Error bounds, quadratic growth, and linear convergence of proximal methods

Dmitriy Drusvyatskiy Adrian S. Lewis

Abstract

We show that the the error bound property, postulating that the step lengths of the proximal gradient method linearly bound the distance to the solution set, is equivalent to a standard quadratic growth condition. We exploit this equivalence in an analysis of asymptotic linear convergence of the proximal gradient algorithm for structured problems, which lack strong convexity. More generally still, we analyze local linear convergence guarantees of a proximal method for minimizing compositions of convex functions with smooth mappings.

1 Introduction

Under favorable conditions, many fundamental optimization algorithms converge linearly: the distance of the iterates to the optimal solution set (the "error") is bounded by a decreasing geometric sequence. Classical optimization literature highlights how quadratic growth properties of the objective function, typically guaranteed through second-order optimality conditions, ensure such linear convergence. Central examples traditionally include the method of steepest descent for smooth minimization [5, Theorem 3.4] and, more abstractly, the proximal point method for nonsmooth convex problems [28, Theorem 2, Proposition 7].

More recent techniques, originally highlighted in the work of Luo and Tseng [17], postulate that the step length at each iteration of the algorithm linearly bounds the error. Such "error bounds" are commonly used in the analysis of first-order methods for strongly convex functions, popular in modern applications such as machine learning and high-dimensional statistics, including in particular the proximal gradient method and its variants; see for example Nesterov [23] and Beck-Teboulle [7]. Convergence analysis based only on the error bound property is appealingly simple even without strong convexity, but the underlying assumption on the optimization problem is opaque, and seems far removed from the classical quadratic growth condition.

Our aim here is to show, in several interesting contemporary optimization frameworks, the equivalence between, on the one hand, the intuitive notion of quadratic growth of the objective function away from the set of minimizers, and on the other

University of Washington, Department of Mathematics, Seattle, WA 98195; www.math.washington.edu/~ddrusv. Research of Drusvyatskiy was partially supported by the AFOSR YIP award FA9550-15-1-0237.

School of Operations Research and Information Engineering, Cornell University, Ithaca, New York, USA; http://people.orie.cornell.edu/~aslewis/. Research supported in part by National Science Foundation Grant DMS-1208338.

hand, the powerful analytic tool furnished by an error bound. Rockafellar already foreshadowed this possibility with his original analysis of the proximal point method [28]. We extend that relationship here to the *proximal gradient method* for problems

$$\min_x \; f(x) + g(x),$$

with g convex and f convex and smooth, and more generally to the *prox-linear algorithm* for convex-composite problems

$$\min_x \; h(c(x)),$$

where h is a finite-valued convex function and c is a smooth mapping. In essence, our analysis depends on viewing these two methods as approximations of the original proximal point algorithm – a perspective of an independent interest. Our assumptions are mild: we rely primarily on a natural strict complementarity condition. In particular, we simplify and extend some of the novel convergence guarantees established in the recent preprint [33] for the prox-gradient method.[1]

The iterative algorithms we consider assign a "gradient-like" step to each potential iterate, as in the analysis of proximal methods in [23, Section 2.1.5]; the step length is zero at stationary points and otherwise serves as a surrogate measure of optimality. For steepest descent, the step is simply a multiple of the negative gradient, for the proximal point method it is determined by a subdifferential relationship, while the prox-gradient and prox-linear methods combine the two. In the language of variational analysis, the existence of an error bound is exactly "metric subregularity" of the gradient-like mapping; see Dontchev-Rockafellar [29]. We will show that subregularity of the gradient-like mapping is equivalent to subregularity of the subdifferential of the objective function itself, thereby allowing us to call on extensive literature relating the quadratic growth of a function to metric subregularity of its subdifferential [10–12,14]. Given the generality of these techniques, we expect that the approach we describe here, rooted in understanding linear convergence through quadratic growth, should extend broadly.

When analyzing the prox-linear algorithm, we encounter a surprise. The error bound condition yields a linear converge rate that is an order of magnitude worse than the analogous rate for the prox-gradient method in the convex setting. The difficulty is that in the nonconvex case, the "linearizations" used by the method do not lower bound the objective function. Nonetheless, we show that the method does converge with a natural rate if the objective function satisfies the stronger condition of quadratic growth that is uniform with respect to tilt-perturbations – a property equivalent to the well-studied notions of tilt-stability [11] and strong metric regularity of the subdifferential [29].

The outline of the manuscript is as follows. Section 2 briefly records some elementary preliminaries. Section 3 contains a detailed analysis of linear convergence of the prox-gradient method for convex functions through the lens of error bounds and quadratic growth. In Section 4, we show that quadratic growth holds in concrete applications under a mild condition of dual strict complementarity; our analysis aims to illuminate and extend some of the results in [33] by dispensing with strong convexity of component functions. Section 5 is dedicated to the local linear convergence of the prox-linear algorithm for minimizing compositions of convex functions with smooth mappings. The final Section 6 explains how a uniform notion of quadratic growth implies linear convergence of the prox-linear method with the natural rate.

[1] While finalizing a first version of this work, the authors became aware of a concurrent, independent and nicely complementary approach [16], based on a related calculus of Kurdyka-Lojasiewicz exponents.

2 Preliminaries

Unless otherwise stated, we follow the terminology and notation of [23,29,30]. Throughout $\mathbf{R}^n$ will denote an n-dimensional Euclidean space with inner-product $\langle \cdot, \cdot \rangle$ and corresponding norm $\| \cdot \|$. The closed unit ball will be written as $\mathbf{B}$, while the open ball of radius r around a point x will be denoted by $B_r(x)$. For any set $Q \subset \mathbf{R}^n$, we define the *distance function*

$$\mathrm{dist}(x; Q) := \inf_{z \in Q} \|z - x\|.$$

The functions we consider will take values in the extended real line $\overline{\mathbf{R}} := \mathbf{R} \cup \{\pm\infty\}$. We call a function $f \colon \mathbf{R}^n \to \overline{\mathbf{R}}$ *closed* if the inequality $\liminf_{x \to \bar{x}} f(x) \geq f(\bar{x})$ holds for any point $\bar{x} \in \mathbf{R}^n$. The symbol $[f \leq \nu] := \{x : f(x) \leq \nu\}$ will denote the ν-sublevel set of f. The Fenchel conjugate of a convex function $f \colon \mathbf{R}^n \to \overline{\mathbf{R}}$ is the closed convex function $f^\star \colon \mathbf{R}^n \to \overline{\mathbf{R}}$ defined by

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

The *subdifferential* of a convex function f at a point x is the set, denoted by $\partial f(x)$, consisting of all vectors $v \in \mathbf{R}^n$ satisfying $f(z) \geq f(x) + \langle v, z - x \rangle$ for all $z \in \mathbf{R}^n$. For any function $f \colon \mathbf{R}^n \to \overline{\mathbf{R}}$ and a real number $t > 0$, we define the *Moreau envelope*

$$g^t(x) := \min_y \left\{ g(y) + \frac{1}{2t}\|y - x\|^2 \right\},$$

and the *proximal mapping*

$$\mathrm{prox}_{tg}(x) := \operatorname*{argmin}_{x \in \mathbf{R}^n} \{g(y) + \frac{1}{2t}\|y - x\|^2\}.$$

A set-valued mapping $F \colon \mathbf{R}^n \rightrightarrows \mathbf{R}^m$ is a mapping assigning to each point $x \in \mathbf{R}^n$ the subset $F(x)$ of $\mathbf{R}^m$. The *graph* of such a mapping is the set

$$\mathrm{gph}\, F := \{(x, y) \in \mathbf{R}^n \times \mathbf{R}^m : y \in F(x)\}.$$

The inverse map $F^{-1} \colon \mathbf{R}^m \rightrightarrows \mathbf{R}^n$ is defined by setting $F^{-1}(y) = \{x : y \in F(x)\}$. Every mapping $F \colon \mathbf{R}^n \rightrightarrows \mathbf{R}^n$ obeys the identity [30, Lemma 12.14]:

$$(I + F)^{-1} + (I + F^{-1})^{-1} = I. \tag{2.1}$$

Note that for any convex function g and a real $t > 0$, equality $\mathrm{prox}_{tg} = (I + t\partial g)^{-1}$ holds.

3 Linear convergence of the prox-gradient method

To motivate the discussion, consider the optimization problem

$$\min_x \varphi(x) := f(x) + g(x) \tag{3.1}$$

where $g \colon \mathbf{R}^n \to \overline{\mathbf{R}}$ is a closed convex function and $f \colon \mathbf{R}^n \to \mathbf{R}$ is a convex C^1-smooth function with a β-Lipschitz continuous gradient:

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta\|x - y\| \qquad \text{for all } x, y \in \mathbf{R}^n.$$

The *proximal-gradient method* is the recurrence

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathbf{R}^n} \; f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2t}\|x - x_k\|^2 + g(x),$$

where the constant $t > 0$ is appropriately chosen. More succinctly, the method can be stated as

$$x_{k+1} = \mathrm{prox}_{tg}(x_k - t\nabla f(x_k)).$$

In order to see the parallel between the proximal gradient method and classical gradient descent for smooth minimization, it is convenient to rewrite the recurrence yet again as $x_{k+1} = x_k - t\mathcal{G}_t(x_k)$ where

$$\mathcal{G}_t(x) := t^{-1}\Big(x - \mathrm{prox}_{tg}(x_k - t\nabla f(x))\Big)$$

is the *prox-gradient mapping*. In particular, equality $\mathcal{G}_t(x) = 0$ holds if and only if x is optimal.

Let S be the set of minimizers of φ and let φ^* be the minimal value of φ. Supposing now $t \le \beta^{-1}$, the following two inequalities are standard [23, Theorem 2.2.7, Corollary 2.2.1] and [7, Lemma 2.3]:

$$\varphi(x_k) - \varphi(x_{k+1}) \ge \frac{1}{2\beta}\|\mathcal{G}_t(x_k)\|^2 \tag{3.2}$$

$$\varphi(x_{k+1}) - \varphi^* \le \langle \mathcal{G}_t(x_k), x_k - x^* \rangle - \frac{1}{2\beta}\|\mathcal{G}_t(x_k)\|^2 \tag{3.3}$$

Here x^* denotes an arbitrary element of S. Hence equation (3.3) immediately implies

$$\varphi(x_{k+1}) - \varphi^* \le \|\mathcal{G}_t(x_k)\|^2 \left(\frac{\|x^* - x_k\|}{\|\mathcal{G}_t(x_k)\|} - \frac{1}{2\beta} \right).$$

Defining $\gamma_k := \frac{\|x^* - x_k\|}{\|\mathcal{G}_t(x_k)\|}$ and using inequality (3.2), along with some trivial algebraic manipulations, yields the geometric decrease guarantee

$$\varphi(x_{k+1}) - \varphi^* \le \left(1 - \frac{1}{2\beta\gamma_k}\right)(\varphi(x_k) - \varphi^*). \tag{3.4}$$

Hence if the quantities γ_k are bounded for all large k, asymptotic Q-linear convergence in function values is assured. This observation motivates the following definition, originating in [17].

