An accelerated algorithm for minimizing convex compositions *

D. Drusvyatskiy † C. Kempton ‡

April 30, 2016

Abstract

We describe a new proximal algorithm for minimizing compositions of finite-valued convex functions with smooth mappings. When applied to convex optimization problems having an additive composite form, the algorithm reduces to FISTA. The method both realizes the best known complexity bound of $\mathcal{O}(1/\epsilon)$ in optimality conditions whenever the objective function has bounded domain, and achieves the accelerated rate $\mathcal{O}(1/\sqrt{\epsilon})$ in function values under standard convexity assumptions. A natural convexity parameter of the composition quantifies the transition between the two modes of convergence.

Key words. Composite minimization, Fast gradient methods, Gauss-Newton, prox-gradient

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

1 Introduction

Our work revolves around minimization problems of the form:

$$\min_x F(x) := g(x) + h(c(x)), \tag{1}$$

where g is a closed convex function (possibly extended-real-valued), h is a finite-valued closed convex function, and c is a smooth mapping. Problems of this form afford great modeling flexibility. For instance, nonlinear least squares problems [14, Section 10] correspond to the setting $h = \frac{1}{2}\|\cdot\|^2$. The function g allows to incorporate side constraints or

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

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

‡E-mail: yumiko88@uw.edu;

structure inducing regularization, such as sparsity or low-rank. Such considerations appear often in recovery problems where the design matrix is corrupted; e.g. structure-aware total least squares [19, 23]. The setting where c maps to the real line and h is the identity function is now ubiquitous in optimization [1,5,13,20]. We will call this situation "additive composite minimization" to distinguish it from the more general problem class (1).

Given the expressive power of (1), it is important to develop algorithms for the problem class with provable global convergence guarantees. One such method was investigated in [9]. The ideas behind the scheme (and of its trust-region variants) go back much earlier. See for example [2, 7, 16, 17, 21, 22], or [2] for a historical discussion. The *prox-linear method* iterates the steps:

$$x_{k+1} = \operatorname*{argmin}_x \Big\{ g(x) + h\big(c(x_k) + \nabla c(x_k)(x - x_k)\big) + \tfrac{1}{2t_k}\|x - x_k\|^2 \Big\}.$$

Here, the control sequence $t_k > 0$ can be chosen by a back-tracking line search for example. In the setting of additive composite minimization, the prox-linear method reduces to the popular prox-gradient algorithm (e.g. [1, Section 2.1], [13]), while for nonlinear least squares, the scheme is the Levenberg-Marquardt method – a damped variant of the Gauss-Newton algorithm [14, Section 10].

For simplicity, suppose that we set $t_k := L\beta$, where L is the Lipschitz constant of h and β is the Lipschitz constant of the Jacobian ∇c. It is natural to measure the progress made by the prox-linear method by the scaled step lengths $\rho_k := t^{-1}\|x_k - x_{k+1}\|$. Indeed, the quantities ρ_k measure the approximate stationarity of the iterates.[1] The prox-linear algorithm has the complexity bound $\min_{i=1,\dots,k} \rho_i^2 \leq \mathcal{O}(\frac{1}{k})$; see e.g. [6, Section 5].

Within the class of additive composite minimization problems, where c is in addition a convex function, the prox-gradient method is suboptimal from the viewpoint of computational complexity [10, 11]. Accelerated gradient methods, beginning with [12] and [1, 13] achieve the superior and best possible rate, $F(x_k) - F^* \leq \mathcal{O}(\frac{1}{k^2})$. Left standing was an intriguing question of designing an acceleration scheme realizing optimal convergence rates for convex problems, while performing no worse than the prox-gradient method on problems lacking convexity. One would expect such an algorithm to significantly outperform the prox-gradient method on nonconvex instances. In the recent manuscript [8], Ghadimi and Lan answered this question in the affirmative. Acceleration techniques have also been used for the broad problem class (1) with numerical success, but without formal justification [4].

