different technique, described in Appendix C.

the finite-sum problem is itself an instance of (1.1) under the identification $h(z_i, \ldots, z_m) := \frac{1}{m}\sum_{i=1}^m h_i(z_i)$ and $c(x) := (c_1(x), \ldots, c_m(x))$. In this structured context, however, the complexity of an algorithm is best measured in terms of the number of individual evaluations $c_i(x)$ and $\nabla c_i(x)$, dot-product evaluations $\nabla c_i(x)^T v$, and proximal operations prox_{th_i} and prox_{tg} the algorithm needs to find a point x satisfying $\|\mathcal{G}_{\frac{1}{L\beta}}(x)\| \leq \varepsilon$. A routine computation shows that the efficiency estimate (1.4) of the basic inexact prox-linear method described above leads to the complexity

$$\widetilde{\mathcal{O}}\left(\frac{m \cdot L^2\beta\|\nabla c\|}{\varepsilon^3}\left(F(x_0) - \inf F\right)\right), \tag{1.5}$$

where abusing notation, we use $\|\nabla c\|$ to now denote an upper bound on $\|\nabla c_i(x)\|$ over all $i = 1, \ldots, m$ and $x \in \mathrm{dom}\, g$. We show that a better complexity in expectation is possible by incorporating (accelerated)-incremental methods [1, 29, 36, 40, 65] for the proximal subproblems. The resulting randomized algorithm will generate a point x satisfying

$$\mathbb{E}[\|\mathcal{G}_{\frac{1}{L\beta}}(x)\|] \leq \varepsilon,$$

after at most

$$\widetilde{\mathcal{O}}\left(\left(\frac{mL\beta}{\varepsilon^2} + \frac{\sqrt{m} \cdot L^2\beta\|\nabla c\|}{\varepsilon^3}\right) \cdot \left(F(x_0) - \inf F\right)\right)$$

basic operations. Notice that the coefficient of $1/\varepsilon^3$ scales at worst as $\sqrt{m}$ — a significant improvement over (1.5). We note that a different complementary approach, generalizing stochastic subgradient methods, has been recently pursued by Duchi-Ruan [25].

5. **(Acceleration)** The final contribution of the paper concerns acceleration of the (exact) prox-linear method. For additive composite problems, with c in addition convex, the prox-gradient method is suboptimal from the viewpoint of computational complexity [47, 48]. Accelerated gradient methods, beginning with Nesterov [50] and extended by Beck-Teboulle [4] achieve a superior rate in terms of function values. Later, Nesterov in [49, Page 11, item 2] showed that essentially the same accelerated schemes also achieve a superior rate of $\mathcal{O}((\frac{\beta}{\varepsilon})^{2/3})$ in terms of stationarity, and even a faster rate is possible by first regularizing the problem [49, Page 11, item 3].[4] 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 [30], 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 [10] is a good example.

   In this work, we extend the accelerated algorithm of Ghadimi-Lan [30] for additive composite problems to the entire problem class (1.1), with inexact subproblem solves. Assuming the diameter $M := \mathrm{diam}(\mathrm{dom}\, g)$ is finite, 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. The inexact analysis of the proposed accelerated method based on functional errors is inspired

[4] The short paper [54] only considered smooth unconstrained minimization; however, a minor modification of the proof technique extends to the convex additive composite setting.

by and shares many features with the seminal papers [40, 62] for convex additive composite problems (1.2).

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, first-order stationarity, and the basic prox-linear method. Section 4 discusses weak-convexity of the composite function and the relationship of the prox-gradient with the gradient of the Moreau envelope. Section 5 analyzes inexact prox-linear methods based on two models of inexactness: near-minimality and dual near-stationarity. In Section 6, we derive efficiency estimates of first-order methods for the composite problem class, based on a smoothing strategy. Section 7 extends the aforementioned results to problems where one seeks to minimize a finite average of composite functions. The final Section 8 discusses an inertial prox-linear algorithm that is adaptive to convexity.

2 Notation

The notation we follow is standard. Throughout, we consider a Euclidean space, denoted by $\mathbf{R}^d$, with an inner product $\langle \cdot, \cdot \rangle$ and the induced norm $\| \cdot \|$. Given a linear map $A \colon \mathbf{R}^d \to \mathbf{R}^l$, the adjoint $A^* \colon \mathbf{R}^l \to \mathbf{R}^d$ is the unique linear map satisfying

$$\langle Ax, y \rangle = \langle x, A^* y \rangle \qquad \text{for all } x \in \mathbf{R}^d, y \in \mathbf{R}^l.$$

The operator norm of A, defined as $\|A\|_{\mathrm{op}} := \max_{\|u\| \leq 1} \|Au\|$, coincides with the maximal singular value of A and satisfies $\|A\|_{\mathrm{op}} = \|A^*\|_{\mathrm{op}}$. For any map $F \colon \mathbf{R}^d \to \mathbf{R}^m$, we set

$$\mathrm{lip}\,(F) := \sup_{x \neq y} \frac{\|F(y) - F(x)\|}{\|y - x\|}.$$

In particular, we say that F is L-Lipschitz continuous, for some real $L \geq 0$, if the inequality $\mathrm{lip}\,(F) \leq L$ holds. Given a set Q in $\mathbf{R}^d$, the *distance* and *projection* of a point x onto Q are given by

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

respectively. The extended-real-line is the set $\overline{\mathbf{R}} := \mathbf{R} \cup \{\pm\infty\}$. The *domain* and the *epigraph* of any function $f \colon \mathbf{R}^d \to \overline{\mathbf{R}}$ are the sets

$$\mathrm{dom}\, f := \{x \in \mathbf{R}^d : f(x) < +\infty\}, \qquad \mathrm{epi}\, f := \{(x, r) \in \mathbf{R}^d \times \mathbf{R} : f(x) \leq r\},$$

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

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

$$f(y) \geq f(x) + \langle v, y - x \rangle \qquad \text{holds for all } y \in \mathbf{R}^d. \tag{2.1}$$

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 the *Fenchel conjugate* $f^\star \colon \mathbf{R}^d \to \overline{\mathbf{R}}$, defined by

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

If f is closed and convex, then equality $f = f^{\star\star}$ holds and we have the equivalence

$$y \in \partial f(x) \qquad \Longleftrightarrow \qquad x \in \partial f^\star(y). \tag{2.2}$$

For any function f and real $\nu > 0$, 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)**.** *Let $f \colon \mathbf{R}^d \to \mathbf{R}$ be a closed, convex function. Then f_ν is convex and C^1-smooth with*

$$\nabla f_\nu(x) = \nu^{-1}(x - \operatorname{prox}_{\nu f}(x)) \qquad \textit{and} \qquad \operatorname{lip}(\nabla f_\nu) \le \tfrac{1}{\nu}.$$

If in addition f is L-Lipschitz, then the envelope $f_\nu(\cdot)$ is L-Lipschitz and satisfies

$$0 \le f(x) - f_\nu(x) \le \frac{L^2\nu}{2} \qquad \textit{for all } x \in \mathbf{R}^d. \tag{2.3}$$

Proof. The expression $\nabla f_\nu(x) = \nu^{-1}(x - \operatorname{prox}_{\nu f}(x)) = \nu^{-1} \cdot \operatorname{prox}_{(\nu f)^*}(x)$ can be found in [60, Theorem 31.5]. The inequality $\operatorname{lip}(\nabla f_\nu) \le \frac{1}{\nu}$ then follows since the proximal mapping of a closed convex function is 1-Lipschitz [60, pp. 340]. The expression (2.3) 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. [60, Theorem 16.4]) and noting that the domain of $f^\star$ is bounded in norm by L. Finally, to see that f_ν is L-Lipschitz, observe $\nabla f_\nu(x) \in \partial f(\operatorname{prox}_{\nu f}(x))$ for all x, and hence $\|\nabla f_\nu(x)\| \le \sup\{\|v\| : y \in \mathbf{R}^d, v \in \partial f(y)\} \le L$. □

3 The composite problem class

This 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. $g \colon \mathbf{R}^d \to \overline{\mathbf{R}}$ is a proper, closed, convex function;

2. $h\colon \mathbf{R}^m \to \mathbf{R}$ is a convex and L-Lipschitz continuous function:

$$|h(x) - h(y)| \le L\|x - y\| \qquad \text{for all } x, y \in \mathbf{R}^m;$$

3. $c\colon \mathbf{R}^d \to \mathbf{R}^m$ is a C^1-smooth mapping with a β-Lipschitz continuous Jacobian map:

$$\|\nabla c(x) - \nabla c(y)\|_{\text{op}} \le \beta\|x - y\| \qquad \text{for all } x, y \in \mathbf{R}^d.$$

The values L and β will often multiply each other; hence, we define the constant $\mu := L\beta$.

3.1 Motivating examples

It is instructive to consider some motivating examples fitting into the 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:

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

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 [4,54]. We will often compare and contrast techniques for general composite problems (3.1) with those specialized to this additive composite setting.

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

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

Gauss-Newton type algorithm [37, 43, 45] 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 $f\colon \mathbf{R}^d \to \mathbf{R}$ and $G\colon \mathbf{R}^d \to \mathbf{R}^m$ are smooth mappings and $\mathcal{K} \subseteq \mathbf{R}^m$ is a closed convex cone. An accompanying *penalty formulation* – ubiquitous in nonlinear optimization [9, 12, 15, 21, 27] – 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 (Statistical estimation)**.** 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 used in the Least Absolute Deviations (LAD) technique in regression [46, 66], Kalman smoothing with impulsive disturbances [2], and for robust phase retrieval [25].

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 [14, 26, 34, 39], 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 $\mathrm{lip}\,(\nabla h) \sim 1/\kappa$. 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 better 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 $\nabla c(x)$ 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 hard-to-compute mapping c and the user chosen penalty h can help in designing algorithms. See for example Conn-Scheinberg-Vicente [16] and Wild [69], and references therein.

3.2 First-order stationary points for composite problems

Let us now explain the goal of algorithms for the problem class (3.1). 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 (or generalized derivatives), which have a very explicit representation for our problem class. We recall here the relevant definitions, following the monographs of Mordukhovich [44] and Rockafellar-Wets [61].

Consider an arbitrary function $f\colon \mathbf{R}^d \to \overline{\mathbf{R}}$ and a point $\bar{x}$ with $f(\bar{x})$ finite. The *Fréchet subdifferential* of f at $\bar{x}$, denoted $\hat{\partial} f(\bar{x})$, is the set of all vectors v satisfying

$$f(x) \geq f(\bar{x}) + \langle v, x - \bar{x}\rangle + o(\|x - \bar{x}\|) \qquad \text{as } x \to \bar{x}.$$

Thus the inclusion $v \in \hat{\partial} f(\bar{x})$ holds precisely when the affine function $x \mapsto f(\bar{x}) + \langle v, x - \bar{x}\rangle$ underestimates f up to first-order near $\bar{x}$. In general, the limit of Fréchet subgradients $v_i \in \hat{\partial} f(x_i)$, along a sequence $x_i \to \bar{x}$, may not be a Fréchet subgradient at the limiting point $\bar{x}$. Hence, one formally enlarges the Fréchet subdifferential and defines the *limiting subdifferential* of f at $\bar{x}$, denoted $\partial f(\bar{x})$, to consist of all vectors v for which there exist sequences x_i and v_i, satisfying $v_i \in \partial f(x_i)$ and $(x_i, f(x_i), v_i) \to (\bar{x}, f(\bar{x}), v)$. We say that x is *stationary* for f if the inclusion $0 \in \partial f(x)$ holds.

For convex functions f, the subdifferentials $\hat{\partial} f(x)$ and $\partial f(x)$ coincide with the subdifferential in the sense of convex analysis (2.1), while for C^1-smooth functions f, they consist only of the

gradient $\nabla f(x)$. Similarly, the situation simplifies for the composite problem class (3.1): the two subdifferentials $\hat{\partial} F$ and ∂F coincide and admit an intuitive representation through a chain-rule [61, Theorem 10.6, Corollary 10.9].

Theorem 3.1 (Chain rule)**.** *For the composite function F, defined in* (3.1)*, the Fréchet and limiting subdifferentials coincide and admit the representation*

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

In summary, the algorithms we consider aim to find stationary points of F, i.e. those points x satisfying $0 \in \partial F(x)$. In "primal terms", it is worth noting that a point x is stationary for F if and only if the directional derivative of F at x is nonnegative in every direction [61, Proposition 8.32]. More precisely, the equality holds:

$$\operatorname{dist}(0; \partial F(x)) = - \inf_{v:\, \|v\| \leq 1} F'(x; v), \tag{3.3}$$

where $F'(x; v)$ is the directional derivative of F at x in direction v [61, Definition 8.1].

3.3 The prox-linear method

The basic algorithm we rely on for the composite problem class is the so-called prox-linear method. To motivate this scheme, let us first consider the setting of additive composite minimization (3.2). The most basic algorithm in this setting is the *proximal gradient method* [4, 54]

$$x_{k+1} := \operatorname*{argmin}_x \left\{ c(x_k) + \langle \nabla c(x_k), x - x_k \rangle + g(x) + \frac{1}{2t} \|x - x_k\|^2 \right\}, \tag{3.4}$$

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 (3.4) 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 method, we use the following notation. For any points $z, y \in \mathbf{R}^d$ and a real $t > 0$, define

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

Throughout the manuscript, we will routinely use the following estimate on the error in approximation $|F(z) - F(z; y)|$. We provide a quick proof for completeness.

Lemma 3.2. *For all $x, y \in \operatorname{dom} g$, the inequalities hold:*

$$-\frac{\mu}{2} \|z - y\|^2 \leq F(z) - F(z; y) \leq \frac{\mu}{2} \|z - y\|^2. \tag{3.5}$$

Proof. Since h is L-Lipschitz, we have $|F(z) - F(z;y)| \le L\big\|c(z) - \big(c(y) + \nabla c(y)(z-y)\big)\big\|$. The fundamental theorem of calculus, in turn, implies

$$\begin{aligned}\big\|c(z) - \big(c(y) + \nabla c(y)(z-y)\big)\big\| &= \left\|\int_0^1 \Big(\nabla c(y + t(z-y)) - \nabla c(y)\Big)(z-y)\,dt\right\| \\ &\le \int_0^1 \|\nabla c(y + t(z-y)) - \nabla c(y)\|_{\mathrm{op}}\,\|z-y\|\,dt \\ &\le \beta\|z-y\|^2\left(\int_0^1 t\,dt\right) = \frac{\beta}{2}\|z-y\|^2.\end{aligned}$$

The result follows. □

In particular, Lemma 3.2 implies that $F_t(\cdot\,;y)$ is an upper model for F for any $t \le \mu^{-1}$, meaning $F_t(z;y) \ge F(z)$ for all points $y, z \in \operatorname{dom} g$. The *prox-linear method*, formalized in Algorithm 1, is then simply the recurrence $x_{k+1} = S_t(x_k)$. Notice that we are implicitly assuming here that the proximal subproblem (3.6) is solvable. We will discuss the impact of an inexact evaluation of $S_t(\cdot)$ in Section 5. Specializing to 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 (3.4).

Algorithm 1: Prox-linear method

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

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

$$x_{k+1} = \operatorname*{argmin}_x \left\{ g(x) + h\Big(c(x_k) + \nabla c(x_k)(x - x_k)\Big) + \frac{1}{2t}\|x - x_k\|^2 \right\}. \tag{3.6}$$

The convergence rate of the prox-linear method is 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 $\min_z F_t(z;x)$ read

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

In particular, it is straightforward to check that with any $t > 0$, a point x is 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". In Section 4, we will establish a much more rigorous justification for why the norm $\|\mathcal{G}_t(x)\|$ provides a reliable basis for judging the quality of the point x. Let us review here the rudimentary convergence guarantees of the method in terms of the prox-gradient, as presented for example in [23, Section 5]. We provide a quick proof for completeness.

Proposition 3.3 (Efficiency of the pure prox-linear method)**.** *Supposing $t \le \mu^{-1}$, the iterates generated by Algorithm 1 satisfy*

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

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

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

Summing the inequalities yields

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

as claimed. □

4 Prox-gradient size $\|\mathcal{G}_t\|$ and approximate stationarity

Before continuing the algorithmic development, let us take a closer look at what the measure $\|\mathcal{G}_t(x)\|$ tells us about "near-stationarity" of the point x. Let us first consider the additive composite setting (3.2), where the impact of the measure $\|\mathcal{G}_t(x)\|$ on near-stationarity is well-understood. As discussed on page 10, the prox-linear method reduces to the prox-gradient recurrence

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

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

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

or equivalently

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

Notice that the right-hand-side is exactly $\partial F(x_{k+1})$. Taking into account that ∇c is β-Lipschitz, we deduce

$$\begin{aligned}\operatorname{dist}(0; \partial F(x_{k+1})) &\leq \|\mathcal{G}_{\frac{1}{\beta}}(x_k)\| + \|\nabla c(x_{k+1}) - \nabla c(x_k)\| \\ &\leq 2\|\mathcal{G}_{\frac{1}{\beta}}(x_k)\|.\end{aligned} \tag{4.1}$$

Thus the inequality $\|\mathcal{G}_{\frac{1}{\beta}}(x_k)\| \leq \varepsilon/2$ indeed guarantees that x_{k+1} is nearly stationary for F in the sense that $\operatorname{dist}(0; \partial F(x_{k+1})) \leq \varepsilon$. Taking into account (3.3), we deduce the bound on directional derivative $F'(x; u) \geq -\varepsilon$ in any unit direction u. With this in mind, the guarantee of Proposition 3.3 specialized to the prox-gradient method can be found for example in [54, Theorem 3].

