Mathematical Programming (2021) 185:357–383
https://doi.org/10.1007/s10107-019-01432-w

FULL LENGTH PAPER

Series A

Nonsmooth optimization using Taylor-like models: error bounds, convergence, and termination criteria

D. Drusvyatskiy[1] · A. D. Ioffe[2] · A. S. Lewis[3]

Received: 4 November 2016 / Accepted: 6 September 2019 / Published online: 27 September 2019

Abstract

We consider optimization algorithms that successively minimize simple Taylor-like models of the objective function. Methods of Gauss–Newton type for minimizing the composition of a convex function and a smooth map are common examples. Our main result is an explicit relationship between the step-size of any such algorithm and the slope of the function at a nearby point. Consequently, we (1) show that the step-sizes can be reliably used to terminate the algorithm, (2) prove that as long as the step-sizes tend to zero, every limit point of the iterates is stationary, and (3) show that conditions, akin to classical quadratic growth, imply that the step-sizes linearly bound the distance of the iterates to the solution set. The latter so-called error bound property is typically used to establish linear (or faster) convergence guarantees. Analogous results hold when the step-size is replaced by the square root of the decrease in the model's value. We complete the paper with extensions to when the models are minimized only inexactly.

Keywords Taylor-like model · Error-bound · Slope · Subregularity · Kurdyka–Łojasiewicz inequality · Ekeland's principle

Research of Drusvyatskiy was partially supported by the AFOSR YIP award FA9550-15-1-0237. Research of Lewis was supported in part by National Science Foundation Grant DMS-1208338. Research of all three authors was supported in part by by the US-Israel Binational Science Foundation Grant 2014241.

✉ D. Drusvyatskiy
ddrusv@uw.edu
http://www.math.washington.edu/∼ddrusv

A. D. Ioffe
ioffe@tx.technion.ac.il

A. S. Lewis
http://people.orie.cornell.edu/∼aslewis

[1] Department of Mathematics, University of Washington, Seattle, WA 98195, USA

[2] Department of Mathematics, Technion-Israel Institute of Technology, 32000 Haifa, Israel

[3] School of Operations Research and Information Engineering, Cornell University, Ithaca, NY, USA

Mathematics Subject Classification 65K05 · 90C30 · 49M37 · 65K10

1 Introduction

A basic algorithmic strategy for minimizing a function f on $\mathbf{R}^n$ is to successively minimize simple "models" of the function, agreeing with f at least up to first-order near the current iterate. We will broadly refer to such models as "Taylor-like". Some classical examples will help ground the exposition. When f is smooth, common algorithms given a current iterate x_k declare the next iterate x_{k+1} to be a minimizer of the quadratic model

$$m_k(x) := f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2}\langle B_k(x - x_k), x - x_k \rangle. \tag{1.1}$$

When the matrix B_k is a multiple of the identity, the scheme reduces to gradient descent; when B_k is the Hessian $\nabla^2 f(x_k)$, one recovers Newton's method; adaptively changing B_k based on accumulated information covers Quasi-Newton algorithms. Higher-order models can also appear; the cubicly regularized Newton's method of Nesterov–Polyak [42] uses the models

$$\begin{aligned} m_k(x) := f(x_k) &+ \langle \nabla f(x_k), x - x_k \rangle \\ &+ \frac{1}{2}\langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle + \frac{M}{6}\|x - x_k\|^3. \end{aligned}$$

For more details on Taylor-like models in smooth minimization, see Nocedal-Wright [47].