Our goals are succinct. We propose an accelerated scheme for the problem class (1), generalizing the algorithm proposed by Ghadimi and Lan [8] for additive composite minimization. Assuming that the domain of g is bounded, the algorithm has worst-case convergence guarantees analogous to those of the prox-linear method and achieving optimal rates for composite functions satisfying a convexity condition. We precisely quantify the balance between the two modes of convergence, based on an intuitive constant measuring convexity of the composition, or lack thereof.

[1] For any index k there always exists a point $\hat{x}$, satisfying $\|\hat{x} - x_{k+1}\| \leq t\rho_k$ and $F'(\hat{x}; v) \geq -5\beta\rho_k$ for all unit vectors v [6, Theorem 5.9].

1.1 Notation

The notation we follow is standard. The *domain* and the *epigraph* of any function $f\colon \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ are the sets

$$\begin{aligned} \operatorname{dom} f &:= \{x \in \mathbf{R}^n : f(x) < +\infty\}, \\ \operatorname{epi} f &:= \{(x, r) \in \mathbf{R}^n \times \mathbf{R} : f(x) \leq r\}, \end{aligned}$$

respectively. We say that f is *closed* if $\operatorname{epi} f$ is a closed set. The *subdifferential* of a convex function f at a point $x \in \operatorname{dom} f$ is the set

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

The *Fenchel conjugate* of a function $f\colon \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ is the function

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

Whenever f is closed and convex, equality $f = f^{\star\star}$ holds. For any convex function f and parameter $t > 0$, the *proximal mapping* is the assignment

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

2 Main results

Our work centers around optimization problems of the form

$$\min_x \; F(x) := g(x) + h(c(x)). \tag{2}$$

Throughout, we make the following assumptions:

- (*i*) $g\colon \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ is a closed convex function, with the diameter of its domain bounded by some constant $M > 0$;
- (*ii*) $h\colon \mathbf{R}^m \to \mathbf{R}$ is a finite-valued L-Lipschitz continuous convex function;
- (*iii*) $c\colon \mathbf{R}^n \to \mathbf{R}^m$ is a C^1-smooth mapping with the Jacobian $x \mapsto \nabla c(x)$ that is β-Lipschitz continuous.

Since the optimization problem (2) is nonconvex, it is natural to seek points x that are only *first-order stationary*, meaning that the directional derivative

$$F'(x; \bar{v}) := \liminf_{\substack{\tau \searrow 0 \\ v \to \bar{v}}} \frac{F(x + \tau v) - F(x)}{t}$$

is nonnegative in every direction $\bar{v}$. Equivalently, these are the points x satisfying the inclusion

$$0 \in \partial g(x) + \nabla c(x)^* \partial h(c(x)).$$

See for example [15] or [18, Example 10.8] for the stated equivalence.

Linearizing c, the *prox-linear mapping* appears naturally:

$$S_t(x) := \operatorname*{argmin}_z \big\{ g(z) + h\big(c(x) + \nabla c(x)(z-x)\big) + \frac{1}{2t}\|z-x\|^2 \big\}.$$

The *prox-linear algorithm* is then simply the recurrence:

$$x_{k+1} = S_{\frac{1}{L\beta}}(x_k).$$

In the additive composite setting, where c maps to the real line and h is the identity function, 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 $x_{k+1} = \operatorname{prox}_{\beta^{-1}g}(x_k - \beta^{-1}\nabla c(x_k))$.