Definition 3.1 (Error bound condition)**.** Given real numbers $\gamma, \nu > 0$, we say that the *error bound condition holds with parameters* (γ, ν) if the inequality

$$\mathrm{dist}(x, S) \le \gamma \|\mathcal{G}_t(x)\| \quad \text{is valid for all} \quad x \in [\varphi \le \varphi^* + \nu].$$

Hence we have established the following.

Theorem 3.2 (Linear convergence)**.** *Suppose the error bound condition holds with parameters (γ, ν). Then the proximal gradient method with $t \le \beta^{-1}$ satisfies $\varphi(x_k) - \varphi^* \le \epsilon$ after at most*

$$k \le \frac{\beta}{2\nu}\mathrm{dist}^2(x_0, S) + 2\beta\gamma \ln\left(\frac{\varphi_0 - \varphi^*}{\epsilon}\right) \qquad \textit{iterations.} \tag{3.5}$$

Moreover, if the iterates x_k have some limit point x^*, then x_k asymptotically converge R-linearly, that is there exists an index r such that the inequality

$$\|x_{r+k} - x^*\|^2 \leq \left(1 - \frac{1}{2\beta\gamma}\right)^k C \cdot (\varphi(x_r) - \varphi^*),$$

holds for all $k \geq 1$, where we set $C := \frac{2}{\beta(1-\sqrt{1-(2\beta\gamma)^{-1}})^2}$.

Proof. From the the standard sublinear estimate $\varphi(x_k) - \varphi^* \leq \frac{\beta \cdot \mathrm{dist}^2(x_0,S)}{2k}$ (see e.g. [7, Theorem 3.1]), we deduce that after $k \leq \frac{\beta}{2\nu}\mathrm{dist}^2(x_0, S)$ iterations the inequality $\varphi(x_k) - \varphi^* \leq \nu$ holds. The second summand in (3.5) is then immediate from the linear rate (3.4) and the fact that the values $\varphi(x_k)$ decrease monotonically.

Now suppose that x^* is a limit point of x_k. Note that if an iterate x_r lies in the set $[\varphi \leq \varphi^* + \nu]$, then we have

$$\begin{aligned}\|x_{r+k} - x^*\| &\leq \sum_{i=r+k}^{\infty} \|x_i - x_{i+1}\| \leq \sqrt{2/\beta} \sum_{i=r+k}^{\infty} \sqrt{\varphi(x_i) - \varphi(x_{i+1})} \\ &\leq \sqrt{2/\beta}\sqrt{\varphi(x_r) - \varphi^*} \sum_{i=r+k}^{\infty} \left(1 - \frac{1}{2\beta\gamma}\right)^{\frac{i-r}{2}} \leq \left(1 - \frac{1}{2\beta\gamma}\right)^{k/2} D\sqrt{\varphi(x_r) - \varphi^*},\end{aligned}$$

where we set $D := \frac{\sqrt{2}}{\sqrt{\beta}(1-\sqrt{1-(2\beta\gamma)^{-1}})}$. Squaring both sides, the result follows. □

Convergence guarantees of Theorem 3.2 are expressed in terms of the error bound parameters (γ, ν) – quantities not stated in terms of the initial data of the problem, f and g. Indeed, the error bound condition is a property of the prox-gradient mapping $\mathcal{G}_t(x)$, a nontrivial object to understand. In contrast, in the current work we will show that the error bound condition is simply equivalent to the objective function φ growing quadratically away from its minimizing set S – a familiar, transparent, and largely classical property in nonsmooth optimization.

To gain some intuition, consider the simplest case $g = 0$. Then the prox-gradient method reduces to gradient descent $x_{k+1} = x_k - t\nabla f(x_k)$. Suppose now that f grows quadratically (globally) away from its minimizing set, meaning there is a real number $\alpha > 0$ such that

$$f(x) \geq f^* + \frac{\alpha}{2}\mathrm{dist}^2(x, S) \qquad \text{for all } x \in \mathbf{R}^n.$$

Notice this property is weaker than strong convexity even for C^1-smooth functions; e.g $f(x) = (\max\{|x| - 1, 0\})^2$. Then convexity implies

$$\frac{\alpha}{2}\mathrm{dist}^2(x, S) \leq f(x) - f^* \leq \langle \nabla f(x_k), x^* - x\rangle \leq \|\nabla f(x)\|\|x - x^*\|.$$

Thus the error bound condition holds with parameters $(\gamma, \nu) = (\frac{2}{\alpha}, \infty)$, and the complexity bound of Theorem 3.2 becomes $k \leq \frac{4\beta}{\alpha} \ln\left(\frac{\varphi_0 - \varphi^*}{\epsilon}\right)$. This is the familiar linear rate of gradient descent (up to a constant).

Our goal is to elucidate the quantitative relationship between quadratic growth and the error bound condition in full generality. The strategy we follow is very natural; we will interpret the proximal gradient method as an approximation to the true proximal point algorithm $y_{k+1} = \mathrm{prox}_{t\varphi}(y_k)$ on the function $\varphi = f + g$, and show a linear relationship between the corresponding step sizes (Theorem 3.5). This will allows us to ignore the linearization appearing the definition of the proximal gradient method and focus on the relationship between quadratic growth of φ, properties of the mapping $\mathrm{prox}_{t\varphi}$, and of the subdifferential $\partial\varphi$ (Theorems 3.3 and 3.4). We believe this interpretation of the prox-gradient method is of interest in its own right.

The following is a central result we will need. It establishes a relationship between quadratic growth properties and a "global error bound property" of the function $x \mapsto d(0, \partial\varphi(x))$. Variants of this result have appeared in [1, Theorem 3.3], [4, Theorem 6.1], [10, Theorem 4.3], [12, Theorem 3.1], and [14].

Theorem 3.3 (Subdifferential error bound and quadratic growth)**.** *Consider a closed convex function $h\colon \mathbf{R}^n \to \overline{\mathbf{R}}$ with minimal value h^* and let S be its set of minimizers. Consider the conditions*

$$h(x) \geq h(\bar{x}) + \frac{\alpha}{2} \cdot \mathrm{dist}^2(x; S) \qquad \textit{for all } x \in [h \leq h^* + \nu] \tag{3.6}$$

and

$$\mathrm{dist}(x; S) \leq L \cdot \mathrm{dist}(0; \partial h(x)) \qquad \textit{for all } x \in [h \leq h^* + \nu]. \tag{3.7}$$

If condition (3.6) *holds, then so does condition* (3.7) *with $L = 2\alpha^{-1}$. Conversely, condition* (3.7) *implies condition* (3.6) *with any $\alpha \in (0, \frac{1}{L})$.*

The proof of the implication (3.6) ⇒ (3.7) is identical to the proof of the analogous implication in [1, Theorem 3.3]; the proof of the implication (3.7) ⇒ (3.6) is the same as that of [10, Theorem 4.3], [12, Theorem 3.1]. Hence we omit the arguments.

Given the equality $\mathrm{prox}_{th} = (I + t\partial h)^{-1}$, it is clear that the subdifferential error bound condition (3.7) is related to an analogous property of the proximal mapping. This is the content of the following elementary result.

Theorem 3.4 (Proximal and subdifferential error bounds)**.** *Consider a closed convex function $h\colon \mathbf{R}^n \to \overline{\mathbf{R}}$ with minimal value h^* and let S be its set of minimizers. Consider the conditions*

$$\mathrm{dist}(x; S) \leq L \cdot \mathrm{dist}(0; \partial h(x)) \qquad \textit{for all } x \in [h \leq h^* + \nu]. \tag{3.8}$$

and

$$\mathrm{dist}(x; S) \leq \widehat{L} \cdot t^{-1}\|x - \mathrm{prox}_{th}(x)\| \qquad \textit{for all } x \in [h \leq h^* + \nu]. \tag{3.9}$$

If condition (3.8) *holds, then so does condition* (3.9) *with $\widehat{L} = L + t$. Conversely, condition* (3.9) *implies condition* (3.8) *with $L = \widehat{L}$.*

Proof. Suppose condition (3.8) holds and consider a point $x \in [h \leq h^* + \nu]$. Then clearly the inequality $h(\mathrm{prox}_{th}(x)) \leq h(x) \leq h^* + \nu$ holds. Taking into account the inclusion $t^{-1}(x - \mathrm{prox}_{th}(x)) \in \partial h(\mathrm{prox}_{th}(x))$, we obtain

$$\begin{aligned} \mathrm{dist}(x, S) &\leq \|x - \mathrm{prox}_{th}(x)\| + \mathrm{dist}(\mathrm{prox}_{th}(x), S) \\ &\leq \|x - \mathrm{prox}_{th}(x)\| + L \cdot \mathrm{dist}(0; \partial h(\mathrm{prox}_{th}(x))) \\ &\leq (1 + t^{-1}L)\|x - \mathrm{prox}_{th}(x)\|, \end{aligned}$$

as claimed. Conversely suppose condition (3.9) holds and fix a point $x \in [h \leq h^* + \nu]$. Then for any subgradient $v \in \partial h(x)$, equality $\mathrm{prox}_{th}(x + tv) = x$ holds. Hence, we obtain

$$\mathrm{dist}(x, S) \leq \widehat{L}t^{-1}\|x - \mathrm{prox}_{th}(x)\| \leq \widehat{L}t^{-1}\left\|\mathrm{prox}_{th}(x + tv) - \mathrm{prox}_{th}(x)\right\| \leq \widehat{L}\|v\|,$$

where we have used the fact that the proximal mapping is 1-Lipschitz continuous. Since the subgradient $v \in \partial h(x)$ is arbitrary, the result follows. □