The situation is dramatically different for the general composite class (3.1). When h is nonsmooth, the quantity $\operatorname{dist}(0; \partial F(x_{k+1}))$ will typically not even tend to zero in the limit, in spite of $\|\mathcal{G}_{\frac{1}{\beta}}(x_k)\|$ tending to zero. For example, the prox-linear algorithm applied to the univariate function $f(x) = |x^2 - 1|$ and initiated at $x > 1$, will generate a decreasing sequence $x_k \to 1$ with $f'(x_k) \to 2$.[5]

Thus we must look elsewhere for an interpretation of the quantity $\|\mathcal{G}_{\frac{1}{\mu}}(x_k)\|$. We will do so by focusing on the Moreau envelope $x \mapsto F_{\frac{1}{2\mu}}(x)$ — a function that serves as a C^1-smooth

[5]Notice f has three stationary points $\{-1, 0, 1\}$. Fix $y > 1$ and observe that x minimizes $f_t(\cdot; y)$ if and only if $\frac{y-x}{2ty} \in \partial|\cdot|(y^2 - 1 + 2y(x-y))$. Hence $\frac{y-x}{2ty}\cdot(y^2 - 1 + 2y(x-y)) \geq 0$. The inequality $x \leq 1$ would immediately imply a contradiction. Thus the inequality $x_0 > 1$ guarantees $x_k > 1$ for all k. The claim follows.

approximation of F with the same stationary points. We argue in Theorem 4.5 that the norm of the prox-gradient $\|\mathcal{G}_{\frac{1}{\mu}}(x_k)\|$ is informative because $\|\mathcal{G}_{\frac{1}{\mu}}(x_k)\|$ is proportional to the norm of the true gradient of the Moreau envelope $\|\nabla F_{\frac{1}{2\mu}}(x)\|$. Before proving this result, we must first establish some basic properties of the Moreau envelope, which will follow from weak convexity of the composite function F; this is the content of the following section.

4.1 Weak convexity and the Moreau envelope of the composition

We will need the following standard definition.

Definition 4.1 (Weak convexity)**.** We say that a function $f\colon \mathbf{R}^d \to \overline{\mathbf{R}}$ 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:

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

It is well-known that for a locally Lipschitz function $f\colon \mathbf{R}^d \to \mathbf{R}$, the following are equivalent; see e.g. [17, Theorem 3.1].

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

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

In particular, the following is true.

Lemma 4.2 (Weak convexity of the composition)**.**
The function $h \circ c$ is ρ-weakly convex on $\mathbf{R}^d$ for some $\rho \in [0, \mu]$.

Proof. To simplify notation, set $\Phi := h \circ c$. Fix two points $x, y \in \mathbf{R}^d$ 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 and the inequality $\|c(y) - c(x) - \nabla c(x)(y-x)\| \leq \frac{\beta}{2}\|y-x\|^2$, 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 of F has an immediate consequence on the Moreau envelope F_ν.

Lemma 4.3 (Moreau envelope of the composite function)**.** *Fix $\nu \in (0, 1/\mu)$. Then the proximal map $\operatorname{prox}_{\nu F}(x)$ is well-defined and single-valued, while the Moreau envelope F_ν is C^1-smooth with gradient*

$$\nabla F_\nu(x) = \nu^{-1}(x - \operatorname{prox}_{\nu F}(x)). \tag{4.2}$$

Moreover, stationary points of F_ν and of F coincide.

Proof. Fix $\nu \in (0, 1/\mu)$. Lemma 4.2 together with [57, Theorem 4.4] immediately imply that $\operatorname{prox}_{\nu F}(x)$ is well-defined and single-valued, while the Moreau envelope F_ν is C^1-smooth with gradient given by (4.2). Equation (4.2) then implies that x is stationary for F_ν if and only if x minimizes the function $\varphi(z) := F(z) + \frac{1}{2\nu}\|z - x\|^2$. Lemma 4.2 implies that φ is strongly convex, and therefore the unique minimizer z of φ is characterized by $\nu^{-1}(x - z) \in \partial F(z)$. Hence stationary points of F_ν and of F coincide. □

Thus for $\nu \in (0, 1/\mu)$, stationary points of F coincide with those of the C^1-smooth function F_ν. More useful would be to understand the impact of $\|\nabla F_\nu(x)\|$ being small, but not zero. To this end, observe the following. Lemma 4.3 together with the definition of the Moreau envelope implies that for any x, the point $\hat{x} := \operatorname{prox}_{\nu F}(x)$ satisfies

$$\begin{cases} \|\hat{x} - x\| & \le \nu\|\nabla F_\nu(x)\|, \\ F(\hat{x}) & \le F(x), \\ \operatorname{dist}(0; \partial F(\hat{x})) & \le \|\nabla F_\nu(x)\|. \end{cases} \tag{4.3}$$

Thus a small gradient $\|\nabla F_\nu(x)\|$ implies that x is *near* a point $\hat{x}$ that is *nearly stationary* for F.

4.2 Prox-gradient and the gradient of the Moreau envelope

The final ingredient we need to prove Theorem 4.5 is the following lemma [6, Theorem 2.4.1]; we provide a short proof for completeness.

Lemma 4.4 (Quadratic penalization principle)**.** *Consider a closed function $f\colon \mathbf{R}^d \to \overline{\mathbf{R}}$ and suppose the inequality $f(x) - \inf f \le \varepsilon$ holds for some point x and real $\varepsilon > 0$. Then for any $\lambda > 0$, the inequality holds:*

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

If f is α-strongly convex (possibly with $\alpha = 0$), then the estimate improves to

$$\|\lambda^{-1}(x - \operatorname{prox}_{\lambda f}(x))\| \le \sqrt{\frac{\varepsilon}{\lambda(1 + \frac{\lambda\alpha}{2})}}.$$

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

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

Hence we deduce $\lambda^{-1}\|y - x\| \le \sqrt{\frac{2\varepsilon}{\lambda}}$, as claimed. If f is α-strongly convex, then the function $z \mapsto f(z) + \frac{1}{2\lambda}\|z - x\|^2$ is $(\alpha + \lambda^{-1})$-strongly convex and therefore

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

The claimed inequality follows along the same lines. □

We can now quantify the precise relationship between the norm of the prox-gradient $\|\mathcal{G}_t(x)\|$ and the norm of the true gradient of the Moreau envelope $\|\nabla F_{\frac{t}{1+t\mu}}(x)\|$.

Theorem 4.5 (Prox-gradient and near-stationarity)**.** *For any point x and real constant $t > 0$, the inequality holds:*

$$\tfrac{1}{(1+\mu t)(1+\sqrt{\mu t})}\left\|\nabla F_{\frac{t}{1+t\mu}}(x)\right\| \le \|\mathcal{G}_t(x)\| \le \tfrac{1+2t\mu}{1+t\mu}\left(\sqrt{\tfrac{t\mu}{1+t\mu}}+1\right)\left\|\nabla F_{\frac{t}{1+t\mu}}(x)\right\|. \tag{4.4}$$

Proof. To simplify notation, throughout the proof set

$$\bar{x} := S_t(x) = \operatorname*{argmin}_z F_t(z;x),$$
$$\hat{x} := \operatorname{prox}_{\frac{tF}{1+t\mu}}(x) = \operatorname*{argmin}_z \ \{F(z) + \tfrac{\mu+t^{-1}}{2}\|z-x\|^2\}.$$

Notice that $\hat{x}$ is well-defined by Lemma 4.3.

We begin by establishing the first inequality in (4.4). For any point z, we successively deduce

$$\begin{aligned} F(z) &\ge F_t(z;x) - \tfrac{\mu+t^{-1}}{2}\|z-x\|^2 \ge F_t(\bar{x};x) + \tfrac{1}{2t}\|\bar{x}-z\|^2 - \tfrac{\mu+t^{-1}}{2}\|z-x\|^2 \\ &\ge F(\bar{x}) + \frac{1}{2t}\|\bar{x}-z\|^2 - \tfrac{\mu+t^{-1}}{2}\|z-x\|^2 + \tfrac{t^{-1}-\mu}{2}\|\bar{x}-x\|^2, \end{aligned} \tag{4.5}$$

where the first and third inequalities follow from (3.5) and the second from strong convexity of $F_t(\cdot;x)$.

Define the function $\zeta(z) := F(z) + \frac{\mu+t^{-1}}{2}\|z-x\|^2 - \frac{1}{2t}\|\bar{x}-z\|^2$ and notice that ζ is convex by Lemma 4.2. Inequality (4.5) directly implies

$$\zeta(\bar{x}) - \inf\zeta \le \left(F(\bar{x}) + \tfrac{\mu+t^{-1}}{2}\|\bar{x}-x\|^2\right) - \left(F(\bar{x}) + \tfrac{t^{-1}-\mu}{2}\|\bar{x}-x\|^2\right) = \mu\|\bar{x}-x\|^2.$$

Notice the relation, $\operatorname{prox}_{t\zeta}(\bar{x}) = \operatorname{prox}_{\frac{tF}{1+t\mu}}(x) = \hat{x}$. Setting $\lambda := t$ and $\varepsilon := \mu\|\bar{x}-x\|^2$ and using Lemma 4.4 (convex case $\alpha = 0$) with $\bar{x}$ in place of x, we conclude

$$\sqrt{\tfrac{\mu}{t}}\|\bar{x}-x\| \ge \|t^{-1}(\bar{x} - \operatorname{prox}_{t\zeta}(\bar{x}))\| = \|t^{-1}(\bar{x}-\hat{x})\| \ge \|t^{-1}(x-\hat{x})\| - \|t^{-1}(\bar{x}-x)\|.$$

Rearranging and using (4.2) yields the first inequality in (4.4), as claimed.

We next establish the second inequality in (4.4). The argument is in the same spirit as the previous part of the proof. For any point z, we successively deduce

$$\begin{aligned} F_t(z;x) &\ge (F(z) + \tfrac{\mu+t^{-1}}{2}\|z-x\|^2) - \mu\|z-x\|^2 \\ &\ge F(\hat{x}) + \tfrac{\mu+t^{-1}}{2}\|\hat{x}-x\|^2 + \tfrac{1}{2t}\|\hat{x}-z\|^2 - \mu\|z-x\|^2, \end{aligned} \tag{4.6}$$

where the first inequality follows from (3.5) and the second from t^{-1}-strong convexity of $z \mapsto F(z) + \frac{\mu+t^{-1}}{2}\|z-x\|^2$. Define now the function

$$\Psi(z) := F_t(z;x) - \tfrac{1}{2t}\|\hat{x}-z\|^2 + \mu\|z-x\|^2.$$

Combining (3.5) and (4.6), we deduce

$$\Psi(\hat{x}) - \inf\Psi \le \left(F_t(\hat{x};x) + \mu\|\hat{x}-x\|^2\right) - \left(F(\hat{x}) + \tfrac{\mu+t^{-1}}{2}\|\hat{x}-x\|^2\right) \le \mu\|\hat{x}-x\|^2.$$

Notice that Ψ is strongly convex with parameter $\alpha := 2\mu$. Setting $\varepsilon := \mu\|\hat{x}-x\|^2$ and $\lambda = t$, and applying Lemma 4.4 with $\hat{x}$ in place of x, we deduce

$$\sqrt{\tfrac{\mu}{t(1+t\mu)}}\|\hat{x}-x\| \ge \|t^{-1}(\hat{x} - \operatorname{prox}_{t\Psi}(\hat{x}))\| \ge \|t^{-1}(x - \operatorname{prox}_{t\Psi}(\hat{x}))\| - \|t^{-1}(\hat{x}-x)\|. \tag{4.7}$$

To simplify notation, set $\hat{z} := \mathrm{prox}_{t\Psi}(\hat{x})$. By definition of Ψ, equality

$$\hat{z} = \operatorname*{argmin}_z \ \left\{ F_t(z;x) + \mu\|z-x\|^2 \right\} \qquad \text{holds,}$$

and therefore $2\mu(x-\hat{z}) \in \partial F_t(\hat{z};x)$. Taking into account that $F_t(\cdot;x)$ is t^{-1}-strongly convex, we deduce

$$\|2\mu(x-\hat{z})\| \geq \mathrm{dist}\,(0;\partial F_t(\hat{z};x)) \geq t^{-1}\|\hat{z}-\bar{x}\| \geq \|t^{-1}(x-\bar{x})\| - \|t^{-1}(x-\hat{z})\|.$$

Rearranging and combining the estimate with (4.2), (4.7) yields the second inequality in (4.4). □

In the most important setting $t = 1/\mu$, Theorem 4.5 reduces to the estimate

$$\tfrac{1}{4}\left\|\nabla F_{\frac{1}{2\mu}}(x)\right\| \leq \left\|\mathcal{G}_{1/\mu}(x)\right\| \leq \tfrac{3}{2}\left(1+\tfrac{1}{\sqrt{2}}\right)\left\|\nabla F_{\frac{1}{2\mu}}(x)\right\|. \tag{4.8}$$

A closely related result has recently appeared in [23, Theorem 5.3], with a different proof, and has been extended to a more general class of Taylor-like approximations in [22]. Combining (4.8) and (4.3) we deduce that for any point x, there exists a point $\hat{x}$ (namely $\hat{x} = \mathrm{prox}_{F/2\mu}(x)$)) satisfying

$$\begin{cases} \|\hat{x}-x\| & \leq \frac{2}{\mu}\|\mathcal{G}_{1/\mu}(x)\|, \\ F(\hat{x}) & \leq F(x), \\ \mathrm{dist}(0;\partial F(\hat{x})) & \leq 4\|\mathcal{G}_{1/\mu}(x)\|. \end{cases} \tag{4.9}$$

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 $\mathrm{prox}_{F/2\mu}$. Computing $\hat{x}$ is not the point, however; the sole purpose of $\hat{x}$ is to certify that x is approximately stationary in the sense of (4.9).

5 Inexact analysis of the prox-linear method

In practice, it is often impossible to solve the proximal subproblems $\min_z F_t(z;y)$ exactly. In this section, we explain the effect of inexactness in the proximal subproblems (3.6) on the overall performance of the prox-linear algorithm. 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 (e.g. distance of initial iterate to the optimal solution). Thus saddle-point methods cannot be directly 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 prox-linear method with termination based on functional near-optimality. Indeed, the 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. The arguments are quick modifications of the proof of Proposition 3.3.

5.1 Near-optimality in the subproblems

We first 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 $f\colon \mathbf{R}^d \to \overline{\mathbf{R}}$ whenever the inequality holds:

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

Consider now a sequence of tolerances $\varepsilon_k \geq 0$ 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 $F_t(\cdot; x_k)$. We record this scheme in Algorithm 2.

Algorithm 2: Inexact prox-linear method: near-optimality
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: ($k \geq 0$) Set x_{k+1} to be an 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 used as surrogates for the true stationarity measure $\|\mathcal{G}_t(x_k)\|$.

Lemma 5.1. *Suppose x^+ is an ε-approximate minimizer of $F_t(\cdot; x)$. Then the inequality holds:*

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

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

$$\begin{aligned} \|\mathcal{G}_t(x)\|^2 &\leq \frac{4}{t} \cdot \frac{1}{2t}\left\|x^+ - z^*\right\|^2 + 2\left\|t^{-1}(x^+ - x)\right\|^2 \\ &\leq \frac{4}{t} \cdot \big(F_t(x^+; x) - F_t(z^*; x)\big) + 2\left\|t^{-1}(x^+ - x)\right\|^2 \\ &\leq \frac{4}{t} \cdot \varepsilon + 2\left\|t^{-1}(x^+ - x)\right\|^2, \end{aligned} \tag{5.1}$$

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 $F_t(\cdot; x)$. □

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

Theorem 5.2 (Convergence of the inexact prox-linear algorithm: near-optimality)**.** *Supposing $t \leq \mu^{-1}$, the iterates generated by Algorithm 2 satisfy*

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

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

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

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

Then the inequality $t \leq \mu^{-1}$ along with (3.5) implies that $F_t(\cdot; x_k)$ is an upper model of $F(\cdot)$ and therefore

$$F(x_k) \geq F(x_{k+1}) - \varepsilon_{k+1} + \tfrac{t}{2}\left\|\mathcal{G}_t(x_k)\right\|^2. \tag{5.2}$$

We conclude

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

The proof is complete. □

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 $q > 0$.

5.2 Near-stationarity in the subproblems

In the previous section, we considered the effect of solving the proximal subproblems up to an accuracy in functional error. We now consider instead a model of inexactness for the proximal subproblems based on near-stationarity. A first naive attempt would be to consider a point z to be ε-stationary for the proximal subproblem, $\min F_t(\cdot; x)$, if it satisfies

$$\operatorname{dist}(0; \partial_z F_t(z; x)) \leq \varepsilon.$$

This assumption, however, is not reasonable since first-order methods for this problem do not produce such points z, unless h is smooth. Instead, let us look at the Fenchel dual problem. To simplify notation, write the target subproblem $\min F_t(\cdot; x)$ as

$$\min_z \; h(b - Az) + G(z) \tag{5.3}$$

under the identification $G(z) = g(z) + \frac{1}{2t}\|z - x\|^2$, $A = -\nabla c(x)$, and $b = c(x) - \nabla c(x)x$. Notice that G is t^{-1}-strongly convex and therefore $G^\star$ is C^1-smooth with t-Lipschitz gradient. The Fenchel dual problem, after negation, takes the form [61, Example 11.41]:

$$\min_w \; \varphi(w) := G^\star(A^* w) - \langle b, w\rangle + h^\star(w). \tag{5.4}$$

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 convex term $h^\star$. Later on, when x depends on an iteration counter k, we will use the notation φ_k, G_k, A_k, b_k instead to make precise that these objects depend on k.