The algorithmic strategy generalizes far beyond smooth minimization. One important arena, and the motivation for the current work, is the class of convex composite problems

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

Here g is a closed convex function (possibly taking infinite values), h is a finite-valued Lipschitz convex function, and c is a smooth map. Algorithms for this problem class have been studied extensively, notably in [9,28,52,53,60,61] and more recently in [13,23,25,38]. Given a current iterate x_k, common algorithms declare the next iterate x_{k+1} to be a minimizer of

$$m_k(x) := g(x) + h\Big(c(x_k) + \nabla c(x_k)(x - x_k)\Big) + \frac{1}{2}\langle B_k(x - x_k), x - x_k \rangle. \tag{1.3}$$

The underlying assumption is that the minimizer of m_k can be efficiently computed. This is the case for example, when interior-point methods can be directly applied to the convex subproblem or when evaluating c and ∇c is already the computational bottleneck. The latter setting is ubiquitous in derivative free optimization; see for example the discussion in Wild [59]. The model m_k in (1.3) is indeed Taylor-like, even when g and h are nonconvex, since the inequality $|m_k(y) - f(y)| \le \frac{\mathrm{Lip}(h)\mathrm{Lip}(\nabla c) + \|B_k\|}{2}\|y - x_k\|^2$ holds for all points y, as the reader can readily verify. When B_k is a multiple of the

identity, the resulting method is called the "prox-linear algorithm" in [23,38], and it subsumes a great variety of schemes.

In the setting $h = 0$, the prox-linear algorithm reduces to the proximal-point method on the function g [40,41,55]. When c maps to the real line and h is the identity function, the scheme is the proximal gradient algorithm on the function $g + c$ [4,46]. Setting $g = 0$ and $h = \|\cdot\|$ yields a variant of the Gauss–Newton method for nonlinear least squares. Allowing B_k to vary with accumulated information results in variable metric variants of the aforementioned algorithms; see e.g. [9,12,58]. Extensions where h and g are not necessarily convex, but are nonetheless simple, are also important and interesting, in large part because of nonconvex penalties and regularizers common in machine learning applications. Other important variants interlace the model minimization step with inertial corrector steps, such as in accelerated gradient methods [31,43], cubically regularized Newton [45], and convex composite algorithms [25].

In this work, we take a broader view of nonsmooth optimization algorithms that use Taylor-like models. Rather than developing new algorithms, we aim to elucidate existing algorithms for nonsmooth optimization and their "primal-only" termination criteria.[1] Setting the stage, consider the optimization problem

$$\min_x \ f(x)$$

for an arbitrary lower-semicontinuous function f on $\mathbf{R}^n$. The model-based algorithms we investigate simply iterate the steps: x_{k+1} is a minimizer of some model $f_{x_k}(\cdot)$ based at x_k. In light of the discussion above, we assume that the models f_{x_k} approximate f (uniformly) up to first-order, meaning

$$|f_{x_k}(x) - f(x)| \leq \omega(\|x - x_k\|) \qquad \text{for all } \ k \in \mathbb{N} \text{ and } \ x \in \mathbf{R}^n, \tag{1.4}$$

where ω is any C^1-smooth function satisfying $\omega(0) = \omega'(0) = 0$. We will call ω a *growth function*. The most import growth function is certainly the quadratic $\omega(x, y) = \frac{1}{2}\|y - x\|^2$. More generally, one can allow power functions $\omega(x, y) = \|x - y\|^{1+v}$, which naturally arise when using high-order derivative expansions (e.g. cubic regularization) or when derivates are only Hölder continuous. For uses of a wider class of models for bundle methods, based on cutting planes, see Noll–Prot–Rondepierre [48]. In this great generality, we begin with the following basic question.

When should one terminate an algorithm that uses Taylor-like models?

For smooth nonconvex optimization, the traditional way to reliably terminate the algorithm is to stop when the norm of the gradient at the current iterate is smaller than some tolerance. For nonsmooth problems, termination criteria based on optimality conditions along the iterates may be meaningless as they may never be satisfied even in the limit. For example, one can easily exhibit a convex composite problem so that the iterates generated by the prox-linear algorithm described above converge to a

[1] Since the first version of this work [22], a number of new algorithms were developed building on our viewpoint. For example [16] analyze stochastic subgradient methods, [35,54] consider algorithms for adversarial learning and saddle-point problems, while [49] discuss generic line-search procedures using Taylor-like models built from Bregman divergences.

stationary point, while the optimality conditions at the iterates are not satisfied even in the limit.[2] Such lack of natural stopping criteria for nonsmooth first-order methods has been often remarked (and is one advantage of bundle-type methods).