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

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

Note that for any $t > 0$, a point x is first-order stationary for F if and only if equality $\mathcal{G}_t(x) = 0$ holds. Hence the norm $\|\mathcal{G}_t(x)\|$ serves as a measure of proximity to stationarity. More precisely, [6, Theorem 5.9] shows that for any point x there exists a "nearby approximately stationary point" $\hat{x}$ satisfying

$$\|\hat{x} - S_t(x)\| \le t\|\mathcal{G}_t(x)\| \qquad \text{and} \qquad F'(\hat{x}; v) \ge -(3L\beta t + 2)\|\mathcal{G}_t(x)\|$$

for all unit vectors v. In the additive composite case, one can simply set $\hat{x} := S_t(x)$; the more general setting requires a perturbation. Denoting by F^* the limit of the decreasing sequence $\{F(x_i)\}$, the iterates generated by the prox-linear method satisfy [6, Section 5]

$$\min_{i=1,\ldots,k} \|\mathcal{G}_{\frac{1}{L\beta}}(x_i)\|^2 \le \frac{2L\beta(F(x_1) - F^*)}{k} \qquad \text{for all indices } k.$$

In this paper, we describe an accelerated version of the prox-linear method, in a sense to be made precise shortly. To this end, we will use the following variant of the prox-linear mapping:

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

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

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

Algorithm 1: Accelerated prox-linear method

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

$k \leftarrow 1$

while $\left\|\mathcal{G}_{1/\tilde{\mu}}(y_{k-1})\right\| > \varepsilon$ **do**

  $a_k \leftarrow \frac{2}{k+1}$

  $y_k \leftarrow a_k v_{k-1} + (1-a_k)x_{k-1}$

  $x_k \leftarrow S_{1/\tilde{\mu}}(y_k)$

  $v_k \leftarrow \begin{cases} x_{k-1} + \frac{1}{a_k}(x_k - x_{k-1}), & \|x_k - x_{k-1}\|^2 \leq \frac{M^2 a_k}{(1-a_k)^2} \\ S_{\frac{1}{\tilde{\mu} a_k}, a_k}(y_k, v_{k-1}), & \text{otherwise} \end{cases}$.

  $k \leftarrow k+1$

end

return $x\mathrm{Sol} = x_k$

This is exactly the construction fundamentally used by Ghadimi and Lan [8, Equation 2.37]. We are now ready to state our proposed scheme in Algorithm 1.

We remark that when the constants L and β are unknown, one can instead equip Algorithm 1 with a backtracking line search. More precisely, one can insert Algorithm 2 after line 4 in Algorithm 1. The line search is entirely analogous to the one used in FISTA [1]. For succinctness, we only analyze the complexity of the constant stepsize scheme with the adage that the resulting guarantees for the backtracking line search are completely analogous.

Algorithm 2: Backtracking line search in iteration k

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

$t \leftarrow t_{k-1}$

while $h\big(c(S_t(y_k))\big) > h\big(c(y_k) + \nabla c(y_k)(S_t(y_k) - y_k)\big) + \frac{1}{2t}\|S_t(y_k) - y_k\|^2$ **do**

  $t \leftarrow \eta t$

end

$t_k \leftarrow t$

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

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

Algorithm 1 behaves differently in convex and nonconvex settings, as one expects. Indeed, we will be able to quantify this change in regime precisely. To this end, observe that since h is closed and convex, equality holds:

$$h(c(x)) = \sup_{w \in \mathbf{R}^n} \{\langle w, c(x)\rangle - h^\star(w)\}. \tag{3}$$

It is the behavior of the functions $x \mapsto \langle w, c(x)\rangle$ that will play a decisive roll. To illustrate the significance of these functions on complexity guarantees, suppose that the function

$$x \mapsto \langle w, c(x)\rangle \quad \text{is convex for any} \quad w \in \operatorname{dom} h^\star. \tag{4}$$

Then equation (3) represents the composition $h \circ c$ as a pointwise supremum of convex functions, and hence $h \circ c$ is itself convex. This setting is already of interest.

Example 2.1 (Sufficient conditions for convexity)**.** Condition (4), in particular, captures convex additive composite minimization and minimization of the pointwise maximum of finitely many convex functions.