Typical first-order methods, such as prox-gradient and its accelerated variants can generate a point w for the problem (5.4) satisfying

$$\operatorname{dist}(0; \partial\varphi(w)) \leq \varepsilon \tag{5.5}$$

up to any specified tolerance $\varepsilon > 0$. Such schemes in each iteration only require evaluation of the gradient of the smooth function $G^\star(A^*w) - \langle b, w\rangle$ along with knowledge of a Lipschitz constant of the gradient, and evaluation of the proximal map of $h^\star$. For ease of reference, we record these quantities here in terms of the original functional components of the composite problem (3.1). Since the proof is standard, we have placed it in Appendix A.

Lemma 5.3. *The following are true for all points z and w and real $t > 0$:*

- *The equation holds:*
$$\operatorname{prox}_{th^\star}(w) = t\left(w/t - \operatorname{prox}_{h/t}(w/t)\right). \tag{5.6}$$
- *The equations hold:*
$$G^\star(z) = (g^\star)_{1/t}(z + x/t) - \tfrac{1}{2t}\|x\|^2 \qquad \text{and} \qquad \nabla G^\star(z) = \operatorname{prox}_{tg}(x + tz). \tag{5.7}$$
Consequently, the gradient map $\nabla\Big(G^\star \circ A^* - \langle \cdot, b\rangle\Big)$ is Lipschitz continuous with constant $t\|\nabla c(x)\|_{op}^2$ and admits the representation:
$$\nabla\Big(G^\star \circ A^* - \langle b, \cdot\rangle\Big)(w) = \nabla c(x)\Big(x + \operatorname{prox}_{tg}(x - t\nabla c(x)^* w)\Big) - c(x). \tag{5.8}$$

Thus, suppose we have found a point w satisfying (5.5). How can we then generate a primal iterate x^+ at which to form the prox-linear subproblem for the next step? The following lemma provides a simple recipe for doing exactly that. It shows how to generate from w a point that is a true minimizer to a slight perturbation of the proximal subproblem.

Lemma 5.4 (Primal recovery from dual ε-stationarity)**.** *Let φ be the function defined in* (5.4)*. Fix a point $w \in \operatorname{dom}\varphi$ and a vector $\zeta \in \partial\varphi(w)$. Then the point $\bar{x} := \nabla G^\star(A^*w)$ is the true minimizer of the problem*
$$\min_z \; h(\zeta + b - Az) + G(z). \tag{5.9}$$

Proof. Appealing to the chain rule, $\partial\varphi(w) = A\nabla G^\star(A^*w) - b + \partial h^\star(w)$, we deduce

$$\zeta + b \in A\nabla G^\star(A^*w) + \partial h^\star(w) = A\bar{x} + \partial h^\star(w).$$

The relation (2.2) then implies $w \in \partial h(\zeta + b - A\bar{x})$. Applying A^* to both sides and rearranging yields

$$0 \in -A^*\partial h(\zeta + b - A\bar{x}) + A^*w \subseteq -A^*\partial h(\zeta + b - A\bar{x}) + \partial G(\bar{x}),$$

where the last inclusion follows from applying (2.2) to G. The right-hand-side is exactly the subdifferential of the objective function in (5.9) evaluated at $\bar{x}$. The result follows. □

This lemma directly motivates the following inexact extension of the prox-linear algorithm (Algorithm 3), based on dual near-stationary points.

Algorithm 3 is stated in a way most useful for convergence analysis. On the other hand, it is not very explicit. To crystallize the ideas, let us concretely describe how one can implement step k of the scheme. First, we find a point w_{k+1} that is ε_{k+1}-stationary for the dual problem (5.4). More precisely, we find a pair (w_{k+1}, ζ_{k+1}) satisfying $\zeta_{k+1} \in \partial\varphi_k(w_{k+1})$ and $\|\zeta_{k+1}\| \le \varepsilon_{k+1}$. We

Algorithm 3: 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: $(k \geq 0)$ Find (x_{k+1}, ζ_{k+1}) such that $\|\zeta_{k+1}\| \leq \varepsilon_{k+1}$ and x_{k+1} is the minimizer of the function

$$z \mapsto g(z) + h\Big(\zeta_{k+1} + c(x_k) + \nabla c(x_k)(z - x_k)\Big) + \frac{1}{2t}\|z - x_k\|^2. \tag{5.10}$$

can achieve this by a proximal gradient method (or its accelerated variants) on the dual problem (5.4). Then combining Lemma 5.4 with equation (5.7), we conclude that we can simply set

$$x_{k+1} := \nabla G^{\star}(A^* w_{k+1}) = \operatorname{prox}_{tg}(x_k - t\nabla c(x_k)^* w_{k+1}).$$

We record this more explicit description of Algorithm 3 in Algorithm 4. The reader should keep in mind that even though Algorithm 4 is more explicit, the convergence analysis we present will use the description in Algorithm 3.

Algorithm 4: Inexact prox-linear method: near-stationarity (explicit)

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: $(k \geq 0)$ Define the function

$$\varphi_k(w) := (g^{\star})_{1/t}\Big(x_k/t - \nabla c(x_k)^* w\Big) - \big\langle c(x_k) - \nabla c(x_k) x_k, w\big\rangle + h^{\star}(w).$$

Find a point w_{k+1} satisfying $\operatorname{dist}(0; \partial\varphi_k(w_{k+1})) \leq \varepsilon_{k+1}$.

Set $x_{k+1} = \operatorname{prox}_{tg}(x_k - t\nabla c(x_k)^* w_{k+1})$.

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 5.5. *Suppose x^+ is a minimizer of the function*

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

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

$$\|\mathcal{G}_t(x)\|^2 \leq 8Lt^{-1} \cdot \|\zeta\| + 2\left\|t^{-1}(x^+ - x)\right\|^2. \tag{5.11}$$

Proof. Define the function

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

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

$$\begin{aligned}\|\mathcal{G}_t(x)\|^2 &\leq \frac{4}{t} \cdot \frac{1}{2t} \left\|x^+ - z^*\right\|^2 + 2\left\|t^{-1}(x^+ - x)\right\|^2 \\ &\leq \frac{4}{t} \cdot \big(F_t(x^+; x) - F_t(z^*; x)\big) + 2\left\|t^{-1}(x^+ - x)\right\|^2 \\ &\leq \frac{4}{t}(l(x^+) - l(z^*) + 2L\|\zeta\|) + 2\left\|t^{-1}(x^+ - x)\right\|^2 \\ &\leq 8t^{-1}L\|\zeta\| + 2\left\|t^{-1}(x^+ - x)\right\|^2,\end{aligned} \tag{5.12}$$

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 $F_t(\cdot; x)$, and the third follows from Lipschitz continuity of h. □

Theorem 5.6 explains the convergence guarantees of the method; c.f. Proposition 3.3.

Theorem 5.6 (Convergence of the inexact prox-linear method: near-stationarity)**.** *Supposing $t \leq \mu^{-1}$, the iterates generated by Algorithm 3 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 $F^* := \liminf_{k\to\infty} F(x_k)$.

Proof. Observe the inequalities:

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

Since the point x_{k+1} minimizes the $\frac{1}{t}$-strongly convex function in (5.10), we deduce

$$\begin{aligned}F(x_{k+1}) &\leq h\big(\zeta_{k+1} + c(x_k)\big) + g(x_k) + L \cdot \varepsilon_{k+1} - \tfrac{1}{2t}\|x_{k+1} - x_k\|^2 \\ &\leq F(x_k) + 2L \cdot \varepsilon_{k+1} - \tfrac{1}{2t}\|x_{k+1} - x_k\|^2.\end{aligned} \tag{5.13}$$

Summing along the indices $j = 0, \ldots, N-1$ 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 5.5, we deduce

$$\min_{j=0,1,\ldots,N-1} \|\mathcal{G}_t(x_j)\|^2 \leq \frac{1}{N}\sum_{j=0}^{N-1} \|\mathcal{G}_t(x_j)\|^2 \leq \frac{4t^{-1}(F(x_0) - F^* + 4L\sum_{j=1}^N \varepsilon_j)}{N}, \tag{5.14}$$

as claimed. □

In particular, to maintain the same rate in N as the exact prox-linear method in Proposition 3.3, 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 $q > 0$.

6 Overall complexity for the composite problem class

In light of the results of Section 5, we can now use the inexact prox-linear method to derive efficiency estimates for the composite problem class (3.1), 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* – evaluations of $c(x)$, matrix-vector multiplications $\nabla c(x)v$ and $\nabla c(x)^* w$, and evaluations of prox_{th}, prox_{tg} – needed to find a point x satisfying $\|\mathcal{G}_t(x)\| \leq \varepsilon$. To simplify the exposition, we will ignore the cost of the evaluation $c(x)$, as it will typically be dominated by the cost of the matrix-vector products $\nabla c(x)v$ and $\nabla c(x)^* w$.

To make progress, in this section we also assume that we have available a real value, denoted $\|\nabla c\|$, satisfying

$$\|\nabla c\| \geq \sup_{x \in \mathrm{dom}\ g} \|\nabla c(x)\|_{\mathrm{op}}.$$

In particular, we assume that the right-hand-side is finite. Strictly speaking, we only need the inequality $\|\nabla c\| \geq \|\nabla c(x_k)\|_{\mathrm{op}}$ to hold along an iterate sequence x_k generated by the inexact prox-linear method. This assumption is completely expected: even when c is a linear map, convergence rates of first-order methods for the composite problem (3.1) depend on some norm of the Jacobian ∇c.

The strategy we propose can be succinctly summarized as follows:

- (Smoothing+prox-linear+fast-gradient) We will replace h by a smooth approximation (Moreau envelope), with a careful choice of the smoothing parameter. Then we will apply an inexact prox-linear method to the smoothed problem, with the proximal subproblems approximately solved by fast-gradient methods.

The basis for the ensuing analysis is the fast-gradient method of Nesterov [54] for minimizing convex additive composite problems. The following section recalls the scheme and records its efficiency guarantees, for ease of reference.

6.1 Interlude: fast gradient method for additive convex composite problems

This section discusses a scheme from [54] that can be applied to any problem of the form

$$\min_x \ f^p(x) := f(x) + p(x), \tag{6.1}$$

where $f\colon \mathbf{R}^d \to \mathbf{R}$ is a convex C^1-smooth function with L_f-Lipschitz gradient and $p\colon \mathbf{R}^d \to \overline{\mathbf{R}}$ is a closed α-strongly convex function ($\alpha \geq 0$). The setting $\alpha = 0$ signifies that p is just convex.

We record in Algorithm 5 the so-called "fast-gradient method" for such problems [54, Accelerated Method].

The method comes equipped with the following guarantee [54, Theorem 6].

Theorem 6.1. *Let x^* be a minimizer of f^p and suppose $\alpha > 0$. Then the iterates x_j generated by Algorithm 5 satisfy:*

$$f^p(x_j) - f^p(x^*) \leq \left(1 + \sqrt{\frac{\alpha}{2L_f}}\right)^{-2(j-1)} \frac{L_f}{4}\|x^* - x_0\|^2.$$

Algorithm 5: Fast gradient method of Nesterov [54, Accelerated Method]

Initialize: Fix a point $x_0 \in \operatorname{dom} p$, set $\theta_0 = 0$, define the function $\psi_0(x) := \frac{1}{2}\|x - x_0\|^2$.
Step j: ($j \geq 0$) Find $a_{j+1} > 0$ from the equation

$$\frac{a_{j+1}^2}{\theta_j + a_{j+1}} = 2\frac{1+\alpha\theta_j}{L_f}.$$

Compute the following:

$$\begin{aligned} \theta_{j+1} &= \theta_j + a_{j+1}, \\ v_j &= \operatorname*{argmin}_x \ \psi_j(x), \qquad (6.2) \\ y_j &= \tfrac{\theta_j x_j + a_{j+1} v_j}{\theta_{j+1}}, \\ x_{j+1} &= \operatorname*{argmin}_x \ \{f(y_j) + \langle \nabla f(y_j), x - y_j\rangle + \tfrac{L_f}{2}\|x - y_j\|^2 + p(x)\}. \qquad (6.3) \end{aligned}$$

Define the function

$$\psi_{j+1}(x) = \psi_j(x) + a_{j+1}[f(x_{j+1}) + \langle \nabla f(x_{j+1}), x - x_{j+1}\rangle + p(x)]. \qquad (6.4)$$

Let us now make a few observations that we will call on shortly. First, each iteration of Algorithm 5 only requires two gradient computations, $\nabla f(y_j)$ in (6.3) and $\nabla f(x_{j+1})$ in (6.4), and two proximal operations, $\operatorname{prox}_{p/L_f}$ in (6.3) and prox_p in (6.2).

Secondly, let us translate the estimates in Theorem 6.1 to estimates based on desired accuracy. Namely, simple arithmetic shows that the inequality

$$f^p(x_j) - f^p(x^*) \leq \varepsilon$$

holds as soon as the number of iterations j satisfies

$$j \geq 1 + \sqrt{\frac{L_f}{2\alpha}} \cdot \log\left(\frac{L_f\|x^* - x_0\|^2}{4\varepsilon}\right). \qquad (6.5)$$

Let us now see how we can modify the scheme slightly so that it can find points x with small subgradients. Given a point x consider a single prox-gradient iteration $\hat{x} := \operatorname{prox}_{\frac{p}{L_f}}\left(x - \frac{1}{L_f}\nabla f(x)\right)$. Then we successively deduce

$$\operatorname{dist}^2(0; \partial f^p(\hat{x})) \leq 4\|L_f(\hat{x} - x)\|^2 \leq 8L_f(f^p(x) - f^p(\hat{x})) \leq 8L_f(f^p(x) - f^p(x^*))$$

where the first inequality is (4.1) and the second is the descent guarantee of the prox-gradient method (e.g. [54, Theorem 1]). Thus the inequality $f^p(x) - f^p(x^*) \leq \varepsilon^2/(8L_f)$ would immediately imply $\operatorname{dist}(0; \partial f^p(\hat{x})) \leq \varepsilon$. Therefore, let us add an extra prox-gradient step $\hat{x}_j := \operatorname{prox}_{\frac{p}{L_f}}\left(x_j - \frac{1}{L_f}\nabla f(x_j)\right)$ to each iteration of Algorithm 5. Appealing to the linear rate in (6.5), we then deduce that we can be sure of the inequality

$$\operatorname{dist}(0; \partial f^p(\hat{x}_j)) \leq \varepsilon$$

as soon as the number of iterations j satisfies

$$j \geq 1 + \sqrt{\frac{L_f}{2\alpha}} \cdot \log\left(\frac{2L_f^2\|x^* - x_0\|^2}{\varepsilon^2}\right). \tag{6.6}$$

With this modification, each iteration of the scheme requires two gradient evaluations of f and three proximal operations of p.

6.2 Total cost if h is smooth

In this section, we will assume that h is already C^1-smooth with the gradient having Lipschitz constant L_h, and calculate the overall cost of the inexact prox-linear method that wraps a linearly convergent method for the proximal subproblems. As we have discussed in Section 5, the proximal subproblems can either be approximately solved by primal methods or by dual methods. The dual methods are better adapted for a global analysis, since the dual problem has a bounded domain; therefore let us look first at that setting.

Remark 6.2 (Asymptotic notation)**.** To make clear dependence on the problem's data, we will sometimes use asymptotic notation [7, Section 3.5]. For two functions ψ and Ψ of a vector $\omega \in \mathbf{R}^\ell$, the symbol $\psi(\omega) = \mathcal{O}(\Psi(\omega))$ will mean that there exist constants $K, C > 0$ such that the inequality, $|\psi(\omega)| \leq C \cdot |\Psi(\omega)|$, holds for all ω satisfying $\omega_i \geq K$ for all $i = 1, \ldots, \ell$. When using asymptotic notation in this section, we will use the vector ω to encode the data of the problem $\omega = (\|\nabla c\|, L_h, L, \beta, F(x_0) - F^*, 1/\varepsilon)$. In the setting that h is not differentiable, L_h will be omitted from ω.

Total cost based on dual near-stationarity in the subproblems

We consider the near-stationarity model of inexactness as in Section 5.2. Namely, let us compute the total cost of Algorithm 4, when each subproblem $\min_w \varphi_k(w)$ is approximately minimized by the fast-gradient method (Algorithm 5). In the notation of Section 6.1, we set $f(w) = G_k^\star(A_k^* w) - \langle b_k, w\rangle$ and $p = h^\star$. By Lemma 5.3, the function f is C^1-smooth with gradient having Lipschitz constant $L_f := t\|\nabla c(x_k)\|_{\mathrm{op}}^2$. Since ∇h is assumed to be L_h-Lipschitz, we deduce that $h^\star$ is $\frac{1}{L_h}$-strongly convex. Notice moreover that since h is L-Lipschitz, any point in dom $h^\star$ is bounded in norm by L; hence the diameter of dom $h^\star$ is at most $2L$. Let us now apply Algorithm 5 (with the extra prox-gradient step) to the problem $\min_w \varphi_k(w) = f(w) + p(w)$. According to the estimate (6.6), we will be able to find the desired point w_{k+1} satisfying $\mathrm{dist}(0; \partial\varphi_k(w_{k+1}))) \leq \varepsilon_{k+1}$ after at most

$$1 + \left\lceil \sqrt{\frac{t\|\nabla c(x_k)\|_{\mathrm{op}}^2 L_h}{2}} \cdot \log\left(\frac{8t^2\|\nabla c(x_k)\|_{\mathrm{op}}^4 L^2}{\varepsilon_{k+1}^2}\right)\right\rceil. \tag{6.7}$$

iterations of the fast-gradient method. According to Lemma 5.3, each gradient evaluation ∇f requires two-matrix vector multiplications and one proximal operation of g, while the proximal operation of p amounts to a single proximal operation of h. Thus each iteration of Algorithm 5, with the extra prox-gradient step requires 9 basic operations. Finally to complete step k of Algorithm 5, we must take one extra proximal map of g. Hence the number of basic operations needed to complete step k of Algorithm 5 is $9\times$(equation (6.7))$+1$, where we set $t = 1/\mu$.