There are, on the other hand, two appealing stopping criteria one can try: terminate the algorithm when either the step-size $\|x_{k+1} - x_k\|$ or the model decrease $f(x_k) - \inf f_{x_k}$ is sufficiently small. We will prove that both of these simple termination criteria are indeed reliable in the following sense. Theorem 3.1 and Corollary 5.4 show that if either the step-size $\|x_{k+1} - x_k\|$ or the model decrease $f(x_k) - \inf f_{x_k}$ is small, then there exists a point $\hat{x}$ close to x_{k+1} in both distance and in function value, which is nearly stationary for the problem. Determining the point $\hat{x}$ is usually difficult but is not important; the only role of $\hat{x}$ is to certify that the current iterate x_k is "close to near-stationarity" in the sense above. Theorem 3.1 follows quickly from Ekeland's variational principle [27]—a standard variational analytic tool. For other uses of the technique in variational analysis, see for example the survey [33]. Stopping criterion based on small near-by subgradients has appeared in many other contexts such as in descent methods of [32] and gradient sampling schemes of [11].

It is worthwhile to compare the viewpoint we advocate with more classical primal–dual termination criterion. In this work, we explore simple and intuitive stopping criteria that are both independent of the explicit presentation of the objective function and involves only the primal iterates. We justify the use of such criteria in terms of proximity to nearly stationary points. For particular cases, such as the composite class (1.2), KKT-residual-based stopping criteria are often available for algorithms attuned to the special structure in the objective. Such termination criteria, however, can be narrow in scope. For example, KKT based conditions do not allow one to compare such algorithms to methods that ignore the composite structure completely. To illustrate, consider applying a subgradient method to the problem (1.2). The composite structure is irrelevant for the subgradient method, and therefore measuring progress using the KKT residual is unnatural in this context. In contrast, using the primal only guarantees that we advocate here allow for a more fair and direct comparison between the subgradient and the prox-linear methods, as well their stochastic variants [16]. For a discussion, see Sect. 4.

Two interesting consequences for convergence analysis flow from our interpretation of the step-size and model decrease as measuring proximity to near-stationarity. Suppose that the models are chosen in such a way that the steps $\|x_{k+1} - x_k\|$ tend to zero. This assumption is often enforced by ensuring that $f(x_{k+1})$ is smaller than $f(x_k)$ by at least a multiple of $\|x_{k+1} - x_k\|^2$ (a sufficient decrease condition) using a back-tracking procedure or by safeguarding the minimal eigenvalue of B_k. Then assuming for simplicity that f is continuous on its domain, any limit point x^* of the iterate sequence x_k will be stationary for the problem (Corollary 3.3).[3] Analogous results hold with the step-size replaced by $f(x_k) - \inf f_{x_k}$. We note that a concise algorithmic framework, influenced by our techniques, appears in the recent manuscript [49].

[2] One such univariate example is $\min_x f(x) = |\frac{1}{2}x^2 + x|$. The prox-linear algorithm for convex composite minimization [23, Algorithm 5.1] initiated to the right of the origin—a minimizer of f—will generate a sequence $x_k \to 0$ with $|f'(x_k)| \to 1$.

[3] By stationary, we mean that zero is a limiting subgradient of the function at the point.