- *(Monotone composition)* Suppose that the component functions c_i of c are convex and that h is the separable function $h(y) = \sum_i h_i(y_i)$, with each h_i convex and nondecreasing. For example, the functions $h_i(t) = t$ and $h_i(t) = \max\{0, t\}$ fit the bill. The target problem (2) takes the form

  $$\min_x \; g(x) + \sum_i h_i(c_i(x)).$$

  The domain, $\operatorname{dom} h^\star$, is clearly a subset of $[0, +\infty)^m$ and therefore property (4) holds. In particular, convex additive composite problems satisfy this condition, as well as usual exact penalty formulations of convex mathematical programs [3, Lemma 5.1].

- *(Pointwise maximum)* Suppose that the coordinate functions c_i of c are convex and h is the coordinate maximum function $h(y) = \max\limits_{i=1,\ldots,m} y_i$. Then the target optimization problem (2) amounts to

  $$\min_x \{g(x) + \max_i c_i(x)\}.$$

  The domain of $h^\star$ is the simplex $\{w \geq 0 : \sum_i w_i = 1\}$ and hence condition (4) holds.

In general, even if the function $x \mapsto \langle w, c(x)\rangle$ fails to be convex for some $w \in \operatorname{dom} h^\star$, it will become convex after adding the quadratic perturbation $\frac{L\beta}{2}\|\cdot\|^2$.

Proposition 2.1. *For any $w \in \operatorname{dom} h^\star$, the function $x \mapsto \langle w, c(x)\rangle + \frac{L\beta}{2}\|x\|^2$ is convex.*

Proof. Standard convex analysis shows the bound $\|w\| \leq L$ for all $w \in \operatorname{dom} h^\star$. To see this, note that the subdifferential $\partial h^\star$ is nonempty on a dense subset of $\operatorname{dom} h^\star$. On the other hand, a vector x lies in $\partial h^\star(y)$ if and only if y lies in $\partial h(x)$. Hence the inequality $\|y\| \leq L$ holds whenever $\partial h^\star(y)$ is nonempty.

Thus for any vector $w \in \operatorname{dom} h^\star$ and points $x, y \in \mathbf{R}^n$, we successively deduce

$$\begin{aligned}\langle w, c(y)\rangle &\geq \langle w, c(x) + \nabla c(x)(y - x)\rangle - \frac{L\beta}{2}\|x - y\|^2 \\ &= \langle w, c(x)\rangle + \langle \nabla c(x)^* w, y - x\rangle - \frac{L\beta}{2}\|x - y\|^2\end{aligned}$$

This inequality is exactly the derivative characterization of convexity of the function $x \mapsto \langle w, c(x)\rangle + \frac{L\beta}{2}\|x\|^2$. The result follows. □

Thus there always exists a constant $r \in [0, L\beta]$, such that the function $x \mapsto \langle w, c(x)\rangle + \frac{r}{2}\|x\|^2$ is convex. Intuitively, r measures the extent to which $h \circ c$ fails to be convex. It is this constant r that governs the regime of convergence exhibited by Algorithm 1. We are now ready to state the main result of this paper.

Theorem 2.2 (Convergence guarantees)**.** *Define the constant $\mu := L\beta$, fix a real number $\tilde{\mu} > \mu$, and let x^* be any point satisfying $F(x^*) < F(x_k)$ for all indices k. Let $r \in [0, \mu]$ be a real number such that the function*

$$x \mapsto \langle w, c(x)\rangle + \frac{r}{2}\|x\|^2 \quad \textit{is convex for every } w \in \operatorname{dom} h^\star.$$

Then for any index $N \geq 1$, the inequality holds:

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

Moreover, in the case $r = 0$, the inequality above remains true under the convention $0 = 0 \cdot \infty$ and the following complexity bound on function values holds:

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

Succinctly, setting $\tilde{\mu} = 2\mu$, Theorem 2.2 guarantees the bound

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

The fraction r/μ balances the two terms. In particular, if r/μ is near zero, then the method will be less sensitive to the choice of the constant M, which may have been set superficially at the onset.