Let us now compute the total cost across the outer iterations k. Theorem 5.6 shows that if we set $\varepsilon_k = \frac{1}{Lk^2}$ in each iteration k of Algorithm 4, then after N outer iterations we are

guaranteed

$$\min_{j=0,\ldots,N-1} \left\| \mathcal{G}_{\frac{1}{\mu}}(x_j) \right\|^2 \leq \frac{4\mu\big(F(x_0) - F^* + 8\big)}{N}. \tag{6.8}$$

Thus we can find a point x satisfying

$$\left\| \mathcal{G}_{\frac{1}{\mu}}(x) \right\| \leq \varepsilon$$

after at most $\mathcal{N}(\varepsilon) := \left\lceil \frac{4\mu(F(x_0)-F^*+8)}{\varepsilon^2} \right\rceil$ outer-iterations and therefore after

$$\left\lceil \frac{4\mu(F(x_0) - F^* + 8)}{\varepsilon^2} \right\rceil \left(10 + 9 \left\lceil \sqrt{\frac{\|\nabla c\|^2 L_h}{2\mu}} \cdot \log\left(\frac{8\|\nabla c\|^4 L^2 (1+\mathcal{N}(\varepsilon))^4}{\beta^2} \right) \right\rceil \right) \tag{6.9}$$

basic operations in total. Thus the number of basic operations is on the order of

$$\boxed{\mathcal{O}\left(\frac{\sqrt{\|\nabla c\|^2 \cdot L_h \cdot \mu} \cdot (F(x_0) - F^*)}{\varepsilon^2} \log\left(\frac{\|\nabla c\|^2 L^3 \beta (F(x_0) - F^*)^2)}{\varepsilon^4} \right) \right)}. \tag{6.10}$$

Total cost based on approximate minimizers of the subproblems

Let us look at what goes wrong with applying Algorithm 2, with the proximal subproblems $\min_z F_t(z; x)$ approximately solved by a primal only method. To this end, notice that the objective function $F_t(\cdot; x)$ is a sum of the $\frac{1}{t}$-strongly convex and prox-friendly term $g + \frac{1}{2t}\|\cdot - x\|^2$ and the smooth convex function $z \mapsto h(c(x) + \nabla c(x)(z - x))$. The gradient of the smooth term is Lipschitz continuous with constant $\|\nabla c(x)\|_{\text{op}}^2 L_h$. Let us apply the fast gradient method (Algorithm 5) to the proximal subproblem directly. According to the estimate (6.5), Algorithm 5 will find an ε-approximate minimizer z of $F_t(\cdot; x)$ after at most

$$1 + \sqrt{\frac{t\|\nabla c(x)\|_{\text{op}}^2 L_h}{2}} \cdot \log\left(\frac{\|\nabla c(x)\|_{\text{op}}^2 L_h \|x^* - z_0\|^2}{4\varepsilon} \right) \tag{6.11}$$

iterations, where x^* is the minimizer of $F_t(\cdot; x)$ and the scheme is initialized at z_0. The difficulty is that there appears to be no simple way to bound the distance $\|z_0 - z^*\|$ for each proximal subproblem, unless we assume that dom g is bounded. We next show how we can correct for this difficulty by more carefully coupling the inexact prox-linear algorithm and the linearly convergent algorithm for solving the subproblem. In particular, in each outer iteration of the proposed scheme (Algorithm 6), one runs a linearly convergent subroutine $\mathcal{M}$ on the prox-linear subproblem for a fixed number of iterations; this fixed number of inner iterations depends explicitly on $\mathcal{M}$'s linear rate of convergence. The algorithmic idea behind this coupling originates in [41]. The most interesting consequence of this scheme is on so-called finite-sum problems, which we will discuss in Section 7. In this context, the algorithms that one runs on the proximal subproblems are stochastic. Consequently, we adapt our analysis to a stochastic setting as well, proving convergence rates on the expected norm of the prox-gradient $\|\mathcal{G}_t(x_k)\|$. When the proximal subproblems are approximately solved by deterministic methods, the convergence rates are all deterministic as well.

The following definition makes precise the types of algorithms that we will be able to accommodate as subroutines for the prox-linear subproblems.

Definition 6.3 (Linearly convergent subscheme)**.** A method $\mathcal{M}$ is *a linearly convergent subscheme* for the composite problem (3.1) if the following holds. For any points $x \in \mathbf{R}^d$, there exist constants $\gamma \geq 0$ and $\tau \in (0,1)$ so that when $\mathcal{M}$ is applied to $\min F_t(\cdot; x)$ with an arbitrary $z_0 \in \operatorname{dom} g$ as an initial iterate, $\mathcal{M}$ generates a sequence $\{z_i\}_{i=1}^{\infty}$ satisfying

$$\mathbb{E}[F_t(z_i; x) - F_t(x^*; x)] \leq \gamma\,(1-\tau)^i\,\|z_0 - x^*\|^2 \qquad \text{for } i = 1, \ldots, \infty, \tag{6.12}$$

where x^* is the minimizer of $F_t(\cdot; x)$.

We will be applying a linearly convergent subscheme to proximal subproblems $\min F_t(\cdot; x_k)$, where x_k is generated in the previous iteration of an inexact prox-linear method. We will then denote the resulting constants (γ, τ) in the guarantee (6.12) by (γ_k, τ_k).

The overall method we propose is Algorithm 6. It is important to note that in order to implement this method, one must know explicitly the constants (γ, τ) for the method $\mathcal{M}$ on each proximal subproblem.

Algorithm 6: Inexact prox-linear method: primal-only subsolves I

Initialize: A point $x_0 \in \operatorname{dom} g$, real $t > 0$, a linearly convergent subscheme $\mathcal{M}$ for (3.1).
Step k: $(k \geq 1)$
Set $x_{k,0} := x_k$. Initialize $\mathcal{M}$ on the problem $\min_z F_t(z; x_k)$ at $x_{k,0}$, and run $\mathcal{M}$ for

$$T_k := \left\lceil \frac{1}{\tau_k} \log(4t\gamma_k) \right\rceil \qquad \text{iterations,} \tag{6.13}$$

thereby generating iterates $x_{k,1}, \ldots, x_{k,T_k}$.
Set $x_{k+1} = x_{k,T_k}$.

The following lemma shows that the proposed number of inner iterations (6.13) leads to significant progress in the prox-linear subproblems, compared with the initialization. Henceforth, we let $\mathbb{E}_{x_k}[\cdot]$ denote the expectation of a quantity conditioned on the iterate x_k.

Lemma 6.4. *The iterates x_k generated by Algorithm 6 satisfy*

$$\mathbb{E}_{x_k}[F_t(x_{k+1}; x_k) - F_t(x_k^*; x_k)] \leq \frac{1}{4t}\,\|x_k - x_k^*\|^2\,. \tag{6.14}$$

Proof. In each iteration k, the linear convergence of algorithm $\mathcal{M}$ implies

$$\begin{aligned}\mathbb{E}_{x_k}[F_t(x_{k+1}; x_k) - F_t(x_k^*; x_k)] &\leq \gamma_k\,(1-\tau_k)^{T_k}\,\|x_{k,0} - x_k^*\|^2 \\ &\leq \gamma_k e^{-\tau_k T_k}\|x_k - x_k^*\|^2 \leq \frac{1}{4t}\|x_k - x_k^*\|^2,\end{aligned}$$

as claimed. □

With this lemma at hand, we can establish convergence guarantees of the inexact method.

Theorem 6.5 (Convergence of Algorithm 6)**.** *Supposing $t \leq \mu^{-1}$, the iterates x_k generated by Algorithm 6 satisfy*

$$\min_{j=0,\ldots,N-1} \mathbb{E}[\|\mathcal{G}_t(x_j)\|^2] \leq \frac{4t^{-1}\,(F(x_0) - \inf F)}{N}.$$

Proof. The proof follows the same outline as Theorem 5.2. Observe

$$\begin{aligned}\mathbb{E}_{x_k}[F(x_k) - F(x_{k+1})] &= \mathbb{E}_{x_k}[F_t(x_k; x_k) - F(x_{k+1})] \\ &\geq \mathbb{E}_{x_k}[F_t(x_k^*; x_k) - F(x_{k+1}) + \frac{1}{2t}\left\|x_k - x_k^*\right\|^2] \\ &\geq \mathbb{E}_{x_k}[F_t(x_k^*; x_k) - F_t(x_{k+1}; x_k)] + \frac{1}{2t}\left\|x_k - x_k^*\right\|^2 \\ &\geq -\frac{1}{4t}\left\|x_k - x_k^*\right\|^2 + \frac{1}{2t}\left\|x_k - x_k^*\right\|^2 \\ &\geq \frac{t}{4}\left\|\mathcal{G}_t(x_k)\right\|^2,\end{aligned}$$

where the second line follows from strong convexity of $F_t(\cdot; x_k)$, the third from Lemma 3.2, and the fourth from Lemma 6.4. Taking expectations of both sides, and using the tower rule, we deduce

$$\mathbb{E}[F(x_k) - F(x_{k+1})] \geq \frac{t}{4}\mathbb{E}[\|\mathcal{G}_t(x_k)\|^2].$$

Summing up both sides, we deduce

$$\frac{t}{4}\min_{j=0,\ldots,N-1}\mathbb{E}[\|\mathcal{G}_t(x_j)\|^2] \leq \frac{t}{4N}\sum_{j=0}^{N-1}\mathbb{E}[\|\mathcal{G}_t(x_j)\|^2] \leq \frac{1}{N}\sum_{j=0}^{N-1}\mathbb{E}[F(x_j) - F(x_{j+1})] \leq \frac{F(x_0) - \inf F}{N},$$

as claimed. □

It is clear from the proof that if the inner algorithm $\mathcal{M}$ satisfies (6.12) with the expectation $\mathbb{E}_{x_k}$ omitted, then Theorem 6.5 holds with $\mathbb{E}$ omitted as well and with $\inf F$ replaced by $F^* := \operatorname{liminf}_{k\to\infty} F(x_i)$. In particular, let us suppose that we set $t = \mu^{-1}$ and let $\mathcal{M}$ be the fast-gradient method (Algorithm 5) applied to the primal problem. Then in each iteration k, we can set $L_f = \|\nabla c(x_k)\|_{\text{op}}^2 L_h$ and $\alpha = \mu$. Let us now determine γ_k and τ_k. Using the inequality $(1+\sqrt{\frac{\alpha}{2L_f}})^{-1} \leq 1 - \sqrt{\frac{\alpha}{2L_f}}$ along with Theorem 6.1, we deduce we can set $\gamma_k = \frac{L_f}{4}$ and $\tau_k = \sqrt{\frac{\alpha}{2L_f}}$ for all indices k. Then each iteration of Algorithm 6 performs $T = \left\lceil \sqrt{\frac{2\|\nabla c(x_k)\|_{\text{op}}^2 L_h}{\mu}} \log\left(\|\nabla c(x_k)\|_{\text{op}}^2 L_h/\mu\right)\right\rceil$ iterations of the fast-gradient method, Algorithm 5. Recall that each iteration of Algorithm 5 requires 8 basic operations. Taking into account Theorem 6, we deduce that the overall scheme will produce a point x satisfying

$$\left\|\mathcal{G}_{\frac{1}{\mu}}(x)\right\| \leq \varepsilon$$

after at most

$$8\left\lceil \frac{4\mu\left(F(x_0) - F^*\right)}{\varepsilon^2}\right\rceil\left\lceil \sqrt{\frac{2\|\nabla c\|^2 L_h}{\mu}}\,\log\left(\frac{\|\nabla c\|^2 L_h}{\mu}\right)\right\rceil \tag{6.15}$$

basic operations. Thus the number of basic operations is on the order of

$$\boxed{\mathcal{O}\left(\frac{\sqrt{\|\nabla c\|^2 \cdot L_h \cdot \mu}\cdot\left(F(x_0) - F^*\right)}{\varepsilon^2}\,\log\left(\frac{\|\nabla c\|^2 L_h}{\mu}\right)\right)}. \tag{6.16}$$

Notice this estimate is better than (6.10), but only in terms of logarithmic dependence.

Before moving on, it is instructive to comment on the functional form of the linear convergence guarantee in (6.12). The right-hand-side depends on the initial squared distance $\|z_0 - x^*\|^2$. Convergence rates of numerous algorithms, on the other hand, are often stated with the right-hand-side instead depending on the initial functional error $F_t(z_0; x) - \inf_z F_t(z; x)$. In particular, this is the case for algorithms for finite sum problems discussed in Section 7, such as SVRG [35] and SAGA [18], and their accelerated extensions [1,29,40]. The following easy lemma shows how any such algorithm can be turned into a linearly convergent subscheme, in the sense of Definition 6.3, by taking a single extra prox-gradient step. We will use this observation in Section 7, when discussing finite-sum problems.

Lemma 6.6. *Consider an optimization problem having the convex additive composite form* (6.1). *Suppose $\mathcal{M}$ is an algorithm for $\min_z f^p(z)$ satisfying: there exist constants $\gamma \geq 0$ and $\tau \in (0, 1)$ so that on any input z_0, the method $\mathcal{M}$ generates a sequence $\{z_i\}_{i=1}^{\infty}$ satisfying*

$$\mathbb{E}[f^p(z_i) - f^p(z^*)] \leq \gamma (1 - \tau)^i (f^p(z_0) - f^p(z^*)) \qquad \text{for } i = 1, \ldots, \infty, \tag{6.17}$$

where z^* is a minimizer of f^p. Define an augmented method $\mathcal{M}^+$ as follows: given input z_0, initialize $\mathcal{M}$ at the point $\operatorname{prox}_{p/L_f}(z_0 - \frac{1}{L_f}\nabla f(z_0))$ and output the resulting points $\{z_i\}_{i=1}^{\infty}$. Then the iterates generates by $\mathcal{M}^+$ satisfy

$$\mathbb{E}[f^p(z_i) - f^p(z^*)] \leq \frac{\gamma L_f}{2} (1 - \tau)^i \|z_0 - z^*\|^2 \qquad \text{for } i = 1, \ldots, \infty,$$

Proof. Set $\hat{z} := \operatorname{prox}_{p/L_f}(z_0 - \frac{1}{L_f}\nabla f(z_0))$. Then convergence guarantees (6.17) of $\mathcal{M}$, with $\hat{z}$ in place of z_0, read

$$\mathbb{E}[f^p(z_i) - f^p(z^*)] \leq \gamma (1 - \tau)^i (f^p(\hat{z}) - f^p(z^*)) \qquad \text{for } i = 1, \ldots, \infty.$$

Observe the inequality $f^p(\hat{z}) \leq f(z_0) + \langle \nabla f(z_0), \hat{z} - z_0 \rangle + p(\hat{z}) + \frac{L_f}{2}\|\hat{z} - z_0\|^2$. By definition, $\hat{z}$ is the minimizer of the function $z \mapsto f(z_0) + \langle \nabla f(z_0), z - z_0 \rangle + p(z) + \frac{L_f}{2}\|z - z_0\|^2$, and hence we deduce $f^p(\hat{z}) \leq f(z_0) + \langle \nabla f(z_0), z^* - z_0 \rangle + p(z^*) + \frac{L_f}{2}\|z^* - z_0\|^2 \leq f^p(z^*) + \frac{L_f}{2}\|z^* - z_0\|^2$, with the last inequality follows from convexity of f. The result follows. □

6.3 Total cost of the smoothing strategy

The final ingredient is to replace h by a smooth approximation and then minimize the resulting composite function by an inexact prox-linear method (Algorithms 4 or 6). Define the smoothed composite function

$$F^\nu(x) := g(x) + h_\nu(c(x)), \tag{6.18}$$

where h_ν is the Moreau envelope of h. Recall from Lemma 2.1 the three key properties of the Moreau envelope:

$$\operatorname{lip}(h_\nu) \leq L, \qquad \operatorname{lip}(\nabla h_\nu) \leq \frac{1}{\nu},$$

and

$$0 \leq h(z) - h_\nu(z) \leq \frac{L^2 \nu}{2} \qquad \text{for all } z \in \mathbf{R}^m.$$

Indeed, these are the only properties of the smoothing we will use; therefore, in the analysis, any smoothing satisfying the analogous properties can be used instead of the Moreau envelope.