The subsequence convergence result is satisfying, since very little is assumed about the underlying algorithm. A finer analysis of linear, or faster, convergence rates relies on some regularity of the function f near a limit point x^* of the iterate sequence x_k. One of the weakest such regularity assumptions is that for all x near x^*, the "slope" of f at x linearly bounds the distance of x to the set of stationary points S—the "error". Here, we call this property the *slope error-bound*. To put it in perspective, we note that the slope error-bound always entails a classical quadratic growth condition away from S (see [19,24,62]), and is equivalent to it whenever f is convex (see [1,36]). Moreover, as an aside, we observe in Theorem 3.7 and Proposition 3.8 that under mild conditions, the slope error-bound is equivalent to the "Kurdyka–Łojasiewicz inequality" with exponent 1/2—an influential condition also often used to prove linear convergence. To the best of our knowledge, the earliest instance of algorithm analysis based on the latter inequality is [51].

Assuming the slope error-bound, a typical convergence analysis strategy aims to deduce that the step-sizes $\|x_{k+1} - x_k\|$ linearly bound the distance $\operatorname{dist}(x_k; S)$. Following Luo-Tseng [39], we call the latter property the *step-size error-bound*. We show in Theorem 3.5 that the slope error-bound indeed always implies the step-size error-bound, under the common assumption that the growth function $\omega(\cdot)$ is a quadratic. The proof is a straightforward consequence of the relationship we have established between the step-size and the slope at a nearby point—underscoring the power of the technique.

In practice, exact minimization of the model function f_{x_k} can be impossible. Instead, one can obtain a point x_{k+1} that is only nearly optimal or nearly stationary for the problem $\min f_{x_k}$. For example, the efficiency of the inexact prox-linear method is often remarked in the early works of Burke and Ferris [10], Fletcher [29], Nesterov [44], Wright [60], etc. More recent works [3,14,26] have extensive numerical examples of the prox-linear method for phase retrieval, blind deconvolution, and low-rank SDP problems, where the subproblems are solved by specialized first-order methods (ADMM). Even from the worst-case perspective, the complexity of an inexact prox-linear method is superior to its natural competitor, the sugradient method; see Remark 4.1. Section 5 shows that all the results above generalize to this more realistic setting. In particular, somewhat surprisingly, we argue that limit points of the iterates will be stationary even if the tolerances on optimality (or stationarity) and the step-sizes $\|x_{k+1} - x_k\|$ tend to zero at independent rates. The arguments in this inexact setting follow by applying the key result, Theorem 3.1, to small perturbations of f and f_{x_k}, thus illustrating the flexibility of the theorem.

The convex composite problem (1.2) and the prox-linear algorithm (along with its variable metric variants) is a fertile application arena for the techniques developed here. An early variant of the key Theorem 3.1 in this setting appeared recently in [23, Theorem 5.3] and was used there to establish sublinear, linear, and quadratic convergence guarantees for the prox-linear method under appropriate regularity conditions. We review these results in Sect. 4, as an illustration of our techniques. An important deviation of ours from earlier work is the use of the step-size as the fundamental analytic tool, in contrast to the Δ measures of Burke [9] and the criticality measures in Cartis–Gould–Toint [13]. To the best of our knowledge, the derived relationship between the step-size and stationarity at a nearby point is entirely new. The fact that the slope

error-bound implies that both the step-size and the square root of the model decrease linearly bounds the distance to the solution set (step-size and model error-bounds) is entirely new as well; previous related results have assumed that h is polyhedral.