The final step is to relate the step sizes taken by the proximal gradient and the proximal point methods. The ensuing arguments are best stated in terms of monotone operators. To this end, observe that our running problem (3.1) is equivalent to solving the inclusion

$$0 \in \nabla f(x) + \partial g(x).$$

More generally, consider monotone operators $F\colon \mathbf{R}^n \to \mathbf{R}^n$ and $G\colon \mathbf{R}^n \rightrightarrows \mathbf{R}$, meaning that F and G satisfy the inequalities $\langle v_1 - v_2, x_1 - x_2\rangle \geq 0$ and $\langle F(x_1) - F(x_2), x_1 - x_2\rangle \geq 0$ for all $x_i \in \mathbf{R}^n$ and $v_i \in G(x_i)$ with $i = 1, 2$. We now further assume that G is *maximal monotone*, meaning that the graph $\operatorname{gph} G$ is not a proper subset of the graph of any other monotone operator. Along with the operator G and a real $t > 0$, we associate the *resolvent*

$$\operatorname{prox}_{tG} := (I + tG)^{-1}.$$

The mapping $\operatorname{prox}_{tG}\colon \mathbf{R}^n \to \mathbf{R}^n$ is then single-valued and nonexpansive (1-Lipschitz continuous [30, Theorem 12.12]). We aim to solve the inclusion

$$0 \in \Phi(x) := F(x) + G(x)$$

by the *Forward-Backward algorithm*

$$x_{k+1} = \operatorname{prox}_{tG}(x - tF(x)),$$

Equivalently we may write $x_{k+1} = x_k - t\mathcal{G}_t(x_k)$ where $\mathcal{G}_t(x)$ is the *prox-gradient* mapping

$$\mathcal{G}_t(x) := \frac{1}{t}\left(x - \operatorname{prox}_{tG}(x - tF(x))\right).$$

Setting $F = \nabla f$, $G := \partial G$, $\Phi := \partial\varphi$ recovers the proximal gradient method for the problem (3.1).

The following key result shows that the step lengths of the Forward-Backword algorithm and those taken by the proximal point algorithm $z_{k+1} = \operatorname{prox}_{t\Phi}(z_k)$ are proportional.

Theorem 3.5 (Step-lengths comparison)**.** *Let $G\colon \mathbf{R}^n \rightrightarrows \mathbf{R}^n$ be a maximal monotone operator and $F\colon \mathbf{R}^n \to \mathbf{R}^n$ a single-valued mapping, and define $\Phi := F + G$. Then the inequality*

$$\|\mathcal{G}_t(x)\| \leq \operatorname{dist}(0; \Phi(x)) \qquad \textit{holds.}$$

Suppose in addition that F is maximal monotone and β-Lipschitz continuous. Then

$$(1 - \beta t) \cdot \|\mathcal{G}_t(x)\| \leq \|t^{-1}\left(x - \operatorname{prox}_{t\Phi}(x)\right)\| \leq (1 + \beta t) \cdot \|\mathcal{G}_t(x)\| \tag{3.10}$$

Proof. Fix a point $x \in \mathbf{R}^n$ and a vector $v \in \Phi(x)$. Then clearly the inclusion

$$(x - tF(x)) + tv \in x + tG(x) \qquad \text{holds,}$$

or equivalently $x = \operatorname{prox}_{tG}((x - tF(x)) + tv)$. Since the proximal mapping is nonexpensive, we deduce

$$t\|\mathcal{G}_t(x)\| = \|x - \operatorname{prox}_{tG}(x - tF(x))\| \leq t\|v\|.$$

Letting v be the minimal norm element of $\Phi(x)$, we deduce the claimed inequality $\|\mathcal{G}_t(x)\| \leq \operatorname{dist}(0; \Phi(x))$.

Now suppose that F is monotone and β-Lipschitz. Consider a point $x \in \mathbf{R}^n$ and define $z := \mathcal{G}_t(x)$. Observe the chain of equivalences:

$$\begin{aligned} z = \mathcal{G}_t(x) \quad &\Leftrightarrow \quad tz = x - \mathrm{prox}_{tG}(x - tF(x)) \\ &\Leftrightarrow \quad x - tF(x) \in (x - tz) + tG(x - tz) \\ &\Leftrightarrow \quad x + t\left(F(x - tz) - F(x)\right) \in (I + t\Phi)(x - tz) \end{aligned}$$

Define now the vector $w = F(x - tz) - F(x)$ and note $\|w\| \le \beta t\|z\|$. Hence taking into account that resolvents are nonexpansive, we obtain

$$x - tz = \mathrm{prox}_{t\Phi}(x + tw) \subset \mathrm{prox}_{t\Phi}(x) + t\|w\|\mathbf{B},$$

and so deduce

$$z \in \frac{1}{t}\left(x - \mathrm{prox}_{t\Phi}(x)\right) + \beta t\|z\|\mathbf{B}.$$

Hence

$$\Big|\|z\| - t^{-1}\|x - \mathrm{prox}_{t\Phi}(x)\|\Big| \le \|z - t^{-1}\left(x - \mathrm{prox}_{t\Phi}(x)\right)\| \le \beta t\|z\|.$$

The two inequalities in (3.10) follow immediately. □

We now arrive at the main result of this section.

Corollary 3.6 (Error bound and quadratic growth). *Consider a closed, convex function $g\colon \mathbf{R}^n \to \overline{\mathbf{R}}$ and a C^1-smooth convex function $f\colon \mathbf{R}^n \to \mathbf{R}$ with β-Lipschitz continuous gradient. Suppose that the function $\varphi := f + g$ has a nonempty set S of minimizers and consider the following conditions:*

- *(Quadratic growth)*

$$\varphi(x) \ge \varphi^\star + \frac{\alpha}{2}\cdot \mathrm{dist}^2(x; S) \qquad \textit{for all } x \in [\varphi \le \varphi^* + \nu] \tag{3.11}$$

- *(Error bound condition)*

$$\mathrm{dist}(x, S) \le \gamma\|\mathcal{G}_t(x)\| \quad \textit{is valid for all} \quad x \in [\varphi \le \varphi^* + \nu], \tag{3.12}$$

Then property (3.11) *implies property* (3.12) *with $\gamma = (2\alpha^{-1} + t)(1 + \beta t)$. Conversely, condition* (3.12) *implies condition* (3.11) *with any $\alpha \in (0, \gamma^{-1})$.*

Proof. Suppose condition (3.11) holds. Then for any $x \in [\varphi \le \varphi^* + \nu]$, we deduce

$$\begin{aligned} \mathrm{dist}(x, S) &\le 2\alpha^{-1}\cdot \mathrm{dist}(0; \partial\varphi(x)) && \text{(Theorem 3.3)} \\ &\le (2\alpha^{-1} + t)\|t^{-1}(x - \mathrm{prox}_{t\varphi}(x))\| && \text{(Theorem 3.4)} \\ &\le (2\alpha^{-1} + t)(1 + \beta t)\|\mathcal{G}_t(x)\| && \text{(Inequality 3.10)} \end{aligned}$$

This establishes (3.12) with $\gamma = (2\alpha^{-1} + t)(1 + \beta t)$. Conversely suppose (3.12) holds. Then for any $x \in [\varphi \le \varphi^* + \nu]$ we deduce using Theorem 3.5 the inequality $\mathrm{dist}(x, S) \le \gamma\|\mathcal{G}_t(x)\| \le \gamma \cdot \mathrm{dist}(0, \partial\varphi(x))$. An application of Theorem 3.3 completes the proof. □

The following convergence result is now immediate from Theorem 3.2 and Corollary 3.6. Notice that the complexity bound matches (up to a constant) the linear rate of convergence of the proximal gradient method when applied to strongly convex functions.

Corollary 3.7 (Quadratic growth and linear convergence)**.** *Consider a closed, convex function $g\colon \mathbf{R}^n \to \overline{\mathbf{R}}$ and a C^1-smooth function $f\colon \mathbf{R}^n \to \mathbf{R}$ with β-Lipschitz continuous gradient. Suppose that the function $\varphi := f + g$ has a nonempty set S of minimizers and that the quadratic growth condition holds:*

$$\varphi(x) \geq \varphi^* + \frac{\alpha}{2} \cdot \operatorname{dist}^2(x; S) \qquad \text{for all } x \in [\varphi \leq \varphi^* + \nu].$$

Then the proximal gradient method with $t \leq \beta^{-1}$ satisfies $\varphi(x_k) - \varphi^* \leq \epsilon$ after at most

$$k \leq \frac{\beta}{2\nu} \operatorname{dist}^2(x_0, S) + 12 \cdot \frac{\beta}{\alpha} \ln\left(\frac{\varphi_0 - \varphi^*}{\epsilon}\right) \qquad \text{iterations.}$$

4 Quadratic growth in structured optimization

Recently, the authors of [33] proved that the error bound condition holds under very mild assumptions, thereby explaining asymptotic linear converge of the proximal gradient method often observed in practice. In this section, we aim to use the equivalence between the error bound condition and quadratic growth, established in Theorem 3.6, to streamline and illuminate the arguments in [33], while also extending their results to a wider setting. To this end, consider the problem

$$\min_x \ \varphi(x) := f(Ax) + g(x)$$

where $f\colon \mathbf{R}^m \to \mathbf{R}$ is convex and C^1-smooth, $g\colon \mathbf{R}^n \to \overline{\mathbf{R}}$ is closed and convex, and $A\colon \mathbf{R}^m \to \mathbf{R}^n$ is a linear mapping. We assume that g is proper, meaning that its domain is nonempty, and that the primal objective φ is bounded below. Consider now the Fenchel dual problem

$$\max_y -\Psi(y) := -f^\star(y) - g^\star(-A^T y). \tag{4.1}$$

Since f is C^1-smooth, the conjugate $f^\star$ is essentially strictly convex and therefore the dual has a unique maximizer $\bar{y}$, characterized by the inclusion $0 \in \partial\Psi(\bar{y})$. Notice now that the primal optimality conditions for a point $\bar{x}$ read $0 \in A^T \nabla f(A\bar{x}) + \partial g(\bar{x})$, or equivalently $0 \in \partial f^\star(\nabla f(A\bar{x})) - A\partial g^\star(-A^T \nabla f(A\bar{x})) \subseteq \partial\Psi(\nabla f(A\bar{x}))$. Hence if $\bar{x}$ is a minimizer of the primal, then $\nabla f(A\bar{x})$ is the unique maximizer of the dual. We record this observation in the following lemma. Throughout, we let S be the set of minimizers of φ.