Let us next see how to choose the smoothing parameter $\nu > 0$ based on a target accuracy ε on the norm of the prox-gradient $\|\mathcal{G}_t(x)\|$. Naturally, we must establish a relationship between the step-sizes of the prox-linear steps on the original problem and its smooth approximation. To distinguish between these two settings, we will use the notation

$$\begin{aligned} x^+ &= \operatorname*{argmin}_z \left\{ h\big(c(x) + \nabla c(x)(z-x)\big) + g(z) + \tfrac{1}{2t}\|z-x\|^2 \right\}, \\ \widehat{x} &= \operatorname*{argmin}_z \left\{ h_\nu\big(c(x) + \nabla c(x)(z-x)\big) + g(z) + \tfrac{1}{2t}\|z-x\|^2 \right\}, \\ \mathcal{G}_t(x) &= t^{-1}(x^+ - x), \\ \mathcal{G}_t^\nu(x) &= t^{-1}(\widehat{x} - x). \end{aligned}$$

Thus $\mathcal{G}_t(x)$ is the prox-gradient on the target problem (3.1) as always, while $\mathcal{G}_t^\nu(x)$ is the prox-gradient on the smoothed problem (6.18). The following theorem will motivate our strategy for choosing the smoothing parameter ν.

Theorem 6.7 (Prox-gradient comparison)**.** *For any point x, the inequality holds:*

$$\|\mathcal{G}_t(x)\| \le \|\mathcal{G}_t^\nu(x)\| + \sqrt{\frac{L^2\nu}{2t}}.$$

Proof. Applying Lemma 2.1 and strong convexity of the proximal subproblems, we deduce

$$\begin{aligned} F_t(x^+;x) &\le F_t(\widehat{x};x) - \frac{1}{2t}\left\|\widehat{x} - x^+\right\|^2 \\ &\le \Big(h_\nu\big(c(x) + \nabla c(x)(\widehat{x}-x)\big) + g(\widehat{x}) + \frac{1}{2t}\|\widehat{x}-x\|^2\Big) + \frac{L^2\nu}{2} - \frac{1}{2t}\left\|\widehat{x} - x^+\right\|^2 \\ &\le \Big(h_\nu\big(c(x) + \nabla c(x)(x^+-x)\big) + g(x^+) + \frac{1}{2t}\left\|x^+ - x\right\|^2\Big) + \frac{L^2\nu}{2} - t^{-1}\left\|\widehat{x} - x^+\right\|^2 \\ &\le F_t(x^+;x) + \frac{L^2\nu}{2} - t^{-1}\left\|\widehat{x} - x^+\right\|^2. \end{aligned}$$

Canceling out like terms, we conclude $t^{-1}\|\widehat{x} - x^+\|^2 \le \frac{L^2\nu}{2}$. The triangle inequality then yields

$$t^{-1}\left\|x^+ - x\right\| \le t^{-1}\|\widehat{x} - x\| + \sqrt{\frac{L^2\nu}{2t}},$$

as claimed.‘ □

Fix a target accuracy $\varepsilon > 0$. The strategy for choosing the smoothing parameter ν is now clear. Let us set $t = \frac{1}{\mu}$ and then ensure $\frac{\varepsilon}{2} = \sqrt{\frac{L^2\nu}{2t}}$ by setting $\nu := \frac{\varepsilon^2}{2L^3\beta}$. Then by Theorem 6.7, any point x satisfying $\|\mathcal{G}_{1/\mu}^\nu(x)\| \le \frac{\varepsilon}{2}$ would automatically satisfy the desired condition $\|\mathcal{G}_{1/\mu}(x)\| \le \varepsilon$. Thus we must only estimate the cost of obtaining such a point x. Following the discussion in Section 6.2, we can apply either of the Algorithms 4 or 6, along with the fast-gradient method (Algorithm 5) for the inner subsolves, to the problem $\min_x F^\nu(x) = g(x) + h_\nu(c(x))$. We note that for a concrete implementation, one needs the following formulas, complementing Lemma 5.3.

Lemma 6.8. *For any point x and real ν, $t > 0$, the following are true:*

$$\operatorname{prox}_{th_\nu}(x) = \left(\tfrac{\nu}{t+\nu}\right)\cdot x + \left(\tfrac{t}{t+\nu}\right)\cdot \operatorname{prox}_{(t+\nu)h}(x) \qquad \textit{and} \qquad \nabla h_\nu(x) = \tfrac{1}{\nu}(x - \operatorname{prox}_{\nu h}(x)).$$

Proof. The expression $\nabla h_\nu(x) = \frac{1}{\nu}(x - \mathrm{prox}_{\nu h}(x))$ was already recorded in Lemma 2.1. Observe the chain of equalities

$$\min_y \left\{ h_\nu(y) + \frac{1}{2t}\|y-x\|^2 \right\} = \min_y \min_z \left\{ h(z) + \frac{1}{2\nu}\|z-y\|^2 + \frac{1}{2t}\|y-x\|^2 \right\} \tag{6.19}$$
$$= \min_z \left\{ h(z) + \frac{1}{2(t+\nu)}\|z-x\|^2 \right\},$$

where the last equality follows by exchanging the two mins in (6.19). By the same token, taking the derivative with respect to y in (6.19), we conclude that the optimal pair (y, z) must satisfy the equality $0 = \nu^{-1}(y-z) + t^{-1}(y-x)$. Since the optimal y is precisely $\mathrm{prox}_{th_\nu}(x)$ and the optimal z is given by $\mathrm{prox}_{(t+\nu)h}(x)$, the result follows. □

Let us apply Algorithm 4 with the fast-gradient dual subsolves, as described in Section 6.2. Appealing to (6.9) with $L_h = \frac{1}{\nu} = \frac{2L^3\beta}{\varepsilon^2}$ and ε replaced by $\varepsilon/2$, we deduce that the scheme will find a point x satisfying $\|\mathcal{G}_{1/\mu}(x)\| \le \varepsilon$ after at most

$$\mathcal{N}(\varepsilon) \cdot \left(10 + 9 \left\lceil \frac{\|\nabla c\| L}{\varepsilon} \cdot \log\left(\frac{8\|\nabla c\|^4 L^2 (1+\mathcal{N}(\varepsilon))^4}{\beta^2} \right) \right\rceil \right)$$

basic operations, where $\mathcal{N}(\varepsilon) := \left\lceil \frac{16\mu\left(F(x_0) - \inf F + 8 + \frac{\varepsilon^2}{4\mu}\right)}{\varepsilon^2} \right\rceil$. Hence the total cost is on the order[6] of

$$\boxed{\mathcal{O}\left(\frac{L^2\beta\|\nabla c\| \cdot (F(x_0) - \inf F)}{\varepsilon^3} \log\left(\frac{\|\nabla c\|^2 L^3 \beta (F(x_0) - \inf F)^2}{\varepsilon^4} \right) \right)}. \tag{6.20}$$

Similarly, let us apply Algorithm 6 with fast-gradient primal subsolves, as described in Section 6.2. Appealing to (6.15), we deduce that the scheme will find a point x satisfying $\|\mathcal{G}_{1/\mu}(x)\| \le \varepsilon$ after at most

$$8 \left\lceil \frac{16\mu\left(F(x_0) - \inf F + \frac{\varepsilon^2}{4\mu}\right)}{\varepsilon^2} \right\rceil \left\lceil \frac{2\|\nabla c\| L}{\varepsilon} \log\left(\frac{2\|\nabla c\|^2 L^2}{\varepsilon^2} \right) \right\rceil$$

basic operations. Thus the cost is on the order[6] of

$$\boxed{\mathcal{O}\left(\frac{L^2\beta\|\nabla c\| \cdot (F(x_0) - \inf F)}{\varepsilon^3} \log\left(\frac{\|\nabla c\| L}{\varepsilon} \right) \right)}. \tag{6.21}$$

Notice that the two estimates (6.20) and (6.21) are identical up to a logarithmic dependence on the problem data. To the best of our knowledge, these are the best-known efficiency estimates of any first-order method for the composite problem class (3.1).

The logarithmic dependence in the estimates (6.20) and (6.21) can be removed entirely, by a different technique, provided we have available an accurate estimate on $F(x_0) - \inf F$ and an a priori known estimate $\|\nabla c\|$ to be used throughout the procedure. Since we feel that the resulting scheme is less practical than the ones outlined in the current section, we have placed the details in Appendix C.

[6] Here, we use the asymptotic notation described in Remark 6.2 with $\omega = (\|\nabla c\|, L, \beta, F(x_0) - \inf F, 1/\varepsilon)$.

7 Finite sum problems

In this section, we extend the results of the previous sections to so-called "finite sum problems", also often called "regularized empirical risk minimization". More precisely, throughout the section instead of minimizing a single composite function, we will be interested in minimizing an average of m composite functions:

$$\min_x \; F(x) := \frac{1}{m}\sum_{i=1}^m h_i(c_i(x)) + g(x) \tag{7.1}$$

In line with the previous sections, we make the following assumptions on the components of the problem:

1. g is a closed convex function;

2. $h_i \colon \mathbf{R} \to \mathbf{R}$ are convex, and L-Lipschitz continuous;

3. $c_i \colon \mathbf{R}^d \to \mathbf{R}$ are C^1-smooth with the gradient map ∇c_i that is β-Lipschitz continuous.

We also assume that we have available a real value, denoted $\|\nabla c\|$, satisfying

$$\|\nabla c\| \geq \sup_{x \in \operatorname{dom} g} \max_{i=1,\dots,m} \|\nabla c_i(x)\|.$$

The main conceptual premise here is that m is large and should be treated as an explicit parameter of the problem. Moreover, notice the Lipschitz data is stated for the individual functional components of the problem. Such finite-sum problems are ubiquitous in machine learning and data science, where m is typically the (large) number of recorded measurements of the system. Notice that we have assumed that c_i maps to the real line. This is purely for notational convenience. Completely analogous results, as in this section, hold when c_i maps into a higher dimensional space.

Clearly, the finite-sum problem (7.1) is an instance of the composite problem class (3.1) under the identification

$$h(z_i, \dots, z_m) := \frac{1}{m}\sum_{i=1}^m h_i(z_i) \qquad \text{and} \qquad c(x) := (c_1(x), \dots, c_m(x)). \tag{7.2}$$

Therefore, given a target accuracy $\varepsilon > 0$, we again seek to find a point x with a small prox-gradient $\|\mathcal{G}_t(x)\| \leq \varepsilon$. In contrast to the previous sections, by a *basic operation* we will mean individual evaluations of $c_i(x)$ and $\nabla c_i(x)$, dot-products $\nabla c_i(x)^T v$, and proximal operations $\operatorname{prox}_{th_i}$ and prox_{tg}.

Let us next establish baseline efficiency estimates by simply using the inexact prox-linear schemes discussed in Sections 6.2 and 6.3. To this end, the following lemma derives Lipschitz constants of h and ∇c from the problem data L and β. The proof is elementary and we have placed it in Appendix A. Henceforth, we set $\operatorname{lip}(\nabla c) := \sup_{x \neq y} \frac{\|\nabla c(x) - \nabla c(y)\|_{\mathrm{op}}}{\|x-y\|}$.

Lemma 7.1 (Norm comparison)**.** *The inequalities hold:*

$$\operatorname{lip}(h) \leq L/\sqrt{m}, \qquad \operatorname{lip}(\nabla c) \leq \beta\sqrt{m}, \qquad \|\nabla c(x)\|_{op} \leq \sqrt{m}\left(\max_{i=1,\dots,m} \|\nabla c_i(x)\|\right) \quad \forall x.$$

If in addition each h_i is C^1-smooth with L_h-Lipschitz derivative $t \mapsto h_i'(t)$, then the inequality, $\operatorname{lip}(\nabla h) \leq L_h/m$, holds as well.

Remark 7.2 (Notational substitution)**.** We will now apply the results of the previous sections to the finite sum problem (7.1) with h and c defined in (7.2). In order to correctly interpret results from the previous sections, according to Lemma 7.1, we must be mindful to replace L with $L/\sqrt{m}$, β with $\beta\sqrt{m}$, $\|\nabla c\|$ with $\sqrt{m}\|\nabla c\|$, and L_h with L_h/m. In particular, observe that we are justified in setting $\mu := L\beta$ without any ambiguity. Henceforth, we will be using this substitution routinely.

Baseline efficiency when h_i are smooth:

Let us first suppose that h_i are C^1-smooth with L_h-Lipschitz derivative and interpret the efficiency estimate (6.16). Notice that each gradient evaluation ∇c requires m individual gradient evaluations ∇c_i. Thus multiplying (6.16) by m and using Remark 7.2, the efficiency estimate (6.16) reads:

$$\boxed{\mathcal{O}\left(\frac{m\sqrt{\|\nabla c\|^2 \cdot L_h \cdot L \cdot \beta}\cdot (F(x_0)-\inf F)}{\varepsilon^2}\ \log\left(\frac{\|\nabla c\|^2 L_h}{L\beta}\right)\right)} \tag{7.3}$$

basic operations.

Baseline efficiency when h_i are nonsmooth:

Now let us apply the smoothing technique described in Section 6.3. Multiplying the efficiency estimate (6.21) by m and using Remark 7.2 yields:

$$\boxed{\mathcal{O}\left(\frac{m\cdot L^2\beta\|\nabla c\|\cdot (F(x_0)-\inf F)}{\varepsilon^3}\ \log\left(\frac{\|\nabla c\|L}{\varepsilon}\right)\right)} \tag{7.4}$$

basic operations.

The two displays (7.3) and (7.4) serve as baseline efficiency estimates for obtaining a point x satisfying $\|\mathcal{G}_{1/\mu}(x)\| \le \varepsilon$. We will now see that one can improve these guarantees in expectation. The strategy is perfectly in line with the theme of the paper. We will replace h by a smooth approximation, then apply an inexact prox-linear Algorithm 6, while approximately solving each subproblem by an "(accelerated) incremental method". Thus the only novelty here is a different scheme for approximately solving the proximal subproblems.

7.1 An interlude: incremental algorithms

There are a number of popular algorithms for finite-sum problems, including SAG [63], SAGA [19], SDCA [64], SVRG [35,71], FINITO [20], and MISO [42]. All of these methods have similar linear rates of convergence, and differ only in storage requirements and in whether one needs to know explicitly the strong convexity constant. For the sake of concreteness, we will focus on SVRG following [71]. This scheme applies to finite-sum problems

$$\min_x\ f^p(x) := \frac{1}{m}\sum_{i=1}^m f_i(x) + p(x), \tag{7.5}$$

where p is a closed, α-strongly convex function ($\alpha > 0$) and each f_i is convex and C^1-smooth with ℓ-Lipschitz gradient ∇f_i. For notational convenience, define the condition number $\kappa := l/\alpha$. Observe that when each h_i is smooth, each proximal subproblem indeed has this form:

$$\min_z \; F_t(z;x) := \frac{1}{m}\sum_{i=1}^m h_i\Big(c_i(x) + \langle \nabla c_i(x), z - x\rangle\Big) + g(z) + \frac{1}{2t}\|z-x\|^2. \tag{7.6}$$

In Algorithm 7, we record the Prox-SVRG method of [71] for minimizing the function (7.5).

Algorithm 7: The Prox-SVRG method [71]

Initialize: A point $\widetilde{x}_0 \in \mathbf{R}^d$, a real $\eta > 0$, a positive integer J.
Step s: ($s \geq 1$)
$\widetilde{x} = \widetilde{x}_{s-1}$;
$\widetilde{v} = \frac{1}{m}\sum_{i=1}^m \nabla f_i(\widetilde{x})$;
$x_0 = \widetilde{x}$
for $j = 1, 2, \ldots, J$ **do**
  pick $i_j \in \{1, \ldots, m\}$ uniformly at random
  $v_j = \widetilde{v} + (\nabla f_{i_j}(x_{j-1}) - \nabla f_{i_j}(\widetilde{x}))$
  $x_j = \mathrm{prox}_{\eta p}(x_{j-1} - \eta v_j)$
end
$\widetilde{x}_s = \frac{1}{J}\sum_{j=1}^J x_j$

The following theorem from [71, Theorem 3.1] summarizes convergence guarantees of Prox-SVRG.

Theorem 7.3 (Convergence rate of Prox-SVRG)**.** *Algorithm 7, with the choices $\eta = \frac{1}{10\ell}$ and $J = \lceil 100\kappa \rceil$, will generate a sequence $\{\widetilde{x}_s\}_{s\geq 1}$ satisfying*

$$\mathbb{E}[f^p(\widetilde{x}_s) - f^p(x^*)] \leq 0.9^s (f^p(\widetilde{x}_0) - f^p(x^*)),$$

where x^* is the minimizer of f^p. Moreover, each step s requires $m + 2\lceil 100\kappa \rceil$ individual gradient ∇f_i evaluations.

Thus Prox-SVRG will generate a point x with $\mathbb{E}[f^p(x) - f^p(x^*)] \leq \varepsilon$ after at most

$$\mathcal{O}\left((m+\kappa)\;\log\left(\frac{f^p(\widetilde{x}_0) - f^p(x^*)}{\varepsilon}\right)\right) \tag{7.7}$$

individual gradient ∇f_i evaluations. It was a long-standing open question whether there is a method that improves the dependence of this estimate on the condition number κ. This question was answered positively by a number of algorithms, including Catalyst [40], accelerated SDCA [65], APPA [29], RPDG [36], and Katyusha [1]. For the sake of concreteness, we focus only on one of these methods, Katyusha [1]. This scheme follows the same epoch structure as SVRG, while incorporating iterate history. We summarize convergence guarantees of this method, established in [1, Theorem 3.1], in the following theorem.