The assumption (1.4) is appealing in its simplicity and modeling flexibility. To illustrate further, let us briefly mention two examples beyond the convex composite setting. Consider the problem, $\min_x h(c(x))$, where h is now smooth while $c(\cdot)$ is Lipschitz continuous. Then an easy computation shows that the model $f_x(y) = h(c(x)) + \langle h(c(x)), c(y) - c(x)\rangle + \frac{L}{2}\|y - x\|^2$ is Taylor-like for any $L > 0$. The minimization of the model $f_x(\cdot)$ may be straightforward. For instance, if $c(\cdot)$ takes the form $c(x) = (c_1(x_1), \ldots, c_1(x_n))$, then the model function is completely separable and therefore the minimization can be done in parallel. Indeed, such separable problems are common in a variety of nonlinear regression tasks (see, e.g. [30, Section 4]). If in addition $c(\cdot)$ is smooth, one can linearize the map for the purpose of simplifying computation, all the while preserving the Taylor-like behavior. As the second example, consider the problem, $\min_x f(x) = \max_{\lambda\in\Lambda} f(x, \lambda)$, where the functions $f(\cdot, \lambda)$ are smooth with gradients that are β-Lipschitz. Then we may simply take as the models the function $f_x(y) = \max_{\lambda\in\Lambda} f(x, \lambda) + \langle \nabla f(x, \lambda), y - x\rangle + \frac{L}{2}\|y - x\|^2$. More generally still, one can imagine a min-max problem where $f(\cdot, \lambda)$ are convex composite. One can then create a model by using the standard convex-composite models within the maximization, thereby underscoring the flexibility of the framework. Working out the algorithmic implications for all such examples would take us far off field. Instead, our goal here is succinct: to highlight the Taylor-like models as the unifying principle in nonsmooth optimization, while emphasizing the role of the stepsize as a termination criterion.

Though the discussion above takes place over the Euclidean space $\mathbf{R}^n$, the most appropriate setting for most of our development is over an arbitrary complete metric space. This is the setting of the paper. The outline is as follows. In Sect. 2, we establish basic notation and recall Ekeland's variational principle. Section 3 contains our main results. Section 4 illustrates the techniques for the prox-linear algorithm in composite minimization, while Sect. 5 explores extensions when the subproblems are solved inexactly.

2 Notation

Fix a complete metric space $\mathcal{X}$ with the metric $d(\cdot, \cdot)$. We denote the open unit ball of radius $r > 0$ around a point x by $\mathbf{B}_r(x)$. The distance from x to a set $Q \subset \mathcal{X}$ is

$$\operatorname{dist}(x; Q) := \inf_{y\in Q} d(x, y).$$

We will be interested in minimizing functions mapping $\mathcal{X}$ to the extended real line $\overline{\mathbf{R}} := \mathbf{R} \cup \{\pm\infty\}$. A function $f : \mathcal{X} \to \overline{\mathbf{R}}$ is called *lower-semicontinuous* (or *closed*) if the inequality $\liminf_{x\to\bar{x}} f(x) \geq f(\bar{x})$ holds for all points $\bar{x} \in \mathcal{X}$. We always assume that the functions we consider are *proper*, meaning that they are never $-\infty$ and are not always $+\infty$.

Consider a closed function $f\colon \mathcal{X} \to \overline{\mathbf{R}}$ and a point $\bar{x}$ with $f(\bar{x})$ finite. The *slope* of f at $\bar{x}$ is simply its maximal instantaneous rate of decrease:

$$|\nabla f|(\bar{x}) := \limsup_{x \to \bar{x}} \frac{(f(\bar{x}) - f(x))^+}{d(\bar{x}, x)}.$$

Here, we use the notation $r^+ = \max\{0, r\}$. If f is a differentiable function on a Euclidean space, the slope $|\nabla f|(\bar{x})$ simply coincides with the norm of the gradient $\|\nabla f(\bar{x})\|$, and hence the notation. For a convex function f, the slope $|\nabla f|(\bar{x})$ equals the norm of the shortest subgradient $v \in \partial f(\bar{x})$. The slope originates in the work of De Giorgi et al. [17]; for more details on the slope and its uses in optimization, see the survey [33], monograph [34], or the thesis [18].

The function $x \mapsto |\nabla f|(x)$ lacks basic lower-semicontinuity properties. As a result, it is important to introduce the *limiting slope*

$$\overline{|\nabla f|}(\bar{x}) := \liminf_{x \to \bar{x},\ f(x) \to f(\bar{x})} |\nabla f|(x).$$

In particular, if f is continuous on its domain, then $\overline{|\nabla f|}$ is simply the lower-semicontinuous envelope of $|\nabla f|$. We say that a point $\bar{x}$ is *stationary* for f if equality $\overline{|\nabla f|}(\bar{x}) = 0$ holds.