The rest of the paper is devoted to proving Theorem 2.2. In large part, our arguments refine the analysis of FISTA [1].

3 Convergence Analysis

Henceforth, we let a_k, y_k, x_k, and v_k be the iterates generated by Algorithm 1. Notice that the points x_k always lie in the domain of g, while y_k and v_k may lie outside of the domain. We begin the analysis with the following two elementary lemmas.

Lemma 3.1. *Consider the point $z := S_{t,\alpha}(y, v)$ for some points $y, v \in \mathbf{R}^n$ and real numbers $t, \alpha > 0$. Then for all $w \in \mathbf{R}^n$ the inequality holds:*

$$\begin{aligned} g(z) + \frac{1}{\alpha}h\big(c(y) + \alpha\nabla c(y)(z-v)\big) \leq\; & g(w) + \frac{1}{\alpha}h\big(c(y) + \alpha\nabla c(y)(w-v)\big) \\ & + \frac{1}{2t}\left(\|w-v\|^2 - \|w-z\|^2 - \|z-v\|^2\right). \end{aligned}$$

Proof. This follows immediately by noting that the function

$$w \mapsto g(w) + \frac{1}{\alpha} h\big(c(y) + \alpha \nabla c(y)(w - v)\big) + \frac{1}{2t} \|w - v\|^2$$

is t^{-1}-strongly convex and z is its minimizer by definition. □

Lemma 3.2. *For any index $k \geq 0$, the inequality $\|x_k - v_k\|^2 \leq \frac{M^2}{a_{k+1}}$ holds.*

Proof. We consider two cases. Suppose first $\|x_k - x_{k-1}\|^2 \leq \frac{M^2 a_k}{(1-a_k)^2}$. Then according to the definition of v_k, we deduce

$$\|x_k - v_k\|^2 = \frac{(1-a_k)^2}{a_k^2} \|x_k - x_{k-1}\|^2,$$

and the result follows. Assuming on the other hand $\|x_k - x_{k-1}\|^2 > \frac{M^2 a_k}{(1-a_k)^2}$, we obtain

$$\|x_k - v_k\|^2 \leq M^2 \leq \frac{M^2}{a_{k+1}},$$

thereby completing the proof. □

Finally, we will need the following key lemma.

Lemma 3.3. *Define $\mu := L\beta$ and fix a real number $\tilde{\mu} > \mu$. Then for any point $x \in \mathbf{R}^n$ and any index k, one has*

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

Proof. Taking into account that h is L-Lipschitzness and ∇c is β-Lipschitz, we obtain

$$h(c(x_k)) \leq h\big(c(y_k) + \nabla c(y_k)(x_k - y_k)\big) + \frac{\mu}{2} \|x_k - y_k\|^2,$$

and hence

$$F(x_k) \leq h\big(c(y_k) + \nabla c(y_k)(x_k - y_k)\big) + g(x_k) + \frac{\mu}{2} \|x_k - y_k\|^2. \tag{5}$$

We now consider two cases. For the simpler case, suppose $\|x_k - x_{k-1}\|^2 \leq \frac{M^2 a_k}{(1-a_k)^2}$. We apply now Lemma 3.1 to $x_k = S_{1/\tilde{\mu},1}(y_k, y_k)$ with $w = a_k x + (1-a_k)x_{k-1}$, yielding the inequalities

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

Combining this inequality with the bound in (5) yields

$$F(x_k) \leq h\big(c(y_k) + \nabla c(y_k)(a_k x + (1-a_k)x_{k-1} - y_k)\big) + a_k g(x) + (1-a_k)g(x_{k-1}) \\ + \frac{\tilde{\mu} a_k^2}{2}\left(\|x - v_{k-1}\|^2 - \|x - v_k\|^2\right) - \frac{\tilde{\mu} - \mu}{2}\|x_k - y_k\|^2, \tag{6}$$

as claimed.