Theorem 7.4 (Convergence rate of Katyusha)**.** *The Katyusha algorithm of [1] generates a sequence of iterates* $\{\widetilde{x}_s\}_{s\geq 1}$ *satisfying*

$$\frac{\mathbb{E}[f^p(\widetilde{x}_s) - f^p(x^*)]}{f^p(\widetilde{x}_0) - f^p(x^*)} \leq \begin{cases} 4 \cdot \left(1 + \sqrt{1/(6\kappa m)}\right)^{-2sm} & , \quad \text{if } \frac{m}{\kappa} \leq \frac{3}{8} \\ 3 \cdot (1.5)^{-s} & , \quad \text{if } \frac{m}{\kappa} > \frac{3}{8} \end{cases}$$

where x^* *is the minimizer of* f^p*. Moreover, each step* s *requires* $3m$ *individual gradient* ∇f_i *evaluations.*[7]

To simplify the expression for the rate, using the inequality $(1+z)^m \geq 1 + mz$ observe

$$\left(1 + \sqrt{\tfrac{1}{6\kappa m}}\right)^{-2sm} \leq \left(1 + \sqrt{\tfrac{2m}{3\kappa}}\right)^{-s}.$$

Using this estimate in Theorem 7.4 simplifies the linear rate to

$$\frac{\mathbb{E}[f^p(\widetilde{x}_s) - f^p(x^*)]}{f^p(\widetilde{x}_0) - f^p(x^*)} \leq 4 \cdot \max\left\{\left(1 + \sqrt{\tfrac{2m}{3\kappa}}\right)^{-s}, 1.5^{-s}\right\}.$$

Recall that each iteration of Katyusha requires $3m$ individual gradient ∇f_i evaluations. Thus the method will generate a point x with $\mathbb{E}[f^p(x) - f^p(x^*)] \leq \varepsilon$ after at most

$$\mathcal{O}\left(\left(m + \sqrt{m\kappa}\,\right) \ \log\left(\frac{f^p(\widetilde{x}_0) - f^p(x^*)}{\varepsilon}\right)\right)$$

individual gradient ∇f_i evaluations. Notice this efficiency estimate is significantly better than the guarantee (7.7) for Prox-SVRG only when $m \ll \kappa$. This setting is very meaningful in the context of smoothing. Indeed, since we will be applying accelerated incremental methods to proximal subproblems after a smoothing, the condition number κ of each subproblem can be huge.

Improved efficiency estimates when h_i are smooth:

Let us now suppose that each h_i is C^1-smooth with L_h-Lipschitz derivative h_i'. We seek to determine the efficiency of the inexact prox-linear method (Algorithm 6) that uses either Prox-SVRG or Katyusha as the linearly convergent subscheme $\mathcal{M}$. Let us therefore first look at the efficiency of Prox-SVRG and Katyusha on the prox-linear subproblem (7.6). Clearly we can set

$$\ell := L_h \cdot \left(\max_{i=1,\ldots,m} \|\nabla c_i(x)\|^2\right) \qquad \text{and} \qquad \alpha = t^{-1}.$$

Notice that the convergence guarantees for Prox-SVRG and Katyusha are not in the standard form (6.12). Lemma 6.6, however, shows that they can be put into standard form by taking a single extra prox-gradient step in the very beginning of each scheme; we'll call these slightly modified schemes Prox-SVRG$^+$ and Katyusha$^+$. Taking into account Lemma 7.1, observe that

[7] The constants 4 and 3 are hidden in the $\mathcal{O}$ notation in [1, Theorem 3.1]. They can be explicitly verified by following along the proof.

the gradient of the function $z \mapsto h(c(x)+\nabla c(x)(z-x))$ is l-Lipschitz continuous. Thus according to Lemma 6.6, Prox-SVRG$^+$ and Katyusha$^+$ on input $\widetilde{z}_0$ satisfy

$$\mathbb{E}[F_t(\widetilde{z}_s;x) - F_t(z^*;x)] \leq \frac{\ell}{2} \cdot 0.9^s \|\widetilde{z}_0 - z^*\|^2,$$
$$\mathbb{E}[F_t(\widetilde{z}_s;x) - F_t(z^*;x)] \leq \frac{4\ell}{2} \cdot \max\left\{\left(1+\sqrt{\tfrac{2m}{3\kappa}}\right)^{-s}, 1.5^{-s}\right\} \cdot \|\widetilde{z}_0 - z^*\|^2,$$

for $s = 1, \ldots, \infty$, respectively, where z^* is the minimizer of $F_t(\cdot;x)$.

We are now ready to compute the total efficiency guarantees. Setting $t = 1/\mu$, Theorem 6.5 shows that Algorithm 6 will generate a point x with

$$\mathbb{E}\left[\|\mathcal{G}_{1/\mu}(x)\|^2\right] \leq \varepsilon^2$$

after at most $\left\lceil \frac{4\mu(F(x_0)-\inf F)}{\varepsilon^2} \right\rceil$ iterations. Each iteration k in turn requires at most

$$\left\lceil \frac{1}{\tau_k}\log(4t\gamma_k) \right\rceil \leq \left\lceil \frac{1}{0.1}\log\left(4 \cdot \frac{1}{\mu} \cdot \frac{L_h \cdot \|\nabla c\|^2}{2}\right) \right\rceil$$

iterations of Prox-SVRG$^+$ and at most

$$\left\lceil \frac{1}{\tau_k}\log(4t\gamma_k) \right\rceil \leq \left\lceil \max\left\{3, \left(1+\sqrt{\tfrac{3L_h\|\nabla c\|^2}{2m\mu}}\right)\right\} \log\left(4 \cdot \frac{1}{\mu} \cdot \frac{4 \cdot L_h \cdot \|\nabla c\|^2}{2}\right) \right\rceil$$

iterations of Katyusha$^+$. Finally recall that each iteration s of Prox-SVRG$^+$ and of Katyusha$^+$, respectively, requires $m + 2\left\lceil \frac{100L_h\|\nabla c\|^2}{\mu} \right\rceil$ and $3m$ evaluations of $\nabla c_i(x)^T v$. Hence the overall efficiency is on the order of

$$\boxed{\mathcal{O}\left(\frac{\left(\mu m + L_h\|\nabla c\|^2\right)\cdot(F(x_0)-\inf F)}{\varepsilon^2} \log\left(\frac{L_h \cdot \|\nabla c\|^2}{\mu}\right)\right)} \tag{7.8}$$

when using Prox-SVRG$^+$ and on the order of

$$\boxed{\mathcal{O}\left(\frac{\left(\mu m + \sqrt{\mu m L_h\|\nabla c\|^2}\right)\cdot(F(x_0)-\inf F)}{\varepsilon^2} \log\left(\frac{L_h \cdot \|\nabla c\|^2}{\mu}\right)\right)} \tag{7.9}$$

when using Katyusha$^+$. Notice that the estimate (7.9) is better than (7.8) precisely when $m \ll \frac{L_h\|\nabla c\|^2}{\mu}$.

Improved efficiency estimates when h_i are nonsmooth:

Finally, let us now no longer suppose that h_i are smooth in the finite-sum problem (7.1) and instead apply the smoothing technique. To this end, observe the equality

$$h_\nu(z) = \inf_y \left\{\frac{1}{m}\sum_{i=1}^m h_i(y_i) + \frac{1}{2\nu}\|y-z\|^2\right\} = \sum_{i=1}^m (h_i/m)_\nu(z_i).$$

Therefore the smoothed problem in (6.18) is also a finite-sum problem with

$$\min_x \ \frac{1}{m}\sum_{i=1}^{m} m\cdot(h_i/m)_\nu(c_i(x)) + g(x).$$

Thus we can can apply the convergence estimates we have just derived in the smooth setting with $h_i(t)$ replaced by $\phi_i(t) := m\cdot(h_i/m)_\nu(t)$. Observe that ϕ_i is L-Lipschitz by Lemma 2.1, while the derivative $\phi_i'(t) = m\cdot\nu^{-1}(t - \operatorname{prox}_{\frac{\nu}{m}h_i}(t))$ is Lipschitz with constant $L_h := \frac{m}{\nu}$. Thus according to the recipe following Theorem 6.7, given a target accuracy $\varepsilon > 0$ for the norm of the prox-gradient $\|\mathcal{G}_{\frac{1}{\mu}}(x)\|$, we should set

$$\nu := \frac{m\varepsilon^2}{2L^3\beta},$$

where we have used the substitutions dictated by Remark 7.2. Then Theorem 6.7 implies

$$\left\|\mathcal{G}_{1/\mu}(x)\right\| \le \left\|\mathcal{G}^\nu_{1/\mu}(x)\right\| + \frac{\varepsilon}{2} \qquad \text{for all } x,$$

where $\left\|\mathcal{G}^\nu_{1/\mu}(x)\right\|$ is the prox-gradient for the smoothed problem. Squaring and taking expectations on both sides, we can be sure $\mathbb{E}[\left\|\mathcal{G}_{1/\mu}(x)\right\|^2] \le \varepsilon^2$ if we find a point x satisfying $\mathbb{E}\left[\left\|\mathcal{G}^\nu_{1/\mu}(x)\right\|^2\right] \le \frac{\varepsilon^2}{4}$. Thus we must simply write the estimates (7.8) and (7.9) for the smoothed problem in terms of the original problem data. Thus to obtain a point x satisfying

$$\mathbb{E}[\left\|\mathcal{G}_{1/\mu}(x)\right\|^2] \le \varepsilon^2,$$

it suffices to perform

$$\boxed{\mathcal{O}\left(\left(\frac{L\beta m}{\varepsilon^2} + \frac{L^2\beta\|\nabla c\|}{\varepsilon^3}\cdot\min\left\{\sqrt{m}, \frac{L\|\nabla c\|}{\varepsilon}\right\}\right)\cdot(F(x_0) - \inf F)\ \log\left(\frac{L\cdot\|\nabla c\|}{\varepsilon}\right)\right)} \tag{7.10}$$

basic operations. The min in the estimate corresponds to choosing the better of the two, Prox-SVRG$^+$ and Katyusha$^+$, in each proximal subproblem in terms of their efficiency estimates. Notice that the $1/\varepsilon^3$ term in (7.10) scales only as $\sqrt{m}$. Therefore this estimate is an order of magnitude better than our baseline (7.4), which we were trying to improve. The caveat is of course that the estimate (7.10) is in expectation while (7.4) is deterministic.

8 An accelerated prox-linear algorithm

Most of the paper thus far has focused on the setting when the proximal subproblems (1) can only be approximately solved by first-order methods. On the other hand, in a variety of circumstances, it is reasonable to expect to solve the subproblems to a high accuracy by other means. For example, one may have available specialized methods for the proximal subproblems, or interior-point points methods may be available for moderate dimensions d and m, or it may be that case that computing an accurate estimate of $\nabla c(x)$ may already be the bottleneck (see e.g. Example 3.5). In this context, it is interesting to see if the basic prox-linear method can in some sense be "accelerated" by using inertial information. In this section, we do exactly that.

We propose an algorithm, motivated by the work of Ghadimi-Lan [30], that is adaptive to some natural constants measuring convexity of the composite function. This being said, the reader should keep in mind a downside the proposed scheme: our analysis (for the first time in the paper) requires the domain of g to be bounded. Henceforth, define

$$M := \sup_{x,y \in \operatorname{dom} g} \|x - y\|$$

and assume it to be finite.

To motivate the algorithm, let us first consider the additive composite setting (3.2) with $c(\cdot)$ in addition convex. Algorithms in the style of Nesterov's second accelerated method (see [51] or [67, Algorithm 1]) incorporate steps of the form $v_{k+1} = \operatorname{prox}_{tg}(v_k - t\nabla c(y_k))$. That is, one moves from a point v_k in the direction of the negative gradient $-\nabla c(y_k)$ evaluated at a different point y_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 [30, 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

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

The scheme we propose is summarized in Algorithm 8.

Algorithm 8: 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

$$a_k = \tfrac{2}{k+1} \tag{8.1}$$

$$y_k = a_k v_{k-1} + (1 - a_k) x_{k-1} \tag{8.2}$$

$$x_k = S_{1/\tilde{\mu}}(y_k) \tag{8.3}$$

$$v_k = S_{\frac{1}{\tilde{\mu} a_k}, a_k}(y_k, v_{k-1}) \tag{8.4}$$

Remark 8.1 (Interpolation weights)**.** When L and β are unknown, one can instead equip Algorithm 8 with a backtracking line search. A formal description and the resulting convergence guarantees appear in Appendix B. We also note that instead of setting $a_k = \frac{2}{k+1}$, one may use the interpolation weights used in FISTA [4]; namely, the sequence a_k may be chosen to satisfy the relation $\frac{1-a_k}{a_k^2} = \frac{1}{a_{k-1}^2}$, with similar convergence guarantees.

8.1 Convergence guarantees and convexity moduli

We will see momentarily that convergence guarantees of Algorithm 8 are adaptive to convexity (or lack thereof) of the composition $h \circ c$. To simplify notation, henceforth set

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

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 8. The first is weak-convexity. Recall from Lemma 4.2 that Φ is ρ-weakly convex for some $\rho \in [0, \mu]$. Thus there is some $\rho \in [0, \mu]$ such that for any points $x, y \in \mathbf{R}^d$ and $a \in [0, 1]$, the approximate secant inequality holds:

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

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 also play a role.

Definition 8.2 (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 (3.5) show that the pair (h, c) indeed has a convexity constant $r \in [0, \mu]$ on $\mathbf{R}^d$. The following relationship between convexity of the pair (h, c) and weak convexity of Φ will be useful.

Lemma 8.3 (Convexity of the pair implies weak 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 [17, Theorem 3.1]. We omit the details. □

Remark 8.4. 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.

Convergence guarantees

Henceforth, let ρ be a weak convexity constant of $h \circ c$ on dom g and let r be a convexity constant of (h, c) on dom g. According to Lemma 8.3, we can always assume $0 \le \rho \le r \le \mu$. We are now ready to state and prove convergence guarantees of Algorithm 8.

Theorem 8.5 (Convergence guarantees)**.** *Fix a real number $\tilde{\mu} > \mu$ and let x^* be any point satisfying $F(x^*) \le F(x_k)$ for all iterates x_k generated by Algorithm 8. Then the efficiency estimate holds:*

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

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:

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

Succinctly, setting $\tilde{\mu} := 2\mu$, Theorem 8.5 guarantees the bound

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

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

Our proof of Theorem 8.5 is based on two basic lemmas, as is common for accelerated methods [67].

Lemma 8.6 (Three-point comparison)**.** *Consider the point $z := S_{t,\alpha}(y, v)$ for some points $y, v \in \mathbf{R}^d$ and real numbers $t, \alpha > 0$. Then for all $w \in \mathbf{R}^d$ the inequality holds:*

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

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 8.7 (Telescoping)**.** *Let a_k, y_k, x_k, and v_k be the iterates generated by Algorithm 8. Then for any point $x \in \mathbf{R}^d$ and any index k, the inequality holds:*

$$\begin{aligned} F(x_k) \le a_k F(x) + (1 - a_k) F(x_{k-1}) + \frac{\tilde{\mu} a_k^2}{2} (\|x - v_{k-1}\|^2 - \|x - v_k\|^2) \\ - \frac{\tilde{\mu} - \mu}{2} \|y_k - x_k\|^2 + \rho a_k \|x - x_{k-1}\|^2 + \frac{r a_k^2}{2} \|x - v_{k-1}\|^2. \end{aligned} \tag{8.5}$$

Proof. Notice that all the points x_k, y_k, and v_k lie in dom g. From inequality (3.5), we have

$$F(x_k) \leq h\big(c(y_k) + \nabla c(y_k)(x_k - y_k)\big) + g(x_k) + \frac{\mu}{2}\|x_k - y_k\|^2. \tag{8.6}$$

Define the point $w_k := a_k v_k + (1 - a_k)x_{k-1}$. Applying Lemma 8.6 to $x_k = S_{1/\tilde{\mu},1}(y_k, y_k)$ with $w = w_k$ yields the inequality

$$\begin{aligned} h(c(y_k) + \nabla c(y_k)(x_k - y_k)) + g(x_k) \leq\ & h(c(y_k) + \nabla c(y_k)(w_k - y_k)) \\ & + \frac{\tilde{\mu}}{2}(\|w_k - y_k\|^2 - \|w_k - x_k\|^2 - \|x_k - y_k\|^2) \\ & + a_k g(v_k) + (1 - a_k)g(x_{k-1}). \end{aligned} \tag{8.7}$$

Note the equality $w_k - y_k = a_k(v_k - v_{k-1})$. Applying Lemma 8.6 again with $v_k = S_{\frac{1}{\tilde{\mu}a_k},a_k}(y_k, v_{k-1})$ and $w = x$ yields

$$\begin{aligned} h\big(c(y_k) + a_k\nabla c(y_k)(v_k - v_{k-1})\big) + a_k g(v_k) \leq\ & h\big(c(y_k) + a_k\nabla c(y_k)(x - v_{k-1})\big) + a_k g(x) \\ & + \frac{\tilde{\mu}a_k^2}{2}\left(\|x - v_{k-1}\|^2 - \|x - v_k\|^2 - \|v_k - v_{k-1}\|^2\right). \end{aligned} \tag{8.8}$$

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

$$\begin{aligned} h(c(y_k) + \nabla c(y_k)(\hat{x} - y_k)) &\leq (h \circ c)(\hat{x}) + \frac{r}{2}\|\hat{x} - y_k\|^2 \\ &\leq a_k h(c(x)) + (1 - a_k)h(c(x_{k-1})) \\ &\quad + \rho a_k(1 - a_k)\|x - x_{k-1}\|^2 + \frac{ra_k^2}{2}\|x - v_{k-1}\|^2. \end{aligned} \tag{8.9}$$

Thus combining inequalities (8.6), (8.7), (8.8), and (8.9), and upper bounding $1 - a_k \leq 1$ and $-\|w_k - x_k\|^2 \leq 0$, we obtain

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

The proof is complete. □

The proof of Theorem 8.5 now quickly follows.