We will be interested in locally approximating functions up to first-order. Seeking to measure the "error in approximation", we introduce the following definition.

Definition 2.1 (*Growth function*) A differentiable univariate function $\omega\colon \mathbf{R}_+ \to \mathbf{R}_+$ is called a *growth function* if it satisfies $\omega(0) = \omega'(0) = 0$ and $\omega' > 0$ on $(0, \infty)$. If in addition, equalities $\lim_{t \to 0} \omega'(t) = \lim_{t \to 0} \omega(t)/\omega'(t) = 0$ hold, we say that ω is a *proper growth function*.

The main examples of proper growth functions are $\omega(t) := \frac{\eta}{r} \cdot t^r$ for real $\eta > 0$ and $r > 1$.

The following result, proved in [27], will be our main tool. The gist of the theorem is that if a point $\bar{x}$ nearly minimizes a closed function, then $\bar{x}$ is close to a true minimizer of a slightly perturbed function.

Theorem 2.2 (Ekeland's variational principle) *Consider a closed function $g\colon \mathcal{X} \to \overline{\mathbf{R}}$ 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 real $\rho > 0$, there exists a point $\hat{x}$ satisfying*

1. $g(\hat{x}) \leq g(\bar{x})$,
2. $d(\bar{x}, \hat{x}) \leq \epsilon/\rho$,
3. *$\hat{x}$ is the unique minimizer of the perturbed function $x \mapsto g(x) + \rho \cdot d(x, \hat{x})$.*

Notice that property 3 in Ekeland's principle directly implies the inequality $|\nabla g|(\hat{x}) \leq \rho$. Thus if a point $\bar{x}$ nearly minimizes g, then the slope of g is small at some nearby point.

2.1 Slope and subdifferentials

The slope is a purely metric creature. However, for a function f on $\mathbf{R}^n$, the slope is closely related to "subdifferentials", which may be more familiar to the audience. We explain the relationship here following [34]. Since the discussion will not be used in the sequel, the reader can safely skip it and move on to Sect. 3.

A vector $\bar{v} \in \mathbf{R}^n$ is called a *Fréchet subgradient* of a function $f\colon \mathbf{R}^n \to \overline{\mathbf{R}}$ at a point $\bar{x}$ if the inequality

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

The set of all Fréchet subgradients of f at $\bar{x}$ is the *Fréchet subdifferential* and is denoted by $\hat{\partial} f(\bar{x})$. The connection of the slope $|\nabla f|(\bar{x})$ to subgradients is immediate. A vector $\bar{v}$ lies in $\hat{\partial} f(\bar{x})$ if and only if the slope of the linearly tilted function $f(\cdot) - \langle \bar{v}, \cdot \rangle$ at $\bar{x}$ is zero. Moreover the inequality

$$|\nabla f|(\bar{x}) \leq \operatorname{dist}(0, \hat{\partial} f(\bar{x})) \qquad \text{holds.} \tag{2.1}$$

The *limiting subdifferential* of f at $\bar{x}$, denoted $\partial f(\bar{x})$, consists of all vectors $\bar{v}$ such that there exists sequences x_i and $v_i \in \hat{\partial} f(x_i)$ satisfying $(x_i, f(x_i), v_i) \to (\bar{x}, f(\bar{x}), \bar{v})$. Assuming that f is closed, a vector $\bar{v}$ lies in $\partial f(x)$ if and only if the limiting slope of the linearly tilted function $f(\cdot) - \langle \bar{v}, \cdot \rangle$ at $\bar{x}$ is zero. Moreover, Proposition 8.5 in [34] shows that the exact equality

$$\overline{|\nabla f|}(\bar{x}) = \operatorname{dist}(0, \partial f(\bar{x})) \qquad \text{holds.} \tag{2.2}$$

In particular, stationarity of f at $\bar{x}$ amounts to the inclusion $0 \in \partial f(\bar{x})$.