Lemma 4.1. *The mapping $x \mapsto \nabla f(Ax)$ is constant on the solution set S.*

To make progress, we now impose the following mild assumptions on the dual problem:

1. (dual nondegeneracy) $\quad 0 \in A^T(\operatorname{dom} f^\star) + \operatorname{ri}(\operatorname{dom} g^\star)$,
2. (dual strict complementarity) $\quad 0 \in \operatorname{ri} \partial\Psi(\bar{y})$.

Taken together, these two standard conditions immediately imply

$$\begin{aligned} 0 \in \operatorname{ri} \partial\Psi(\bar{y}) &= \operatorname{ri}\left(\partial f^\star(\bar{y}) - A\partial g^\star(-A^T\bar{y})\right) \\ &= \operatorname{ri} \partial f^\star(\bar{y}) - A\left(\operatorname{ri} \partial g^\star(-A^T\bar{y})\right), \end{aligned} \tag{4.2}$$

where the last equality follows for example from [27, Theorem 6.6]. To elucidate the impact this condition has on error bounds, recall that we must estimate the distance

$\operatorname{dist}(x,S)$ for a putative solution x. Observe that Lemma 4.1 directly implies that S admits the description

$$S = \partial g^\star(-A^T\bar{y}) \cap A^{-1}\partial f^\star(\bar{y}).$$

The inclusion (4.2), combined with [27, Theorem 6.7], guarantees that the relative interiors of the two sets $\partial g^\star(-A^T\bar{y})$ and $A^{-1}\partial f^\star(\bar{y})$ meet and hence for any compact set $\mathcal{X} \subset \mathbf{R}^n$ there exists a constant $\kappa \geq 0$ satisfying

$$\operatorname{dist}(x,S) \leq \kappa\Big(\operatorname{dist}(x,\partial g^\star(-A^T\bar{y})) + \operatorname{dist}(Ax,\partial f^\star(\bar{y}))\Big) \qquad \text{for all } x \in \mathcal{X}. \quad (4.3)$$

This type of an inequality is often called linear regularity; see for example [6, Corollary 4.5]. The final assumption we need to deduce quadratic growth of φ, not surprisingly, is a quadratic growth assumption on the individual functions f and g after tilt perturbations.

Definition 4.2 (Firm convexity)**.** A closed convex function $h\colon \mathbf{R}^n \to \overline{\mathbf{R}}$ is *firmly convex relative to a vector* $v \in \mathbf{R}^n$ if the tilted function $h_v(x) := h(x) - \langle v, x\rangle$ satisfies the quadratic growth condition: for any compact set $\mathcal{X} \subset \mathbf{R}^n$ there is a constant α satisfying

$$h_v(x) \geq (\inf h_v) + \frac{\alpha}{2}\operatorname{dist}^2(x,(\partial h_v)^{-1}(0)) \qquad \text{for all } x \in \mathcal{X}.$$

We say that h is *firmly convex* if h is firmly convex relative to any vector $v \in \mathbf{R}^n$.

We are now ready to prove the main theorem of this section; note that unlike in [33], we do not require strong convexity of the function f. This generalization is convenient since it allows to capture "robust" formulations where f is a translate of the Huber penalty or its asymmetric extensions.

Theorem 4.3 (Quadratic growth in composite optimization)**.** *Consider a closed, convex function $g\colon \mathbf{R}^n \to \overline{\mathbf{R}}$ and a C^1-smooth convex function $f\colon \mathbf{R}^m \to \mathbf{R}$. Suppose that the function $\varphi(x) := f(Ax) + g(x)$ has a nonempty set S of minimizer and let $\bar{y}$ be the optimal solution of the dual problem* (4.1)*. Suppose the conditions hold:*

1. *(Compactness) The solution set S is bounded.*
2. *(Dual nondegeneracy and strict complementarity) Assumptions 1 and 2.*
3. *(Quadratic growth of components) The functions f and g are firmly convex relative to $\bar{y}$ and $-A^T\bar{y}$, respectively.*

Then the error bound condition holds with some parameters (γ,ν).

Proof. Since S is compact, all sublevel sets of φ are compact. Choose a number $\nu > 0$ and set $\mathcal{X} := [\varphi \leq \varphi^* + \nu]$ and $\mathcal{Y} = \operatorname{cl} A(\mathcal{X})$. Le $\bar{x} \in S$ be arbitrary and note the equality $\bar{y} = \nabla f(A\bar{x})$. Then observing that $A\bar{x}$ minimizes $f(\cdot) - \langle \bar{y}, \cdot\rangle$ and $\bar{x}$ minimizes $g(\cdot) + \langle A^T\bar{y}, \cdot\rangle$, property (3) (Quadratic growth of components) guarantees that there exist constants $c, \alpha \geq 0$ such that

$$f(y) \geq f(A\bar{x}) + \langle \bar{y}, y - A\bar{x}\rangle + \frac{c}{2}\operatorname{dist}^2(y,(\partial f)^{-1}(\bar{y})) \qquad \text{for all } y \in \mathcal{Y},$$

and

$$g(x) \geq g(\bar{x}) + \langle -A^T\bar{y}, x - \bar{x}\rangle + \frac{\alpha}{2}\operatorname{dist}^2(x,(\partial g)^{-1}(-A^T\bar{y})) \qquad \text{for all } x \in \mathcal{X}.$$

Letting κ be the constant from (4.3), and setting $y := Ax$ above we deduce

$$\begin{aligned}\varphi(x) = f(Ax) + g(x) &\geq \Big(f(A\bar{x}) + \langle \bar{y}, Ax - A\bar{x}\rangle + \frac{c}{2}\mathrm{dist}^2(Ax, \partial f^\star(\bar{y}))\Big) + \\ &+ \Big(g(\bar{x}) + \langle -A^T\bar{y}, x - \bar{x}\rangle + \frac{\alpha}{2}\mathrm{dist}^2(x, \partial g^\star(-A^T\bar{y}))\Big) \\ &\geq \varphi(\bar{x}) + \frac{\max\{\alpha, c\}}{6\kappa^2}\mathrm{dist}^2(x, S).\end{aligned}$$

This completes the proof. □

Notice that firm convexity requires a certain inequality to hold on compact sets $\mathcal{X}$, rather than on sublevel sets. In any case, firm convexity is intimately tied to error bounds. For example, analogously to Theorem 3.3, one can show that h is firmly convex relative to v if and only if for any compact set $\mathcal{X}$ there exists a constant $L \geq 0$ satisfying

$$\mathrm{dist}(x, (\partial h)^{-1}(v)) \leq L \cdot \mathrm{dist}(v, \partial h(x)) \qquad \text{for all } x \in \mathcal{X}.$$

Indeed this is implicitly shown in the proof of Theorem [1, Theorem 3.3], for example. Moreover, the same argument as in Theorem 3.4 shows that h is firmly convex relative to v if and only if for any compact set $\mathcal{X}$ there exists a constant $\widehat{L} \geq 0$ satisfying

$$\mathrm{dist}(x; S) \leq \widehat{L} \cdot \|t^{-1}(x - \mathrm{prox}_{th}(x))\| \qquad \text{for all } x \in \mathcal{X}.$$

The class of firmly convex functions is large, including for example all strongly convex functions and polyhedral functions. More generally, all convex Piecewise Linear Quadratic (PLQ) functions [30, Section 10.20] are firmly convex, since their subdifferential graphs are finite unions of polyhedra. Indeed, the subclass of affinely composed PLQ penalties [30, Example 11.18] is ubiquitous in optimization. These are functions of the form

$$h(x) := \sup_{z \in Z}\ \langle Bx - b, z\rangle - \langle Az, z\rangle$$

where Z is a polyhedron, B is a linear map, and A is a positive-semidefinite matrix. For more details on the PLQ family, see [2,3]. For example, the elastic net penalty [34], used for group detection, and the soft-insensitive loss [9], used for training Support Vector Machines, fall within this class.

Note that the assumptions of dual nondegeneracy and strict complementarity (Assumptions 1 and 2) were only used in the proof Theorem 4.3 to guarantee inequality (4.3). On the other hand, this inequality holds automatically if the subdifferentials $\partial g^\star(-A^T\bar{y})$ and $\partial f^\star(\bar{y})$ are polyhedral—a common situation.

Corollary 4.4 (Quadratic growth without strict complementarity)**.** *Consider a convex PLQ function $g\colon \mathbf{R}^n \to \overline{\mathbf{R}}$ and a C^1-smooth convex function $f\colon \mathbf{R}^m \to \mathbf{R}$. Suppose that the function $\varphi(x) := f(Ax) + g(x)$ has a nonempty compact set S of minimizers and that either f is strictly convex or f is PLQ. Then the error bound condition holds with some parameters (γ, ν).*

Proof. Since g is PLQ, the subdifferential $\partial g^\star$ at any point is polyhedral. Similarly, if f is PLQ then $\partial f^\star$ is polyhedral at any point, while if f is strictly convex, the subdifferential $\partial f^\star(\bar{y})$ is a singleton. Thus in all cases the inequality (4.3) holds and the proof proceeds as in Theorem 4.3. □

Firm convexity is preserved under separable sums.

Lemma 4.5 (Separable sum)**.** *Consider a family of functions $f_i \colon \mathbf{R}^{n_i} \to \overline{\mathbf{R}}$ for $i = 1, \ldots, m$ with each f_i firmly convex relative to some $v_i \in \mathbf{R}^{n_i}$. Then the separable function $f : \mathbf{R}^{n_1 + \cdots + n_m} \to \overline{\mathbf{R}}$ defined by $f(x) = \sum_{i=1}^m f_i(x_i)$ is firmly convex relative to the vector $(v_1, \ldots, v_n)$.*

Proof. The proof is immediate from definitions. □

Moreover, firmly convex functions are preserved by the Moreau envelope.