Proof of Theorem 8.5. Set $x = x^*$ in inequality (8.5). Rewriting (8.5) by subtracting $F(x^*)$ from both sides, we obtain

$$\begin{aligned} \frac{F(x_k) - F(x^*)}{a_k^2} + \frac{\tilde{\mu}}{2}\|x^* - v_k\|^2 \leq\ & \frac{1 - a_k}{a_k^2}\big(F(x_{k-1}) - F(x^*)\big) + \frac{\tilde{\mu}}{2}\|x^* - v_{k-1}\|^2 \\ & + \frac{\rho M^2}{a_k} + \frac{rM^2}{2} - \frac{\tilde{\mu} - \mu}{2a_k^2}\|x_k - y_k\|^2. \end{aligned} \tag{8.10}$$

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

$$\frac{F(x_N) - F(x^*)}{a_N^2} + \frac{\tilde{\mu}}{2}\|x^* - v_N\|^2 \leq \frac{1-a_1}{a_1^2}\big(F(x_0) - F(x^*)\big) + \frac{\tilde{\mu}}{2}\|x^* - v_0\|^2 + \rho M^2 \left(\sum_{j=1}^N \frac{1}{a_j}\right) + \frac{NrM^2}{2} - \frac{\tilde{\mu}-\mu}{2}\sum_{j=1}^N \frac{\|x_j - y_j\|^2}{a_j^2}. \tag{8.11}$$

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

$$\frac{\tilde{\mu}-\mu}{2}\sum_{j=1}^N \frac{\|x_j - y_j\|^2}{a_j^2} \leq \frac{\tilde{\mu}}{2}\|x^* - v_0\|^2 + \rho M^2 \left(\sum_{j=1}^N \frac{1}{a_j}\right) + \frac{NrM^2}{2} \tag{8.12}$$

and hence

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

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

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

thereby establishing the first claimed efficiency estimate in Theorem 8.5.

Finally suppose $r = 0$, and hence we can assume $\rho = 0$ by Lemma 8.3. Inequality (8.11) 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. □

8.2 Inexact computation

Completely analogously, we can consider an inexact accelerated prox-linear method based on approximately solving the duals of the prox-linear subproblems (Algorithm 9).

Algorithm 9: 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

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

- Find (x_k, ζ_k) such that $\|\zeta_k\| \leq \varepsilon_k$ and x_k is the minimizer of the function

$$z \mapsto g(z) + h\Big(\zeta_k + c(y_k) + \nabla c(y_k)(z - y_k)\Big) + \frac{\tilde{\mu}}{2}\|z - y_k\|^2. \tag{8.13}$$

- Find (v_k, ξ_k) such that $\|\xi_k\| \leq \delta_k$ and v_k is the minimizer of the function

$$v \mapsto g(v) + \frac{1}{a_k} h\Big(\xi_k + c(y_k) + a_k \nabla c(y_k)(v - v_{k-1})\Big) + \frac{\tilde{\mu} a_k}{2}\|v - v_{k-1}\|^2. \tag{8.14}$$

Theorem 8.8 (Convergence of inexact accelerated prox-linear method: near- stationarity)**.** *Fix a real number $\tilde{\mu} \geq \mu$ and let x^* be any point satisfying $F(x^*) \leq F(x_k)$ for iterates x_k generated by Algorithm 9. Then for any $N \geq 1$, the iterates x_k 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:

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

The proof appears in Appendix A. Thus to preserve the rate in N of the exact accelerated prox-linear method in Theorem 8.5, it suffices to 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 $q > 0$.

Similarly, we can consider an inexact version of the accelerated prox-linear method based on approximately solving the primal problems in function value. The scheme is recorded in Algorithm 10.

Theorem 8.9 presents convergence guarantees of Algorithm 10. The statement of Theorem 8.9 is much more cumbersome than the analogous Theorem 8.8. The only take-away message for the reader is that to preserve the rate of the exact accelerated prox-linear method in Theorem 8.5 in terms of N, it sufficies for the sequences $\{\sqrt{i\delta_i}\}$, $\{i\delta_i\}$, and $\{i^2\varepsilon_i\}$ to be summable. Thus it suffices to take $\varepsilon_i, \delta_i \sim \frac{1}{i^{3+q}}$ for some $q > 0$.

The proof of Theorem 8.9 appears in Appendix A. Analysis of inexact accelerated methods of this type for additive convex composite problems has appeared in a variety of papers, including

Algorithm 10: 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 $i = 1, 2, \ldots, \infty$.

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

$$a_k = \tfrac{2}{k+1} \tag{8.15}$$

$$y_k = a_k v_{k-1} + (1 - a_k) x_{k-1} \tag{8.16}$$

$$\text{Set } x_k \text{ to be a } \varepsilon_k\text{-approximate minimizer of } F_{1/\tilde{\mu}}(\cdot; y_k) \tag{8.17}$$

$$\text{Set } v_k \text{ to be a } \delta_k\text{-approximate minimizer of } F_{\frac{1}{\tilde{\mu} a_k}, a_k}(\cdot; y_k, v_{k-1}) \tag{8.18}$$

[40,62,68]. In particular, our proof shares many features with that of [62], relying on approximate subdifferentials and the recurrence relation [62, Lemma 1].

Theorem 8.9 (Convergence of the accelerated prox-linear algorithm: near-optimality)**.** *Fix a real number $\tilde{\mu} > \mu$, and let x^* be any point satisfying $F(x^*) \leq F(x_k)$ for iterates x_k generated by Algorithm 10. Then the iterates x_k satisfy the inequality:*

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

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:

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

Note that with the choices ε_i, $\delta_i \sim \frac{1}{i^{3+q}}$, the quantity A_N remains bounded. Consequently, in the setting $r = 0$, the functional error $F(x_N) - F(x^*)$ is on the order of $\mathcal{O}(1/N^2)$.

Acknowledgements

We thank the two anonymous referees for their meticulous reading of the manuscript. Their comments and suggestions greatly improved the quality and readability of the paper. We also thank Damek Davis and Zaid Harchaoui for their insightful comments on an early draft of the paper.

References

[1] Z. Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *Preprint arXiv:1603.05953 (version 5)*, 2016.

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

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

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

[5] J. Bolte and E. Pauwels. Majorization-minimization procedures and convergence of SQP methods for semi-algebraic and tame programs. *Math. Oper. Res.*, 41(2):442–465, 2016.

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

[7] G. Brassard and P. Bratley. *Fundamentals of algorithmics*. Prentice Hall, Inc., Englewood Cliffs, NJ, 1996.

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

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

[10] 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.

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

[12] 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.

[13] 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.

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

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

[16] 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.

[17] 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.

[18] A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pages 1646–1654, 2014.

[19] A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1646–1654. Curran Associates, Inc., 2014.

[20] Aaron Defazio, Justin Domke, and Tibério S Caetano. Finito: A faster, permutable incremental gradient method for big data problems. In *ICML*, pages 1125–1133, 2014.

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

[22] 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.

[23] D. Drusvyatskiy and A.S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *To appear in Math. Oper. Res., arXiv:1602.06661*, 2016.

[24] J.C. Duchi and F. Ruan. Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *Preprint arXiv:1705.02356*, 2017.

[25] J.C. Duchi and F. Ruan. Stochastic methods for composite optimization problems. *Preprint arXiv:1703.08570*, 2017.

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

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

[28] 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).

[29] R. Frostig, R. Ge, S.M. Kakade, and A. Sidford. Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

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

[31] N. Gillis. The why and how of nonnegative matrix factorization. In *Regularization, optimization, kernels, and support vector machines*, Chapman & Hall/CRC Mach. Learn. Pattern Recogn. Ser., pages 257–291. CRC Press, Boca Raton, FL, 2015.

[32] N. Gillis. Introduction to nonnegative matrix factorization. *SIAG/OPT Views and News*, 25(1):7–16, 2017.

[33] 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.

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

[35] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS'13, pages 315–323, USA, 2013. Curran Associates Inc.

[36] G. Lan. An optimal randomized incremental gradient method. *arXiv:1507.02000*, 2015.

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

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

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

[40] H. Lin, J. Mairal, and Z. Harchaoui. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems*, pages 3366–3374, 2015.

[41] H. Lin, J. Mairal, and Z. Harchaoui. QuickeNing: A generic Quasi-Newton algorithm for faster gradient-based optimization. *Preprint arXiv:1610.00960*, 2016.

[42] J. Mairal. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 25(2):829–855, 2015.

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

[44] 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.

[45] 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.

[46] 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.

[47] 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.

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

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

[50] 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.

[51] Yu. Nesterov. On an approach to the construction of optimal methods of minimization of smooth convex functions. *Ekonom. i. Mat. Metody*, 24:509–517, 1988.

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

[53] Yu. Nesterov. Modified Gauss-Newton scheme with worst case guarantees for global performance. *Optim. Methods Softw.*, 22(3):469–483, 2007.

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

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

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

[57] R.A. Poliquin and R.T. Rockafellar. Prox-regular functions in variational analysis. *Trans. Amer. Math. Soc.*, 348:1805–1838, 1996.

[58] 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.

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

[60] R.T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

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

[62] 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.

[63] M. Schmidt, N. Le Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *arXiv:1309.2388*, 2013.

[64] S. Shalev-Shwartz and T. Zhang. Proximal stochastic dual coordinate ascent. *arXiv:1211.2717*, 2012.

[65] S. Shalev-Shwartz and T. Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Mathematical Programming*, 2015.

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

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

[68] S. Villa, S. Salzo, L. Baldassarre, and A. Verri. Accelerated and inexact forward-backward algorithms. *SIAM J. Optim.*, 23(3):1607–1633, 2013.

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

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

[71] L. Xiao and T. Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM J. Optim.*, 24(4):2057–2075, 2014.

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

A Proofs of Lemmas 5.3, 7.1 and Theorems 8.8, 8.9

In this section, we prove Lemmas 5.3, 7.1 and Theorems 8.8, 8.9 in order.

Proof of Lemma 5.3. Observe for any $t > 0$ and any proper, closed, convex function f, we have

$$\operatorname{prox}_{(tf)^\star}(w) = \operatorname*{argmin}_z \ \{tf^\star(z/t) + \tfrac{1}{2}\|z - w\|^2\} = t \cdot \operatorname{prox}_{f^\star/t}(w/t), \tag{A.1}$$

where the first equation follows from the definition of the proximal map and from [60, Theorem 16.1]. From [60, Theorem 31.5], we obtain $\operatorname{prox}_{th^\star}(w) = w - \operatorname{prox}_{(th^\star)^\star}(w)$, while an application of (A.1) with $f = h^\star$ then directly implies (5.6).

The fact that the gradient map $\nabla\Big(G^\star \circ A^* - \langle b, \cdot\rangle\Big)$ is Lipschitz with constant $t\|\nabla c(x)\|_{\mathrm{op}}^2$ follows directly from $\nabla G^\star$ being t-Lipschitz continuous. The chain rule, in turn, yields

$$\nabla\Big(G^\star \circ A^* - \langle b, \cdot\rangle\Big)(w) = A\nabla G^\star(A^* w) - b.$$

Thus we must analyze the expression $\nabla G^\star(z) = \nabla(g + \frac{1}{2t}\|\cdot - x\|^2)^\star(z)$. Notice that the conjugate of $\frac{1}{2t}\|\cdot - x\|^2$ is the function $\frac{t}{2}\|\cdot\|^2 + \langle\cdot, x\rangle$. Hence, using [60, Theorem 16.4] we deduce

$$(g + \tfrac{1}{2t}\|\cdot - x\|^2)^\star(z) = \inf_y \ \{g^\star(y) + \tfrac{t}{2}\|z - y\|^2 + \langle z - y, x\rangle\} = (g^\star)_{1/t}(z + x/t) - \tfrac{1}{2t}\|x\|^2,$$

where the last equation follows from completing the square. We thus conclude

$$\nabla G^\star(z) = \nabla(g^\star)_{1/t}(z + x/t) = t \cdot \operatorname{prox}_{(g^\star/t)^\star}(z + x/t) = \operatorname{prox}_{tg}(x + tz),$$

where the second equality follows from Lemma 2.1 and the third from (A.1). The expressions (5.7) and (5.8) follow. □

Proof of Lemma 7.1. Observe

$$\|h(y) - h(z)\| \leq \frac{1}{m}\sum_{i=1}^{m} |h_i(y_i) - h_i(z_i)| \leq \frac{L}{m}\sum_{i=1}^{m} \|y - z\|_1 \leq \frac{L}{\sqrt{m}}\|y - z\|,$$

where the last equality follows from the l_p-norm comparison $\|\cdot\|_1 \leq \sqrt{m}\|\cdot\|_2$. This proves $\operatorname{lip}(h) \leq L/\sqrt{m}$. Next for any point x observe

$$\|\nabla c(x)\|_{\mathrm{op}} = \max_{v:\|v\|=1} \|\nabla c(x)v\| \leq \sqrt{\sum_{i=1}^{m} \|\nabla c_i(x)\|^2} \leq \sqrt{m} \max_{i=1,\ldots,m} \|\nabla c_i(x)\|$$

By an analogous argument, we have

$$\|\nabla c(x) - \nabla c(z)\|_{\mathrm{op}} \leq \sqrt{\sum_{i=1}^{m} \|\nabla c_i(x) - \nabla c_i(z)\|^2} \leq \beta\sqrt{m}\|x - z\|,$$

and hence $\operatorname{lip}(\nabla c) \leq \beta\sqrt{m}$. Finally, suppose that each h_i is C^1-smooth with L_h-Lipschitz gradient ∇h_i. Observe then

$$\|\nabla h(y) - \nabla h(z)\| = \frac{1}{m}\sqrt{\sum_{i=1}^{m} |h_i'(y_i) - h_i'(z_i)|^2} \leq \frac{L_h}{m}\|y - z\|^2.$$

The result follows. □

Proof of Theorem 8.8. The proof is a modification of the proof Theorem 8.5; 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 (8.9) 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 (8.5) except for the error term $2L(\delta_k + \varepsilon_k)$. The same argument as in the proof of Theorem 8.5 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 5.5, 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

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

Combining the first and fourth terms and using the inequality $\tilde{\mu} \ge \mu$ 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 8.5. □

We next prove Theorem 8.9. To this end, we will need the following lemma.

Lemma A.1 (Lemma 1 in [62])**.** *Suppose the following recurrence relation is satisfied*

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

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

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

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

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

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

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 [33, 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*

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

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

$$t^{-1}(v - z^+ - \gamma) \in \partial_\varepsilon F_\alpha(z^+; y, v). \tag{A.2}$$

We are now ready to prove Theorem 8.9.

Proof of Theorem 8.9. Let x_k, y_k, and v_k be the iterates generated by Algorithm 10. We imitate the proof of Theorem 8.5, while taking into account inexactness. First, inequality (8.6) 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.2), 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

$$F(x_k) \leq F(w; y_k) + \tfrac{\mu}{2}\|x_k - y_k\|^2 + \langle \tilde{\mu}(y_k - x_k - \gamma_k), x_k - w\rangle + \varepsilon_k. \tag{A.3}$$

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

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

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.4) and dropping the term $\frac{\tilde{\mu}}{2}\|\gamma_k + c_k\|^2$, we deduce

$$\begin{aligned} F(x_k) \leq\ & h(c(y_k) + a_k\nabla c(y_k)(v_k - v_{k-1})) + a_k g(v_k) + (1 - a_k)g(x_{k-1}) \\ & + \tfrac{\tilde{\mu}a_k^2}{2}\|v_k - v_{k-1}\|^2 - \tfrac{\tilde{\mu}-\mu}{2}\|x_k - y_k\|^2 + \varepsilon_k + \tfrac{\tilde{\mu}}{2}\|\gamma_k\|^2. \end{aligned} \tag{A.5}$$

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.2), 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

$$\begin{aligned} F_{a_k}(v_k; y_k, v_{k-1}) \leq\ & \frac{1}{a_k}h(c(y_k) + a_k\nabla c(y_k)(x - v_{k-1}) + g(x) \\ & + \tilde{\mu}a_k\langle v_{k-1} - v_k - \eta_k, v_k - x\rangle + \delta_k. \end{aligned} \tag{A.6}$$

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.5) and (A.6), 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 8.5, 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 8.5, setting $x = x^*$, we deduce

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

In particular, we have

$$\begin{aligned}\frac{\tilde{\mu}-\mu}{2}\sum_{i=1}^N\frac{\|x_i - y_i\|^2}{a_i^2} \leq & \frac{\tilde{\mu}}{2}\|x^* - v_0\|^2 + \frac{\rho M^2 N(N+3)}{4} + \frac{NrM^2}{2} + \sum_{i=1}^N\frac{\delta_i}{a_i} \\ & + 2\sum_{i=1}^N\frac{\varepsilon_i}{a_i^2} + \sqrt{2\tilde{\mu}}\sum_{i=1}^N\|x^* - v_i\|\cdot\sqrt{\frac{\delta_i}{a_i}}.\end{aligned} \tag{A.7}$$

and

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

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

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

Finally, combining inequality (A.7) with Lemma 5.1 we deduce

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

Hence

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

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 8.5. □

B Backtracking

In this section, we present a variant of Algorithm 8 where the constants L and β are unknown. The scheme is recorded as Algorithm 12 and relies on a backtracking line-search, stated in Algorithm 11.