Suppose we are now in the second case $\|x_k - x_{k-1}\|^2 > \frac{M^2 a_k}{(1-a_k)}$. Returning to (5), we apply Lemma 3.1 as above but with $w = a_k v_k + (1-a_k)x_{k-1}$. Similar arithmetic gives the analogous inequality to (6), except where we drop the term $\|w - z\|^2$, namely

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

Taking into account (5), we conclude

$$F(x_k) \leq h\big(c(y_k) + \nabla c(y_k)(a_k v_k + (1-a_k)x_{k-1} - y_k)\big) + a_k g(v_k) + (1-a_k)g(x_{k-1}) \\ + \frac{\tilde{\mu} a_k^2}{2}\|v_k - v_{k-1}\|^2 - \frac{\tilde{\mu} - \mu}{2}\|x_k - y_k\|^2.$$

Note that in contrast to inequality (6), this bound lacks the essential telescoping property in the v_k's. To rectify this, note first the equality $(1-a_k)x_{k-1} - y_k = -a_k v_{k-1}$. We can now apply Lemma 3.1 again with $v_k = S_{1/(\tilde{\mu}a_k),a_k}(y_k, v_{k-1})$ and $w = x$ yielding

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

Thus we obtain

$$F(x_k) \leq h\big(c(y_k) + a_k \nabla c(y_k)(x - v_{k-1})\big) + a_k g(x) + (1-a_k)g(x_{k-1}) \\ + \frac{\tilde{\mu} a_k^2}{2}(\|x - v_{k-1}\|^2 - \|x - v_k\|^2) - \frac{\tilde{\mu} - \mu}{2}\|x_k - y_k\|.$$

Appealing to equality $a_k v_{k-1} = y_k - (1-a_k)x_{k-1}$, we conclude this is precisely the same expression as (6). □

The reader may notice that the proof of the previous lemma did not use the right-hand-side of the inequality $\|x_k - x_{k-1}\|^2 \leq \frac{M^2 a_k}{(1-a_k)^2}$; this expression will be used shortly. We are now ready to prove the main result of the paper.

Proof of Theorem 2.2. We aim to upper bound the right-hand-side of the inequality in Lemma 3.3. For notational convenience, for any point x and index k define

$$z_k^x := c(y_k) + \nabla c(y_k)(a_k x + (1-a_k)x_{k-1} - y_k)$$

and let $w_k^x \in \partial h(z_k^x)$ be any subgradient. Convexity of h implies

$$\begin{aligned} h(z_k^x) - \big(a_k h(c(x)) + (1-a_k)h(c(x_{k-1}))\big) &\leq h(z_k^x) - h(a_k c(x) + (1-a_k)c(x_{k-1})). \\ &\leq \langle w_k^x, z_k^x - (a_k c(x) + (1-a_k)c(x_{k-1}))\rangle \\ &= \langle w_k^x, c(y_k) - c(x_{k-1})\rangle + \langle \nabla c(y_k)^* w_k^x, x_{k-1} - y_k\rangle \\ &\quad + a_k\langle w_k^x, c(x_{k-1}) - c(x)\rangle + a_k\langle \nabla c(y_k)^* w_k^x, x - x_{k-1})\rangle \\ &\leq \frac{r}{2}(a_k\|x - x_{k-1}\|^2 + \|y_k - x_{k-1}\|^2) \\ &\leq \frac{ra_k}{2}(\|x - x_{k-1}\|^2 + a_k\|v_{k-1} - x_{k-1}\|^2). \end{aligned}$$