Theorem 4.6 (Moreau envelope)**.** *Consider a function $h \colon \mathbf{R}^n \to \overline{\mathbf{R}}$ that is firmly convex relative to a vector v. Then the Moreau envelope h^t is itself firmly convex relative to v.*

Proof. Define the tilted functions $h_v(x) := h(x) - \langle v, x\rangle$ and $(h^t)_v(x) := h^t(x) - \langle v, x\rangle$. Observe

$$\begin{aligned}(h^t)_v(x) &= \min_y \left\{ h(y) + \frac{1}{2t}\|y - x\|^2 - \frac{1}{t}\langle tv, x\rangle \right\} \\ &= \min_y \left\{ h(y) - \langle v, y\rangle + \frac{1}{2t}\|y - (x - tv)\|^2 - \frac{t}{2}\|v\|^2 \right\} \\ &= (h_v)^t(x - tv) - \frac{t}{2}\|v\|^2.\end{aligned}$$

Since firm convexity is invariant under translation of the domain, it is now sufficient to show that $(h_v)^t$ is firmly convex relative to the zero vector. To this end, let S be the set of minimizers of h_v, or equivalently the set of minimizers of $(h_v)^t$. Since h is firmly convex relative to v, for any compact set $\mathcal{Z} \subset \mathbf{R}^n$, there exists a constant $\widehat{L} \geq 0$ so that

$$\operatorname{dist}(z, S) \leq \widehat{L}\|t^{-1}(z - \operatorname{prox}_{th_v}(z))\| \qquad \text{for all } z \in \mathcal{Z}.$$

Taking into account the equation

$$\nabla (h_v)^t(z) = t^{-1}[z - \operatorname{prox}_{th_v}(z)],$$

we deduce $\operatorname{dist}(z, S) \leq \widehat{L} \cdot \|\nabla (h_v)^t(z)\|$ for all $z \in \mathcal{Z}$, thereby completing the proof. □

In summary, all typical smooth penalties (e.g. square l_2-norm, logistic loss), polyhedral functions (e.g. l_1 and l_∞-penalties, vapnik, hinge loss, check function, total variation penalty), Moreau envelopes of polyhedral functions (e.g. Huber and quantile huber [3]), and general affinely composed PLQ penalties (e.g. soft-insensitive loss [9], elastic net [34]) are firmly convex. Another important example is the nuclear norm [33].

5 Prox-linear algorithm

Our next algorithmic illustration targets optimization problems of the form

$$\min_x \varphi(x) := h(c(x)), \tag{5.1}$$

where $h \colon \mathbf{R}^m \to \mathbf{R}$ is an L-Lipschitz continuous convex function and $c \colon \mathbf{R}^n \to \mathbf{R}^m$ is a C^1-smooth mapping whose Jacobian $\nabla c(x)$ is β-Lipschitz continuous. Since such problems are typically nonconvex, we seek point x that are only *first-order stationary*, meaning that the directional derivate of φ at x is nonnegative in all directions. The directional derivate of φ is exactly the support function of the *subdifferential set*

$$\partial\varphi(x) := \nabla c(x)^T \partial h(c(x)),$$

and hence stationary of x simply amounts to the inclusion $0 \in \partial\varphi(x)$.

To specify the algorithm we study, define for any points $x, y \in \mathbf{R}^n$ the linearized function

$$\varphi(x; y) := h\big(c(x) + \nabla c(x)(y - x)\big),$$

and for any real $t > 0$ consider the quadratic perturbation

$$\varphi_t(x; y) := \varphi(x; y) + \frac{1}{2t}\|x - y\|^2.$$

Note that the function $\varphi(x; \cdot)$ is always convex, even though φ typically is not convex. Let x^t be the minimizer of the proximal subproblem

$$x^t := \operatorname*{argmin}_y \ \varphi_t(x; y).$$

It is immediate that the linearized function $\varphi(x; \cdot)$ is quadratically close to φ itself:

$$-\frac{L\beta}{2}\|x - y\|^2 \leq \varphi(y) - \varphi(x; y) \leq \frac{L\beta}{2}\|x - y\|^2. \tag{5.2}$$

In particular, $\varphi_t(x, \cdot)$ is a quadratic upper estimator of φ for any $t \leq (L\beta)^{-1}$. We now define the prox-gradient mapping in the natural way

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

It is easily verified that equality $\mathcal{G}_t(x) = 0$ holds if and only if x is stationary for φ. In this section, we consider the well-known Algorithm 1, recently studied for example in [15]. The ideas behind the method (and its trust-region variants) go back a long time, e.g. [8, 13, 25, 26, 31, 32]; see [8] for a historical discussion.

Algorithm 1: ProxDescent for finite convex h

```
Data: A point $x_1 \in \mathbf{R}^n$, and constants $q \in (0,1)$, $t > 0$, and $\sigma > 0$.
$k \leftarrow 1$
while $\|\mathcal{G}_t(x_k)\| > \epsilon$ do
    while $\varphi(x_k^t) > \varphi(x_k) - \frac{\sigma}{2}\|\mathcal{G}_t(x_k)\|^2$ do
        $t \leftarrow qt$                                  [Backtracking line search]
    $x_{k+1} \leftarrow x_k^t$                                     [Iterate update]
    $k \leftarrow k + 1$
return $x_k$
```

Note that the prox-gradient method in Section 3 for minimizing the sum $\min_x f(x) + g(x)$ is an example of Algorithm 1 with the decomposition $c(x) = (f(x), x)$ and $h(t, z) = t + g(z)$. Therefore we now perform an analysis following the same strategy as for the proximal gradient method; there are important and surprising differences, however, both in the conclusions we make and in the proof techniques. We begin with the following lemma; the proof follows that of [23, Lemma 2.3.2].

Lemma 5.1 (Gradient inequality)**.** *For all points $x, y \in \mathbf{R}^n$, the inequality*

$$\varphi(x;y) \geq \varphi_t(x;x^t) + \langle \mathcal{G}_t(x), y - x\rangle + \frac{t}{2}\|\mathcal{G}_t(x)\|^2, \tag{5.3}$$

holds, and consequently we have

$$\varphi(y) \geq \varphi(x^t) + \langle \mathcal{G}_t(x), y - x\rangle + \frac{t}{2}(2 - L\beta t)\|\mathcal{G}_t(x)\|^2 - \frac{L\beta}{2}\|x - y\|^2. \tag{5.4}$$

and

$$\varphi(x) \geq \varphi(x^t) + \frac{t}{2}(2 - L\beta t)\|\mathcal{G}_t(x)\|^2. \tag{5.5}$$

Proof. Noting that the function $\varphi_t(x;y) := \varphi(x;y) + \frac{1}{2t}\|y - x\|^2$ is strongly convex in the variable y, we deduce

$$\begin{aligned}\varphi(x;y) &= \varphi_t(x;y) - \frac{1}{2t}\|y - x\|^2 \\ &\geq \varphi_t(x;x^t) + \frac{1}{2t}\left(\|y - x^t\|^2 - \|y - x\|^2\right) \\ &= \varphi_t(x;x^t) + \langle \mathcal{G}_t(x), y - x\rangle + \frac{t}{2}\|\mathcal{G}_t(x)\|^2,\end{aligned}$$

establishing (5.3). Inequality (5.4) follows by combining (5.2) and (5.3). Finally, we obtain inequality (5.5) from (5.4) by setting $y = x$. □

For simplicity, we assume that the constants L and β are known and we set $t \leq \frac{1}{L\beta}$ and $\sigma = \frac{1}{L\beta}$ in Algorithm 1, so that the line search always accepts the initial step. The more general setting with the backtracking line-search is entirely analogous. Observe now that the inequality (5.5) yields

$$\varphi(x_{k+1}) \leq \varphi(x_k) - \frac{1}{2L\beta}\|\mathcal{G}_t(x_k)\|^2,$$

and hence we obtain the global convergence rate

$$\min_{i=1,\ldots,k} \|\mathcal{G}_t(x_i)\|^2 \leq \frac{2L\beta}{k}\sum_{i=1}^{k} \varphi(x_k) - \varphi(x_{k+1}) = \frac{2L\beta(\varphi(x_1) - \varphi^*)}{k},$$

where we set $\varphi^* := \lim_{k\to\infty} \varphi(x_k)$. In particular, any limit point of the sequence x_k is stationary for the target problem (5.1).

Now suppose that x^* is a limit point of the sequence x_k. Appealing to inequality (5.4) with $y = x^*$, we deduce

$$\varphi(x_{k+1}) - \varphi(x^*) \leq \|\mathcal{G}_t(x_k)\|^2\left(\frac{\|x_k - x^*\|}{\|\mathcal{G}_t(x_k)\|} + \frac{L\beta}{2}\frac{\|x_k - x^*\|^2}{\|\mathcal{G}_t(x_k)\|^2} + \frac{t}{2}(L\beta t - 2)\right).$$

Defining $\gamma_k := \max\{1, \|x - x^*\|/\|\mathcal{G}_t(x)\|\}$ and appealing to (5.5), we deduce the geometric decay

$$\varphi(x_{k+1}) - \varphi(x^*) \leq \left(1 - \frac{1}{L\beta(1 + L\beta)\gamma_k^2}\right)(\varphi(x_k) - \varphi(x^*)).$$

Hence provided that γ_k are bounded for all large k, the function values asymptotically converge Q-linearly. This motivates the following definition, akin to Definition 3.1.

Definition 5.2 (Error bound condition)**.** We say that the *error bound condition holds around a point* $\bar{x}$ *with parameter* $\gamma > 0$ if there exist $\epsilon > 0$ so that the inequality

$$\mathrm{dist}(x, (\partial\varphi)^{-1}(0)) \leq \gamma\|\mathcal{G}_t(x)\| \quad \text{is valid for all} \quad x \in B_\epsilon(\bar{x}).$$

Hence we arrive at the following convergence guarantee.

Theorem 5.3 (Linear convergence of the proximal method)**.** *Consider the sequence* x_k *generated by Algorithm 1 with* $t \leq (L\beta)^{-1}$*, and suppose that* x_k *has some limit point* x^* *around which the error bound condition holds with parameter* $\gamma > 0$*. Then for all large* k*, function values converge Q-linearly*

$$\varphi(x_{k+1}) - \varphi(x^*) \leq \left(1 - \frac{1}{L\beta(1+L\beta)\gamma^2}\right)(\varphi(x_k) - \varphi(x^*)),$$

while the points x_k *asymptotically converge R-linearly, meaning there exists an index* r *such that the inequality*

$$\|x_{r+k} - x^*\|^2 \leq \left(1 - \frac{1}{L\beta(1+L\beta)\gamma^2}\right)^k C \cdot (\varphi(x_r) - \varphi^*),$$

holds for all $k \geq 1$*, where we set* $C := \frac{2}{L\beta(1-\sqrt{1-(L\beta(1+L\beta)\gamma^2)^{-1}})^2}$.