Algorithm 11: Backtracking(η, α, t, y)

Initialize: A point y and real numbers $\eta, \alpha \in (0,1)$ and $t > 0$.
while $F(S_{\alpha t}(y)) > F_t(S_{\alpha t}(y))$ **do**
  $t \leftarrow \eta t$
end
Set $\tilde{\mu} = \frac{1}{\alpha t}$ and $x = S_{\alpha t}(y)$
return $\tilde{\mu}, t, x$;

The backtracking procedure completes after only logarithmically many iterations.

Lemma B.1 (Termination of backtracking line search)**.** *Algorithm 11 on input (η, α, t, y) terminates after at most $1 + \left\lceil \frac{\log(t\mu)}{\log(\eta^{-1})} \right\rceil$ evaluations of $S_\alpha.(y)$.*

Algorithm 12: Accelerated prox-linear method with backtracking

Initialize: Fix two points $x_0, v_0 \in \operatorname{dom} g$ and real numbers $t_0 > 0$ and $\eta, \alpha \in (0,1)$.

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

$$\begin{aligned} a_k &= \tfrac{2}{k+1} \\ y_k &= a_k v_{k-1} + (1-a_k) x_{k-1} \\ (\tilde{\mu}_k, t_k, x_k) &= \text{Backtracking}(\eta, \alpha, t_{k-1}, y_k) \\ v_k &= S_{\frac{1}{\tilde{\mu}_k a_k}, a_k}(y_k, v_{k-1}) \end{aligned}$$

Proof. This follows immediately by observing that the loop in Algorithm 11 terminates as soon as $t \leq \mu^{-1}$. □

We now establish convergence guarantees of Algorithm 12, akin to those of Algorithm 8.

Theorem B.2 (Convergence guarantees with backtracking)**.** *Fix real numbers $t_0 > 0$ and $\eta, \alpha \in (0,1)$ and let x^* be any point satisfying $F(x^*) \leq F(x_k)$ for all iterates x_k generated by Algorithm 12. Define $\tilde{\mu}_{\max} := \max\{(\alpha t_0)^{-1}, (\alpha\eta)^{-1}\mu\}$ and $\tilde{\mu}_0 := (\alpha t_0)^{-1}$. Then the efficiency estimate holds:*

$$\min_{j=1,\ldots,N} \left\| \mathcal{G}_{1/\tilde{\mu}_j}(y_j) \right\|^2 \leq \frac{24\tilde{\mu}_{\max}}{1-\alpha} \left(\frac{\tilde{\mu}_0 \|x^* - v_0\|^2}{N(N+1)(2N+1)} + \frac{M^2\left(r + \frac{\rho}{2}(N+3)\right)}{(N+1)(2N+1)} \right).$$

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:

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

Proof. We closely follow the proofs of Lemma 8.7 and Theorem 8.5, as such, we omit some details. For $k \geq 1$, the stopping criteria of the backtracking algorithm guarantees that analogous inequalities (8.6) and (8.7) hold, namely,

$$F(x_k) \leq h\big(c(y_k) + \nabla c(y_k)(x_k - y_k)\big) + g(x_k) + \frac{1}{2t_k}\|x_k - y_k\|^2 \tag{B.1}$$

and

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

where $w_k := a_k v_k + (1-a_k)x_{k-1}$. By combining (B.1) and (B.2) together with the definition that $\tilde{\mu}_k = (\alpha t_k)^{-1}$, we conclude

$$\begin{aligned} F(x_k) \leq\ & h\big(c(y_k) + \nabla c(y_k)(w_k - y_k)\big) + a_k g(v_k) + (1-a_k) g(x_{k-1}) \\ & + \frac{\tilde{\mu}_k}{2}\left(\|w_k - y_k\|^2 - \|w_k - x_k\|^2\right) + \frac{\left(1 - \alpha^{-1}\right)}{2t_k}\|x_k - y_k\|^2. \end{aligned} \tag{B.3}$$

We note the equality $w_k - y_k = a_k(v_k - v_{k-1})$. Observe that (8.8) holds by replacing $\frac{\tilde{\mu}}{2}$ with $\frac{\tilde{\mu}_k}{2}$; hence, we obtain for all points x

$$\begin{aligned} h\big(c(y_k) + a_k\nabla c(y_k)(v_k - v_{k-1})\big) + a_k g(v_k) \le h\big(c(y_k) + a_k\nabla c(y_k)(x - v_{k-1})\big) + a_k g(x) \\ + \frac{\tilde{\mu}_k a_k^2}{2}\left(\|x - v_{k-1}\|^2 - \|x - v_k\|^2 - \|v_k - v_{k-1}\|^2\right). \end{aligned} \tag{B.4}$$

Notice also that (8.9) holds as stated. Combining the inequalities (8.9), (B.3), and (B.4), we deduce

$$\begin{aligned} F(x_k) \le a_k F(x) + (1 - a_k)F(x_{k-1}) + \frac{\tilde{\mu}_k a_k^2}{2}\left(\|x - v_{k-1}\|^2 - \|x - v_k\|^2\right) \\ - \frac{(\alpha^{-1} - 1)}{2t_k}\|y_k - x_k\|^2 + \rho a_k(1 - a_k)\|x - x_{k-1}\|^2 + \frac{r a_k^2}{2}\|x - v_{k-1}\|^2. \end{aligned} \tag{B.5}$$

Plugging in $x = x^*$, subtracting $F(x^*)$ from both sides, and rearranging yields

$$\begin{aligned} \frac{F(x_k) - F(x^*)}{a_k^2} + \frac{\tilde{\mu}_k}{2}\|x^* - v_k\|^2 \le \frac{1 - a_k}{a_k^2}(F(x_{k-1}) - F(x^*)) + \frac{\tilde{\mu}_k}{2}\|x^* - v_{k-1}\|^2 \\ + \frac{\rho M^2}{a_k} + \frac{rM^2}{2} - \frac{(\alpha^{-1} - 1)}{2t_k a_k^2}\|y_k - x_k\|^2. \end{aligned}$$

This is exactly inequality (8.10) with $\frac{\tilde{\mu}}{2}$ replaced by $\frac{\tilde{\mu}_k}{2}$ and $\frac{\tilde{\mu} - \mu}{2}$ replaced by $\frac{(\alpha^{-1} - 1)}{2t_k}$; Using the fact that the sequence $\{\tilde{\mu}_k\}_{k=0}^{\infty}$ is nondecreasing and $\frac{1 - a_k}{a_k^2} \le \frac{1}{a_{k-1}^2}$, we deduce

$$\begin{aligned} \frac{F(x_k) - F(x^*)}{a_k^2} + \frac{\tilde{\mu}_k}{2}\|x^* - v_k\|^2 \le \frac{\tilde{\mu}_k}{\tilde{\mu}_{k-1}}\Bigg(\frac{F(x_{k-1}) - F(x^*)}{a_{k-1}^2} + \frac{\tilde{\mu}_{k-1}}{2}\|x^* - v_{k-1}\|^2 \\ + \frac{\rho M^2}{a_k} + \frac{rM^2}{2} - \frac{(\alpha^{-1} - 1)}{2t_k a_k^2}\|y_k - x_k\|^2\Bigg). \end{aligned} \tag{B.6}$$

Notice $\tilde{\mu}_k \le \alpha^{-1}\max\{t_0^{-1}, \eta^{-1}\mu\} =: \tilde{\mu}_{\max}$. Recursively applying (B.6) N times, we get

$$\begin{aligned} \frac{F(x_N) - F(x^*)}{a_N^2} + \frac{\tilde{\mu}_N}{2}\|x^* - v_N\|^2 \le \left(\prod_{j=1}^{N}\frac{\tilde{\mu}_j}{\tilde{\mu}_{j-1}}\right)\Bigg(\frac{\tilde{\mu}_0}{2}\|x^* - v_0\|^2 + \sum_{j=1}^{N}\frac{\rho M^2}{a_j} + \frac{NrM^2}{2} \\ - \sum_{j=1}^{N}\frac{(\alpha^{-1} - 1)}{2t_j}\cdot\frac{\|x_j - y_j\|^2}{a_j^2}\Bigg) \end{aligned} \tag{B.7}$$

By the telescoping property of $\prod_{j=1}^{N}\frac{\tilde{\mu}_j}{\tilde{\mu}_{j-1}} \le \frac{\tilde{\mu}_{\max}}{\tilde{\mu}_0}$, we conclude

$$\frac{\tilde{\mu}_{\max}}{\tilde{\mu}_0}\sum_{j=1}^{N}\frac{(\alpha^{-1} - 1)}{2t_j}\cdot\frac{\|x_j - y_j\|^2}{a_j^2} \le \frac{\tilde{\mu}_{\max}}{\tilde{\mu}_0}\left(\frac{\tilde{\mu}_0}{2}\|x^* - v_0\|^2 + \rho M^2\left(\sum_{j=1}^{N}\frac{1}{a_j}\right) + \frac{NrM^2}{2}\right). \tag{B.8}$$

Using the inequality (B.8) and $\alpha t_j = \tilde{\mu}_j^{-1} \ge \tilde{\mu}_{\max}^{-1}$ for all j, we conclude

$$\frac{(\alpha^{-1} - 1)\alpha}{2\tilde{\mu}_0}\cdot\left(\sum_{j=1}^{N}\frac{1}{a_j^2}\right)\min_{j=1,\ldots,N}\|\tilde{\mu}_j(x_j - y_j)\|^2 \le \frac{\tilde{\mu}_{\max}}{\tilde{\mu}_0}\left(\frac{\tilde{\mu}_0}{2}\|x^* - v_0\|^2 + \rho M^2\left(\sum_{j=1}^{N}\frac{1}{a_j}\right) + \frac{NrM^2}{2}\right).$$

The result follows by mimicking the rest of the proof in Theorem 8.5. Finally, suppose $r = 0$, and hence we can assume $\rho = 0$. Inequality (B.7) then implies

$$\frac{F(x_N) - F(x^*)}{a_N^2} + \frac{\tilde{\mu}_N}{2} \|x^* - v_N\|^2 \leq \frac{\tilde{\mu}_{\max}}{\tilde{\mu}_0} \cdot \frac{\tilde{\mu}_0}{2} \|x^* - v_0\|^2 .$$

The claimed efficiency estimate follows. □

C Removing the logarithmic dependence when an estimate on $F(x_0) - \inf F$ is known.

In this section, we show that if a good estimate on the error $F(x_0) - \inf F$ is available, then there is a first-order method for the composite problem class 3.1 with efficiency $\mathcal{O}\left(\frac{L^2\beta\|\nabla c\|\cdot(F(x_0)-\inf F)}{\varepsilon^3}\right)$. Notice that this is an improvement over (6.21) since there is no logarithmic term. The outline is as follows. We will fix at the very beginning a budget of basic operations we are willing to tolerate. We will then perform a constant number of iterations of the inexact prox-linear Algorithm 2 with a constant number of iterations of an accelerated primal-dual first-order method on the proximal subproblem. Before delving into the details, it is important to note two downsides of the scheme, despite the improved worst-case efficiency over the smoothing technique. First, we must have a good estimate on $F(x_0) - \inf F$. Secondly, the number of inner iterations we are willing to tolerate depends on $\|\nabla c\|$, rather than on the norms $\|\nabla c(x_k)\|_{\mathrm{op}}$ along the generated iterate sequences x_k. The reason is that the number of iterations (both outer and inner) must be set a priori, without knowledge of the iterates that will be generated. This is in direct contrast to the algorithms discussed in Section 6, where the dependences on $\|\nabla c\|$ could always be replaced by an upper bound on $\max_k \|\nabla c(x_k)\|_{\mathrm{op}}$ along the generated iterate sequence x_k. Nonetheless, from the complexity viewpoint, the improved efficiency estimate is notable.

We now describe the outlined strategy in detail. In order to find approximate minimizers of the proximal subproblems (5.3), let us instead focus on the dual (5.4), and apply a (fast) primal-dual method with sublinear guarantees. To specify precisely the method we will use on the subproblems, we follow the exposition in [67]. Recall that $G^\star$ is C^1-smooth with t-Lipschitz gradient. Moreover since h is L-Lipschitz, the domain of the function $w \mapsto h^\star(w) - \langle b, w\rangle$ has diameter upper bounded by $2L$. In Algorithm 13, we record the specialization of [67, Algorithm 1] to our target problem (5.4).[8]

Algorithm 13 comes equipped with the following guarantee [67, Corollary 1(b)].

Theorem C.1. *For every index j, the iterates generated by Algorithm 13 satisfy:*

$$F_t(v_j; x) - \inf F_t(\cdot; x) \leq \frac{8lL^2}{(j+2)^2}.$$

Set $t = 1/\mu$ and fix a real $q > 0$, which will appear in the final efficiency estimate. Suppose that we aim to run a total of at most T iterations of Algorithm 13 over all the proximal subproblems. Suppose moreover that T is sufficiently large to satisfy $T \geq \frac{4(1.5)^{3/2}\|\nabla c\|}{\sqrt{2\beta q/L}}$.

[8] In the notation of [67], we set $\phi(w, v) := \langle v, A^*w\rangle - G(v)$ and $p(w) := h^\star(w) - \langle b, w\rangle$, and note $\mathrm{prox}_{tp}(\cdot) = \mathrm{prox}_{th^\star}(\cdot + tb)$.

Algorithm 13: Optimal method (Auslender-Teboulle [3], Tseng [67, Algorithm 1])

Initialize: Fix two points $w_0, z_0 \in \operatorname{dom} p$; choose a real $l \geq t\|A\|^2$; set $v_{-1} := 0$ and $a_0 := 1$.

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

$$\begin{aligned} y_j &= (1-a_j)w_j + a_j z_j \\ z_{j+1} &= \operatorname{prox}_{\frac{h^\star}{a_j l}}\Big(z_j - \tfrac{1}{a_j l}(\nabla G^\star(y_j) - b)\Big) \\ w_{j+1} &= (1-a_j)w_j + a_j z_{j+1} \\ a_{j+1} &= \tfrac{\sqrt{a_j^4 + 4a_j^2} - a_j^2}{2} \end{aligned}$$

Update the primal iterate

$$v_j = (1-a_j)v_{j-1} + a_j \nabla G^\star(A^* y_j)$$

Consider now the following procedure. Define

$$N := \left\lceil \left(\frac{T\sqrt{2\beta q/L}}{4\|\nabla c\|} \right)^{2/3} \right\rceil - 2$$

and note $N \geq 0$. Let us now run the inexact prox-linear Algorithm 2 for $k = 0, \ldots, N$ iterations with each prox-linear subproblem approximately solved by running

$$\left\lceil 4\|\nabla c\| \sqrt{\tfrac{L(N+1)}{2\beta q}} \right\rceil$$

iterations of Algorithm 13; we will determine an estimate on the incurred errors $\varepsilon_k > 0$ shortly. Observe that the total number of iterations of Algorithm 13 is indeed at most

$$(N+1) \cdot \left\lceil 4\|\nabla c\| \sqrt{\tfrac{L(N+1)}{2\beta q}} \right\rceil \leq 4\|\nabla c\| \sqrt{L} (N+1)^{3/2} / \sqrt{2\beta q} \leq T.$$

Appealing to Theorem C.1, we deduce

$$\begin{aligned} F_{1/\mu}(x_{k+1}; x_k) - \inf F_{1/\mu}(\cdot; x_k) &\leq \frac{8\|\nabla c\|^2 L^2/\mu}{\lceil 4\|\nabla c\| \sqrt{L(N+1)/(2\beta q)} \rceil^2} \\ &\leq \frac{8\|\nabla c\|^2 L^2/\mu}{16 L \|\nabla c\|^2 (N+1)/(2\beta q)} = \frac{q}{N+1}. \end{aligned}$$

Thus, in the notation of Algorithm 2 we can set $\varepsilon_k := \frac{q}{N+1}$ for each index k. Theorem 5.2 then yields the estimate

$$\min_{i=0,\ldots,N-1} \big\|\mathcal{G}_{1/\mu}(x_i)\big\|^2 \leq \frac{2\mu\big(F(x_0) - \inf F + q\big)}{N} \leq \frac{2\mu\big(F(x_0) - \inf F + q\big)}{\left(\frac{T\sqrt{2\beta q/L}}{4\|\nabla c\|}\right)^{2/3} - 2},$$

Thus to find a point x with $\|\mathcal{G}_{1/\mu}(x)\| \leq \varepsilon$ it suffices to choose T satisfying

$$T \geq \frac{8\|\nabla c\|}{\sqrt{\beta q/L}} \cdot \left(1 + \frac{\mu(F(x_0) - \inf F + q)}{\varepsilon^2}\right)^{3/2}.$$

Notice that the assumed bound $T \geq \frac{4(1.5)^{3/2}\|\nabla c\|}{\sqrt{2\beta q/L}}$ holds automatically for this choice of T.

In particular, if q can be chosen to satisfy $\frac{q}{F(x_0)-\inf F} \in [\gamma_1, \gamma_2]$ for some fixed constants $\gamma_2 \geq \gamma_1 \geq 1$, the efficiency estimate becomes on the order of

$$\boxed{\mathcal{O}\left(\frac{L^2\beta\|\nabla c\| \cdot (F(x_0) - \inf F)}{\varepsilon^3}\right)},$$

as claimed.