Combining this inequality with Lemma 3.3, we obtain

$$F(x_k) \leq a_k F(x) + (1-a_k)F(x_{k-1}) + \frac{ra_k}{2}(\|x - x_{k-1}\|^2 + a_k\|v_{k-1} - x_{k-1}\|^2) \tag{7}$$

$$+ \frac{a_k^2\tilde{\mu}}{2}\left(\|x - v_{k-1}\|^2 - \|x - v_k\|^2\right) - \frac{\tilde{\mu} - \mu}{2}\|x_k - y_k\|^2. \tag{8}$$

Let x^* be any point satisfying $F(x^*) < F(x_k)$ for all indices k, and set $x := x^*$ above. Noting by Lemma 3.2 the inequality $\|x_{k-1} - v_{k-1}\|^2 \leq \frac{M^2}{a_k}$ and taking into account $\|x^* - x_{k-1}\| \leq M$, we deduce

$$\frac{ra_k}{2}(\|x^* - x_{k-1}\|^2 + a_k\|v_{k-1} - x_{k-1}\|^2) \leq a_k r M^2.$$

Rewriting (8) by subtracting F^* from both sides and using this bound, we conclude

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

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

$$\begin{aligned} \frac{F(x_N) - F(x^*)}{a_N^2} + \frac{\tilde{\mu}}{2}\|x^* - v_N\|^2 &\leq \frac{1-a_1}{a_1^2}\big(F(x_0) - F(x^*)\big) + \frac{\tilde{\mu}}{2}\|x^* - v_0\|^2 \\ &\quad + rM^2\left(\sum_{j=1}^N \frac{1}{a_j}\right) - \frac{\tilde{\mu} - \mu}{2}\sum_{j=1}^N \frac{\|x_j - y_j\|^2}{a_j^2}. \end{aligned} \tag{9}$$

Note that $F(x_N) - F(x^*) > 0$ and $a_1 = 1$ so that

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

and hence

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

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

$$\frac{\tilde{\mu}-\mu}{2}\min_{j=1,\ldots N}\{\|x_j-y_j\|^2\}\leq\frac{12\tilde{\mu}\|x^*-v_0\|^2}{N(N+1)(2N+1)}+\frac{6rM^2(N+3)}{(N+1)(2N+1)},$$

thereby establishing the first claimed complexity bound in Theorem 2.2.

Finally suppose $r=0$, that is we are in the convex setting. Inequality (9) then becomes

$$\begin{aligned}\frac{F(x_N)-F(x^*)}{a_N^2}&+\frac{\tilde{\mu}}{2}\|x^*-v_N\|^2\\&\leq\frac{1-a_1}{a_1^2}\big(F(x_0)-F(x^*)\big)+\frac{\tilde{\mu}}{2}\|x^*-v_0\|^2-\frac{\tilde{\mu}-\mu}{2}\sum_{j=1}^{N}\frac{\|x_j-y_j\|^2}{a_j^2}.\end{aligned}$$

By noting $\tilde{\mu}-\mu\geq 0$ and $a_1=1$, and dropping terms, we deduce

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

The complexity bound on function values follows immediately. □

References

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

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

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

[4] 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.

[5] E.J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509, 2006.

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

[7] 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).

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

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

[10] 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.

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

[12] 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.

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

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

[15] R.A. Poliquin and R.T. Rockafellar. Amenable functions in optimization. In *Nonsmooth optimization: methods and applications (Erice, 1991)*, pages 338–353. Gordon and Breach, Montreux, 1992.

[16] 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.

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

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

[19] Z. Tan, P. Yang, and A. Nehorai. Joint sparse recovery method for compressed sensing with structured dictionary mismatches. *IEEE Trans. Signal Process.*, 62(19):4997–5008, 2014.

[20] J.A. Tropp. Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Trans. Inform. Theory*, 52(3):1030–1051, 2006.

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

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

[23] H. Zhu, G. Leus, and G.B. Giannakis. Sparsity-cognizant total least-squares for perturbed compressive sampling. *IEEE Trans. Signal Process.*, 59(5):2002–2016, 2011.