Proof. Let $\epsilon > 0$ be as in Definition 5.2. We aim to show that if x_r is sufficiently close to x^*, then all following iterates never leave the ball $B_\epsilon(x^*)$. To this end, let r be an index such that x_r lies in $B_\epsilon(x^*)$ and let $k \geq 1$ be the smallest index satisfying $x_{r+k} \notin B_\epsilon(x^*)$. Defining $\zeta := L\beta(1+L\beta)\gamma^2$, we deduce

$$\begin{aligned}\|x_k - x_r\| &\leq \sum_{i=r}^{k-1}\|x_i - x_{i+1}\| \leq \sqrt{2/(L\beta)}\sum_{i=r}^{k-1}\sqrt{\varphi(x_i) - \varphi(x_{i+1})} \\ &\leq \sqrt{2/(L\beta)}\sqrt{\varphi(x_r) - \varphi(x^*)}\sum_{i=r}^{k}\left(1 - \frac{1}{\zeta}\right)^{\frac{i-r}{2}} \leq \sqrt{\frac{2(\varphi(x_r) - \varphi(x^*))}{L\beta(1-\zeta^{-1})}}.\end{aligned}$$

Hence if x_r lies in the ball $B_{\epsilon/2}(x^*)$ and is sufficiently close to x^* so that the right-hand-side is smaller than $\frac{\epsilon}{2}$, we obtain a contradiction. Thus there exists an index r so that for all $k \geq r$, the iterates x_k lie in $B_\epsilon(x^*)$. The claimed Q-Linear rate follows immediately. To obtain the R-linear rate of the iterates, we argue as in the proof of Theorem 3.2:

$$\begin{aligned}\|x_{r+k} - x^*\| &\leq \sum_{i=r+k}^{\infty}\|x_i - x_{i+1}\| \leq \sqrt{2/(L\beta)}\sum_{i=r+k}^{\infty}\sqrt{\varphi(x_i) - \varphi(x_{i+1})} \\ &\leq \sqrt{2/(L\beta)}\sqrt{\varphi(x_r) - \varphi(x^*)}\sum_{i=r+k}^{\infty}\left(1 - \frac{1}{\zeta}\right)^{\frac{i-r}{2}} \leq \left(1 - \frac{1}{\zeta}\right)^{k/2} D\sqrt{\varphi(x_r) - \varphi^*},\end{aligned}$$

where $D = \frac{\sqrt{2}}{\sqrt{L\beta}(1-\sqrt{1-\zeta^{-1}})}$. Squaring both sides the result follows. □

Theorem 5.3 already marks a point of departure from the convex setting: the rate of convergence is an order of magnitude worse than one would expect, in light of the convergence guarantees in Theorem 3.2 and inequality (3.4). The difference is the lack of convexity; the linearizations $\varphi(x; \cdot)$ no longer lower bound the objective function φ, but only do so up to a quadratic deviation, thereby leading to a worse linear rate of convergence. We put this issue aside for the moment and will revisit it in Section 6, where we will show that Algorithm 1 accelerates under a natural *uniform* quadratic growth condition. The ensuing discussion relating the error bound condition and quadratic growth will drive that analysis as well.

Following the pattern of the current work, it is natural to interpret the error bound condition in terms of a natural property of the subdifferential $\partial\varphi$. To this end, as in Section 3, we first prove that the step-lengths of Algorithm 1 are proportional to the step-lengths of the proximal point method $z_{k+1} \in (I + t\partial\varphi)^{-1}(z_k)$. Though the arguments deviate significantly from those in Section 3, the main conclusions remain the same.

Theorem 5.4 (Step-lengths comparison)**.** *Consider a finite convex function $h\colon \mathbf{R}^m \to \mathbf{R}$ and a C^1-smooth mapping $c\colon \mathbf{R}^n \to \mathbf{R}^m$. Define the composite function $\varphi := h \circ c$. Then the inequality*

$$\frac{1}{2}\|\mathcal{G}_t(x)\| \leq \operatorname{dist}(0; \partial\varphi(x)) \qquad \textit{holds.} \tag{5.6}$$

Suppose in addition that h is L-Lipschitz continuous and ∇c is β-Lipschitz continuous, and set $r := L\beta$. Then for $t < r^{-1}$ and any point $x^+ \in (I + t\partial\varphi)^{-1}(x)$ the inequality holds:

$$1 + (1 - tr)^{-1} \geq \frac{\|x^t - x\|}{\|x^+ - x\|} + \frac{\|x^+ - x\|}{\|x^t - x\|}.$$

Proof. Fix a vector $v \in \partial\varphi(x)$. Then there exists $w \in \partial h(c(x))$ satisfying $v = \nabla c(x)^* w$. Convexity yields

$$h\big(c(x) + \nabla c(x)(x^t - x)\big) \geq h(c(x)) + \langle w, \nabla c(x)(x^t - x)\rangle = h(c(x)) + \langle v, x^t - x\rangle.$$

Appealing to the inequality $\varphi(x) \geq \varphi(x; x^t) + \frac{1}{2t}\|x - x^t\|^2$, we deduce $\|v\| \cdot \|x^t - x\| \geq \frac{1}{2t}\|x^t - x\|^2$, completing the proof of (5.6).

Next, again fix a point x and a subgradient $v = \nabla c(x)^* w$ for some $w \in \partial h(c(x))$. Then for all $y \in \mathbf{R}^n$ we successively deduce

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

Let x^+ be a critical point of the function $\varphi(\cdot) + \frac{1}{2t}\|\cdot - x\|^2$. Set $r := L\beta$ and observe then

$$\begin{aligned}
\varphi(y) &\geq \varphi(x^+) + \langle t^{-1}(x - x^+), y - x^+\rangle - \frac{L\beta}{2}\|y - x^+\|^2 \\
&= \varphi(x^+) + \frac{1}{2t}\|x^+ - x\|^2 + \frac{t^{-1} - r}{2}\|x^+ - y\|^2 - \frac{1}{2t}\|y - x\|^2
\end{aligned}$$

Plugging in $y = x^t$, we deduce

$$\begin{aligned}
\frac{t^{-1} - r}{2}\|x^+ - x^t\|^2 &\leq \big(\varphi(x^t) - \frac{1}{2t}\|x^t - x\|^2\big) - \big(\varphi(x^+) + \frac{1}{2t}\|x^+ - x\|^2\big) \\
&\leq \varphi_t(x; x^t) - \varphi_t(x; x^+) + \frac{r}{2}\left(\|x^t - x\|^2 + \|x^+ - x\|^2\right).
\end{aligned}$$

Taking into account the strong convexity inequality $\varphi_t(x;x^+) \geq \varphi_t(x;x^t) + \frac{1}{2t}\|x^+ - x^t\|^2$, we deduce

$$(2t^{-1} - r)\|x^+ - x^t\|^2 \leq r\left(\|x^t - x\|^2 + \|x^+ - x\|^2\right).$$

A short computation then shows

$$1 + (1 - tr)^{-1} \geq \frac{\|x^t - x\|}{\|x^+ - x\|} + \frac{\|x^+ - x\|}{\|x^t - x\|},$$

and the result follows. □

In the above theorem, the assumption $t < (L\beta)^{-1}$ seems unavoidable. In essence, the difficulty is that the base-point x at which the lengths comparison is made remains fixed. This difficulty will not impact our main result Corollary 5.10, whose proof relies on Ekeland's variational principle. We now record the following key definition.

Definition 5.5 (Subregularity)**.** A set-valued mapping $T\colon \mathbf{R}^n \rightrightarrows \mathbf{R}^m$ is *subregular* at $(\bar{x}, \bar{y}) \in \operatorname{gph} T$ with constant $l > 0$ if there exists a neighborhood $\mathcal{X}$ of $\bar{x}$ satisfying

$$\operatorname{dist}\left(x; T^{-1}(\bar{y})\right) \leq l \cdot \operatorname{dist}\left(\bar{y}; T(x)\right) \qquad \text{for all } x \in \mathcal{X}.$$

In this language, the error bound property around a stationary point $\bar{x}$ of φ with parameter γ amounts to subregularity of the prox-gradient mapping $\mathcal{G}_t(\cdot)$ at $(\bar{x}, 0)$ with constant γ. We aim to show that the error bound property is equivalent to subregularity of the subdifferential $\partial\varphi$ itself – a transparent notion closely tied to quadratic growth [10–12]. We first record the following elementary lemma.

Lemma 5.6 (Perturbation by identity)**.** *Consider a set-valued mapping $S\colon \mathbf{R}^n \rightrightarrows \mathbf{R}^m$ and a pair $(\bar{x}, 0) \in \operatorname{gph} S$. Then if S is subregular at $(\bar{x}, 0)$ with constant l, the mapping $(I + S^{-1})^{-1}$ is subregular at $(\bar{x}, 0)$ with constant $1 + l$. Conversely, if $(I + S^{-1})^{-1}$ is subregular at $(\bar{x}, 0)$ with constant $\hat{l}$, then S is subregular at $(\bar{x}, 0)$ with constant $1 + \hat{l}$.*

Proof. Suppose that S is l-subregular at $(\bar{x}, \bar{y})$, and let $\mathcal{X}$ be the corresponding neighborhood of $\bar{x}$. Consider a point $x \in \mathcal{X}$ along with an arbitrary point $z \in (I + S)^{-1}(x)$. For the sake of establishing subregularity of $(I + S^{-1})^{-1} = I - (I + S)^{-1}$ (equation (2.1)), we can clearly assume $z \in \mathcal{X}$. Taking into account the inclusion $x - z \in S(z)$, we obtain

$$\begin{aligned}
\operatorname{dist}\left(x; S^{-1}(0)\right) &\leq \|x - z\| + \operatorname{dist}(z; S^{-1}(0)) \\
&\leq \|x - z\| + l \cdot \operatorname{dist}(0; S(z)) \\
&\leq (1 + l)\|x - z\|,
\end{aligned}$$

as claimed. Conversely, suppose that $(I + S^{-1})^{-1}$ is subregular at $(\bar{x}, 0)$ with constant $\hat{l}$, and let $\mathcal{X}$ be the corresponding neighborhood of $\bar{x}$. Fix an arbitrary point $x \in \mathcal{X}$ and a vector $v \in S(x)$. Aiming to establish subregularity of S we can assume that $x + v$ lies in $\mathcal{X}$. Note the inclusion $x + v \in (I + S)(x)$ and hence $v \in (I + S^{-1})^{-1}(x + v)$. We deduce

$$\operatorname{dist}(x; S^{-1}(0)) \leq \|v\| + \operatorname{dist}(x + v; S^{-1}(0)) \leq (1 + \hat{l})\|v\|.$$

The result follows.

□

The following result analogous to Theorem 3.4 is now immediate.