3 Main results

For the rest of the paper, fix a closed function $f\colon \mathcal{X} \to \overline{\mathbf{R}}$ on a complete metric space $\mathcal{X}$, and a point x with $f(x)$ finite. The following theorem is our main result. It shows that for any function $f_x(\cdot)$ (the "model"), such that the error in approximation $|f_x(y) - f(y)|$ is controlled by a growth function of the norm $d(x, y)$, the distance between x and the minimizer x^+ of $f_x(\cdot)$ prescribes near-stationarity of f at some nearby point $\hat{x}$.

Theorem 3.1 (Perturbation result) *Consider a closed function $f_x\colon \mathcal{X} \to \overline{\mathbf{R}}$ such that the inequality*

$$|f_x(y) - f(y)| \leq \omega(d(x, y)) \quad \textit{holds for all } y \in \mathcal{X}, \tag{3.1}$$

where ω is some growth function, and let x^+ be a minimizer of f_x. If x^+ coincides with x, then the slope $|\nabla f|(x)$ is zero. On the other hand, if x and x^+ are distinct, then there exists a point $\hat{x} \in \mathcal{X}$ satisfying

1. **(point proximity)** $d(x^+, \hat{x}) \leq 2 \cdot \frac{\omega(d(x^+,x))}{\omega'(d(x^+,x))}$,
2. **(value proximity)** $f(\hat{x}) \leq f(x^+) + \omega(d(x^+, x))$,
3. **(near-stationarity)** $|\nabla f|(\hat{x}) \leq \omega'(d(x^+, x)) + \omega'(d(\hat{x}, x))$.

Proof A quick computation shows the equality $|\nabla f_x|(x) = |\nabla f|(x)$. Thus if x^+ coincides with x, the slope $|\nabla f|(x)$ must be zero, as claimed. Therefore, for the remainder of the proof, we will assume that x^+ and x are distinct.

Observe now the inequality

$$f(y) \geq f_x(y) - \omega(d(x, y)) \geq f_x(x^+) - \omega(d(x, y)).$$

Define the function $g(y) := f(y) + \omega(d(x, y))$ and note $\inf g \geq f_x(x^+)$. We deduce

$$g(x^+) - \inf g \leq f(x^+) - f_x(x^+) + \omega(d(x^+, x)) \leq 2 \cdot \omega(d(x^+, x)). \quad (3.2)$$

An easy argument now shows the inequality

$$|\nabla g|(z) \geq |\nabla f|(z) - \omega'(d(z, x)) \qquad \text{for all } z \in \mathcal{X}.$$

Setting $\epsilon := 2\omega(d(x^+, x))$ and applying Ekeland's variational principle (Theorem 2.2), we obtain for any $\rho > 0$ a point $\hat{x}$ satisfying

$$g(\hat{x}) \leq g(x^+), \qquad d(x^+, \hat{x}) \leq \frac{\epsilon}{\rho}, \qquad \text{and} \qquad |\nabla g|(\hat{x}) \leq \rho.$$

We conclude $|\nabla f|(\hat{x}) \leq \rho + \omega'(d(\hat{x}, x))$. Setting $\rho := \omega'(d(x^+, x))$ yields the result. □

Note that the distance $d(\hat{x}, x)$ appears on the right hand-side of the near-stationarity property. By the triangle-inequality and point proximity, however, it can be upper bounded by $d(x^+, x) + 2 \cdot \frac{\omega(d(x^+,x))}{\omega'(d(x^+,x))}$, a quantity independent of $\hat{x}$. To better internalize this result, let us look at the most important setting of Theorem 3.1 where the growth function is a quadratic $\omega(t) = \frac{\eta}{2}t^2$ for some real $\eta > 0$.