Theorem 5.7 (Proximal and subdifferential subregularity)**.** *Consider a finite L-Lipschitz, convex function $h\colon \mathbf{R}^m \to \mathbf{R}$ and a C^1-smooth mapping $c\colon \mathbf{R}^n \to \mathbf{R}^m$. Define the composite function $\varphi := h \circ c$ and let $\bar{x}$ be a stationary point of φ. Consider the conditions*

- *(i) the subdifferential $\partial\varphi$ is subregular at $(x, 0)$ with constant l.*
- *(ii) the mapping $T := t^{-1}\left(I - (I + t\partial\varphi)^{-1}\right)$ is subregular at $(\bar{x}, 0)$ with constant $\hat{l}$.*

If condition (i) holds, then so does condition (ii) with $\hat{l} = l + t$. Conversely, condition (ii) implies condition (i) with $l = \hat{l} + t$.

Proof. Note first the equality $tT = (I + (t\partial\varphi)^{-1})^{-1}$ (equation (2.1)). Suppose that $\partial\varphi$ is l-subregular at $(\bar{x}, 0)$ with constant l. Then by Lemma 5.6, the mapping tT is subregular at $(\bar{x}, 0)$ with constant $1 + l/t$, and hence T is subregular at $(\bar{x}, 0)$ with constant $l + t$, as claimed. The converse argument is entirely analogous. □

The final step in relating subdifferential subregularity to the error bound property relies on Ekeland's variational principle, which we record below.

Theorem 5.8 (Ekeland's variational principle)**.** *Consider a closed function $g\colon \mathbf{R}^n \to \overline{\mathbf{R}}$ that is bounded from below. Suppose that for some $\epsilon > 0$ and $\bar{x} \in \mathbf{R}^n$, we have $g(\bar{x}) \leq \inf g + \epsilon$. Then for any $\rho > 0$, there exists a point $\bar{u}$ satisfying*

- $g(\bar{u}) \leq g(\bar{x})$,
- $\|\bar{x} - \bar{u}\| \leq \epsilon/\rho$,
- $\{\bar{u}\} = \operatorname*{argmin}_u \{g(u) + \rho\|u - \bar{u}\|\}$.

The following perturbation result will play a central role both in the proof of Theorem 5.10 and in Section 6.

Theorem 5.9 (Perturbation)**.** *Consider a finite L-Lipschitz, convex function $h\colon \mathbf{R}^m \to \mathbf{R}$ and a C^1-smooth mapping $c\colon \mathbf{R}^n \to \mathbf{R}^m$ with β-Lipchitz gradient. Define the composite function $\varphi := h \circ c$. Then for any real $t > 0$, there exists a point $\bar{u}$ satisfying the properties*

- *(i) (point proximity)* $\quad \|x^t - \bar{u}\| \leq \|x^t - x\|$,
- *(ii) (near-stationarity)* $\quad \operatorname{dist}(0; \partial\varphi(\bar{u})) \leq (3L\beta t + 2)\|\mathcal{G}_t(x)\|$.

Proof. Define the function $g(y) := \varphi(y) + \frac{1}{2}(L\beta + t^{-1})\|x - y\|^2$ and note that the inequality $g(y) \geq \varphi_t(x, y) \geq \varphi_t(x, x^t)$ holds for all $y \in \mathbf{R}^n$. Hence we have

$$g(x_t) - g^* \leq \varphi(x_t) - \varphi(x; x^t) + \frac{L\beta}{2}\|x^t - x\|^2 \leq L\beta\|x^t - x\|^2,$$

where the last inequality follows from (5.2). Define the constants $\epsilon := L\beta\|x^t - x\|^2$ and $\rho := L\beta\|x^t - x\|$. Applying Ekeland's variational principle, we obtain a point $\bar{u}$ satisfying the inequalities $g(\bar{u}) \leq g(x^t)$ and $\|x^t - \bar{u}\| \leq \epsilon/\rho$, and the inclusion $0 \in \partial g(\bar{u}) + \rho\mathbf{B}$. The proximity condition (i) is immediate. To see near-stationarity (ii), observe

$$\begin{aligned}
\operatorname{dist}(0, \partial\varphi(\bar{u})) &\leq \rho + (L\beta + t^{-1})\|\bar{u} - x\| \\
&\leq L\beta\|x^t - x\| + (L\beta + t^{-1})(\|x^t - \bar{u}\| + \|x^t - x\|) \\
&\leq \left(L\beta + 2(L\beta + t^{-1})\right)\|x^t - x\|.
\end{aligned}$$

The result follows. □

We are now ready to prove the main result of this section.

Theorem 5.10 (Subdifferential subregularity and the error bound property)**.** *Consider a finite L-Lipschitz, convex function $h\colon \mathbf{R}^m \to \mathbf{R}$ and a C^1-smooth mapping $c\colon \mathbf{R}^n \to \mathbf{R}^m$ with β-Lipchitz gradient. Define the composite function $\varphi := h \circ c$ and let $\bar{x}$ be a stationary point of φ. Consider the conditions:*

(i) the subdifferential $\partial\varphi$ is subregular at $(\bar{x}, 0)$ with constant l.

(ii) the prox-gradient mapping $\mathcal{G}_t(\cdot)$ is subregular at $(\bar{x}, 0)$ with constant $\hat{l}$.

If condition (ii) holds, then condition (i) holds with $l := 2\hat{l}$. Conversely, if condition (i) holds, then condition (ii) holds with $\hat{l} := (3L\beta t + 2)l + 2t$, and moreover in the case $t < (L\beta)^{-1}$, we can be sure $\hat{l} \leq (l + t)(1 + (1 - tL\beta)^{-1})$.

Proof. Suppose first that the gradient mapping $\mathcal{G}_t(x)$ is subregular at $(\bar{x}, 0)$ with constant $\hat{l}$. Then by Theorems 5.4 and 5.7, we deduce for all x near $\bar{x}$, the inequalities

$$\operatorname{dist}(x, \big(\partial\varphi)^{-1}(0)\big) \leq \hat{l} \cdot \|\mathcal{G}_t(x)\| \leq 2\hat{l} \cdot \operatorname{dist}(0; \partial\varphi(x)).$$

Conversely, suppose that $\partial\varphi$ is subregular at $(\bar{x}, 0)$ with constant l. Fix a point x, and let $\bar{u}$ be the point guaranteed to exist by Theorem 5.9. For the purpose of establishing subregularity of $\mathcal{G}_t(\cdot)$, we can suppose the x and x_t are arbitrarily close to $\bar{x}$. Then $\bar{u}$ is close to $\bar{x}$, and we deduce

$$\begin{aligned} l \cdot \operatorname{dist}(0, \partial\varphi(\bar{u})) \geq \operatorname{dist}\big(\bar{u}; (\partial\varphi)^{-1}(0)\big) &\geq \operatorname{dist}\big(x; (\partial\varphi)^{-1}(0)\big) - \|x^t - \bar{u}\| - \|x^t - x\| \\ &\geq \operatorname{dist}\big(x; (\partial\varphi)^{-1}(0)\big) - 2\|x^t - x\|. \end{aligned}$$

We conclude $\operatorname{dist}\big(x; (\partial\varphi)^{-1}(0)\big) \leq \big(l(3L\beta + 2t^{-1}) + 2\big) \|x^t - x\|$, as claimed. Finally, suppose $t < L\beta$. Then for all x near $\bar{x}$ we have

$$\operatorname{dist}(x, \big(\partial\varphi)^{-1}(0)\big) \leq (l{+}t){\cdot}t^{-1}\operatorname{dist}(0; x{-}(I{+}t\partial\varphi)^{-1}(x)) \leq (l{+}t)(1{+}(1{-}tL\beta)^{-1})\|\mathcal{G}_t(x)\|,$$

where the first inequality follows from Theorem 5.7 and the last inequality follows from Theorem 5.4. The result follows. □

Thus subregularity of the subdifferential $\partial\varphi$ and the error bound property are identical notions, with a precise relationship between the constants. Subdifferential subregularity at a minimizer, on the other hand, is equivalent to the natural quadratic growth condition when the functions in question are semi-algebraic (or more generally tame) [10]. To the best of our knowledge, it is not yet known if such a relationship persists for all convex-composite functions.

6 Natural rate of convergence under tilt-stability

As we alluded to in Section 5, the linear rate at which Algorithm 1 converges under the error bound condition is an order of magnitude slower than the rate that one would expect. In Section 5, we highlighted the equivalence between the error bound condition, subregularity of the subdifferential, and a quadratic growth condition. We will now show that when these properties hold *uniformly* relative to tilt-perturbations, the algorithm accelerates to the natural rate.

We begin with the following key definition.

Definition 6.1 (Stable strong local minimizer)**.** We say that $\bar{x}$ is an *stable strong local minimizer with constant* $\alpha > 0$ of a function $f\colon \mathbf{R}^n \to \overline{\mathbf{R}}$ if there exists a neighborhood $\mathcal{X}$ of $\bar{x}$ so that for each vector v near the origin, there is a point x_v (necessarily unique) in $\mathcal{X}$, with $x_0 = \bar{x}$, so that in terms of the perturbed functions $f_v := f(\cdot) - \langle v, \cdot \rangle$, the inequality

$$f_v(x) \geq f_v(x_v) + \frac{\alpha}{2}\|x - x_v\|^2 \qquad \text{holds for all } x \in \mathcal{X}.$$

This type of uniform quadratic growth is known to be equivalent to a number of influential notions, such as tilt-stability [24] and strong metric regularity of the subdifferential [11]. Here, we specialize the discussion to the convex-composite case, though the relationships hold much more generally. The following theorem appears in [12, Theorem 3.7, Proposition 4.5]; some predecessors were proved in [11,18].

Theorem 6.2 (Uniform growth, tilt-stability, & strong subdifferential regularity)**.** *Consider a finite convex function* $h\colon \mathbf{R}^m \to \mathbf{R}$ *and a* C^1*-smooth mapping* $c\colon \mathbf{R}^n \to \mathbf{R}^m$*. Define the composite function* $\varphi := h \circ c$ *and let* x^* *be a local minimizer of* φ*. Then the following properties are equivalent.*

1. **(uniform quadratic growth)** *The point* x^* *is a stable strong local minimizer of* φ *with constant* α*.*
2. **(local subdifferential convexity)** *There exists a a neighborhood* $\mathcal{X}$ *of* x^* *so that for any sufficiently small vector* v*, there is a point* $x_v \in \mathcal{X} \cap (\partial\varphi)^{-1}(v)$ *so that the inequality*

$$f(x) \geq f(x_v) + \langle v, x - x_v \rangle + \frac{\alpha}{2}\|x - x_v\|^2 \qquad \textit{holds for all } x \in \mathcal{X}.$$

3. **(tilt-stability)** *There exists a neighborhood* $\mathcal{X}$ *of* x^* *so that the mapping*

$$v \mapsto \operatorname*{argmin}_{x \in \mathcal{X}} \{\varphi(x) - \langle v, x \rangle\}$$

is single-valued and $1/\alpha$*-Lipschitz continuous on some neighborhood of the origin.*

4. **(strong regularity of the subdifferential)** *There exist neighborhoods* $\mathcal{X}$ *of* x^* *and* $\mathcal{V}$ *of* $\bar{v} = 0$ *so that the restriction* $(\partial f)^{-1}\colon \mathcal{V} \rightrightarrows \mathcal{X}$ *is a single-valued* $1/\alpha$*-Lipschitz continuous mapping.*

There has been a lot of recent work aimed at characterizing the above properties in concrete circumstances; see e.g. [19–22]. Suppose that the equivalent conditions in Theorem 6.2 hold. We will now investigate the impact of such an assumption on the linear convergence of Algorithm 1. For simplicity, assume $t \leq (L\beta)^{-1}$ and consider a point x near $\bar{x}$. Let $\bar{u}$ then be the point guaranteed to exist by Theorem 5.9, and set v to be the minimal norm vector in the subdifferential $\partial\varphi(\bar{u})$. Note the inequality $\|v\| \leq (3L\beta t + 2)\|\mathcal{G}_t(x)\| \leq 5\|\mathcal{G}_t(x)\|$. Hence for any point y near x^*, we have

$$\begin{aligned}
\varphi(y) &\geq \varphi(\bar{u}) + \langle v, y - \bar{u} \rangle && \text{(strong-regularity)}\\
&\geq \varphi_t(x, \bar{u}) - L\beta\|x - \bar{u}\|^2 + \langle v, y - \bar{u} \rangle && \text{(inequality (5.2))}\\
&\geq \varphi_t(x, x_t) - L\beta\|x - \bar{u}\|^2 + \langle v, y - \bar{u} \rangle && \text{(definition of } x_t)\\
&\geq \varphi(x^t) - L\beta\|x - \bar{u}\|^2 + \langle v, y - \bar{u} \rangle && \text{(inequality (5.2)).}
\end{aligned}$$

Plugging in $y = x^*$, we deduce

$$\begin{aligned}
\varphi(x^t) - \varphi(x^*) &\le L\beta\|x - \bar{u}\|^2 + \langle v, \bar{u} - x^*\rangle \\
&\le 4L\beta\|x^t - x\|^2 + \|v\| \cdot (\|\bar{u} - x^t\| + \|x^t - x\| + \|x - x^*\|) \\
&\le \frac{4}{L\beta}\|\mathcal{G}_t(x)\|^2 + 5\|\mathcal{G}_t(x)\| \cdot (2\|x^t - x\| + \|x - x^*\|) \\
&\le \frac{14}{L\beta}\|\mathcal{G}_t(x)\|^2 + 5\|\mathcal{G}_t(x)\| \cdot \|x - x^*\| \\
&= \frac{\|\mathcal{G}_t(x)\|^2}{L\beta}\left(14 + 5L\beta\frac{\|x - x^*\|}{\|\mathcal{G}_t(x)\|}\right).
\end{aligned}$$

Hence if while Algorithm 1 is running, the fractions $\frac{\|x_k - x^*\|}{\|\mathcal{G}_t(x_k)\|}$ remain bounded by a constant γ, appealing to the descent inequality (5.5), we obtain the Q-linear convergence guarantee

$$\varphi(x_{k+1}) - \varphi(x^*) \le \left(1 - \frac{1}{25 + 10L\beta\gamma}\right)(\varphi(x_k) - \varphi(x^*)). \tag{6.1}$$

We have thus established the main result of this section.

Theorem 6.3 (Natural rate of convergence)**.** *Consider a finite L-Lipschitz, convex function $h\colon \mathbf{R}^m \to \mathbf{R}$ and a C^1-smooth mapping $c\colon \mathbf{R}^n \to \mathbf{R}^m$ with β-Lipchitz gradient. Define the composite function $\varphi := h \circ c$ and let x_k be the sequence generated by Algorithm 1 with $t \le (L\beta)^{-1}$. Suppose that x_k has some limit point x^* around which one of the equivalent properties in Theorem 6.2 hold. Then for all large k, function values converge Q-linearly*

$$\varphi(x_{k+1}) - \varphi(x^*) \le \left(1 - \frac{1}{45 + 50L\beta/\alpha}\right)(\varphi(x_k) - \varphi(x^*)),$$

while the points x_k asymptotically converge R-linearly, meaning there exists an index r such that the inequality

$$\|x_{r+k} - x^*\|^2 \le \left(1 - \frac{1}{45 + 50L\beta/\alpha}\right)^k C \cdot (\varphi(x_r) - \varphi^*),$$

holds for all $k \ge 1$, where we set $C := \frac{2}{L\beta(1-\sqrt{1-(45+50L\beta/\alpha)^{-1}}))^2}$.

Proof. Strong regularity of the subdifferential in Theorem 6.2, in particular, implies that the subdifferential mapping $\partial\varphi$ is subregular at $(x^*, 0)$ with constant $1/\alpha$. Theorem 5.10 then implies that the error bound condition holds near x^* with constant $\frac{5}{\alpha} + \frac{2}{L\beta}$. Theorem 5.3 then shows that the sequence x_k converges to x^*. Consequently for all large indices k, the ratios $\frac{\|x_k - x^*\|}{\|\mathcal{G}_t(x_k)\|}$ are bounded by $\frac{5}{\alpha} + \frac{2}{L\beta}$. The inequality (6.1) immediately yields the claimed Q-linear rate. The R-linear rate follows easily by a standard argument, as in the proof of Theorem 5.3. □

References

[1] F.J. Aragón Artacho and M.H. Geoffroy. Characterization of metric regularity of subdifferentials. *J. Convex Anal.*, 15(2):365–380, 2008.

[2] A.Y. Aravkin, J.V. Burke, and G. Pillonetto. Sparse/robust estimation and kalman smoothing with nonsmooth log-concave densities: Modeling, computation, and theory. *The Journal of Machine Learning Research*, 14(1):2689–2728, 2013.

[3] A.Y. Aravkin, A. Lozano, R. Luss, and P. Kambadur. Orthogonal matching pursuit for sparse quantile regression. In *Data Mining (ICDM), 2014 IEEE International Conference on*, pages 11–19. IEEE, 2014.

[4] D. Azé and J.-N. Corvellec. Nonlinear local error bounds via a change of metric. *J. Fixed Point Theory Appl.*, 16(1-2):351–372, 2014.

[5] S. Basu, R. Pollack, and M. Roy. *Algorithms in Real Algebraic Geometry (Algorithms and Computation in Mathematics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[6] H.H. Bauschke and J.M. Borwein. On the convergence of von Neumann's alternating projection algorithm for two sets. *Set-Valued Anal.*, 1(2):185–212, 1993.

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

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

[9] W. Chu, S. S. Keerthi, and C. J. Ong. Bayesian support vector regression using a unified loss function. *Trans. Neur. Netw.*, 15(1):29–44, January 2004.

[10] D. Drusvyatskiy and A.D. Ioffe. Quadratic growth and critical point stability of semi-algebraic functions. *Math. Program.*, 153(2, Ser. A):635–653, 2015.

[11] D. Drusvyatskiy and A.S. Lewis. Tilt stability, uniform quadratic growth, and strong metric regularity of the subdifferential. *SIAM J. Optim.*, 23(1):256–267, 2013.

[12] D. Drusvyatskiy, B.S. Mordukhovich, and T.T.A. Nghia. Second-order growth, tilt stability, and metric regularity of the subdifferential. *J. Convex Anal.*, 21(4):1165–1192, 2014.

[13] 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).

[14] D. Klatte and B. Kummer. *Nonsmooth equations in optimization*, volume 60 of *Nonconvex Optimization and its Applications*. Kluwer Academic Publishers, Dordrecht, 2002. Regularity, calculus, methods and applications.

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

[16] G. Li and T.K. Pong. Calculus of the exponent of Kurdyka-łojasiewicz inequality and its applications to linear convergence of first-order methods. *arXiv:1602.02915*, 2016.

[17] Z.-Q. Luo and P. Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Ann. Oper. Res.*, 46/47(1-4):157–178, 1993. Degeneracy in optimization problems.

[18] B.S. Mordukhovich and T.T.A. Nghia. Second-order variational analysis and characterizations of tilt-stable optimal solutions in infinite-dimensional spaces. *Nonlinear Anal.*, 86:159–180, 2013.

[19] B.S. Mordukhovich and T.T.A. Nghia. Second-order characterizations of tilt stability with applications to nonlinear programming. *Math. Program.*, 149(1-2, Ser. A):83–104, 2015.

[20] B.S. Mordukhovich and J.V. Outrata. Tilt stability in nonlinear programming under Mangasarian-Fromovitz constraint qualification. *Kybernetika (Prague)*, 49(3):446–464, 2013.

[21] B.S. Mordukhovich, J.V. Outrata, and H. Ramírez C. Second-order variational analysis in conic programming with applications to optimality and stability. *SIAM J. Optim.*, 25(1):76–101, 2015.

[22] B.S. Mordukhovich and R. T. Rockafellar. Second-order subdifferential calculus with applications to tilt stability in optimization. *SIAM J. Optim.*, 22(3):953–986, 2012.

[23] Y. Nesterov. *Introductory Lectures on Convex Optimization.* Kluwer Academic, Dordrecht, The Netherlands, 2004.

[24] R.A. Poliquin and R.T. Rockafellar. Tilt stability of a local minimum. *SIAM J. Optim.*, 8(2):287–299 (electronic), 1998.

[25] 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.

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

[27] R.T. Rockafellar. *Convex analysis.* Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J., 1970.

[28] R.T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM J. Control Optimization*, 14(5):877–898, 1976.

[29] R.T. Rockafellar and A.L. Dontchev. *Implicit functions and solution mappings.* Monographs in Mathematics, Springer-Verlag, 2009.

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

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

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

[33] Z. Zhou and A.M.-C. So. A unified approach to error bounds for structured convex optimization problems. *arXiv:1512.03518*, 2015.

[34] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(2):301–320, 2005.