Corollary 3.2 (Quadratic error) *Consider a closed function* $f_x \colon \mathcal{X} \to \overline{\mathbf{R}}$ *and suppose that with some real* $\eta > 0$ *the inequality*

$$|f_x(y) - f(y)| \leq \frac{\eta}{2} \cdot d^2(x, y) \quad \textit{holds for all } y \in \mathcal{X}.$$

Define x^+ *to be the minimizer of* f_x. *Then there exists a point* $\hat{x} \in \mathbf{R}^n$ *satisfying*

1. **(point proximity)** $d(x^+, \hat{x}) \leq d(x^+, x)$,
2. **(value proximity)** $f(\hat{x}) \leq f(x^+) + \frac{\eta}{2} \cdot d^2(x^+, x)$,
3. **(near-stationarity)** $|\nabla f|(\hat{x}) \leq 5\eta \cdot d(x^+, x)$.

An immediate consequence of Theorem 3.1 is the following subsequence convergence result.

Corollary 3.3 (Subsequence convergence to stationary points) *Consider a sequence of points x_k and closed functions $f_{x_k} : \mathcal{X} \to \overline{\mathbf{R}}$ satisfying $x_{k+1} = \operatorname{argmin}_y f_{x_k}(y)$ and $d(x_{k+1}, x_k) \to 0$. Suppose moreover that the inequality*

$$|f_{x_k}(y) - f(y)| \leq \omega(d(y, x_k)) \qquad \text{holds for all indices } k \text{ and points } y \in \mathcal{X},$$

where ω is a proper growth function. If $(x^*, f(x^*))$ is a limit point of the sequence $(x_k, f(x_k))$, then x^* is stationary for f.

Proof Fix a subsequence x_{k_i} with $(x_{k_i}, f(x_{k_i})) \to (x^*, f(x^*))$, and consider the points $\hat{x}_{k_i}$ guaranteed to exist by Theorem 3.1. By point proximity, we deduce $d(x_{k_i}, \hat{x}_{k_i-1}) \leq \frac{\omega(d(x_{k_i}, x_{k_i-1}))}{\omega'(d(x_{k_i}, x_{k_i-1}))}$, and the fact that the right hand-side tends to zero, we conclude that $\hat{x}_{k_i-1}$ converge to x^*. The functional proximity, $f(\hat{x}_{k_i-1}) \leq f(x_{k_i}) + \omega(d(x_{k_i}, x_{k_i-1}))$ implies $\limsup_{i\to\infty} f(\hat{x}_{k_i-1}) \leq \limsup_{i\to\infty} f(x_{k_i}) = f(x^*)$. Lower-semicontinuity of f then implies the equality $\lim_{i\to\infty} f(\hat{x}_{k_i-1}) = f(x^*)$. Finally, the near-stationarity,

$$|\nabla f|(\hat{x}_{k_i-1}) \leq \omega'(d(x_{k_i}, x_{k_i-1})) + \omega'(d(\hat{x}_{k_i-1}, x_{k_i-1})),$$

implies $|\nabla f|(\hat{x}_{k_i-1}) \to 0$. Thus x^* is a stationary point of f. □

Remark 3.4 (Asymptotic convergence to critical points) Corollary 3.3 proves something stronger than stated. An unbounded sequence z_k is *asymptotically critical* for f if it satisfies $|\nabla f|(z_k) \to 0$. The proof of Corollary 3.3 shows that if the sequence x_k is unbounded, then there exists an asymptotically critical sequence z_k satisfying $d(x_k, z_k) \to 0$.

Corollary 3.3 is fairly satisfying since very little is assumed about the model functions. More sophisticated linear, or faster, rates of convergence rely on some regularity of the function f near a limit point x^* of the iterate sequence x_k. A classical example in nonlinear programming is the second-order sufficient condition for optimality. The literature on regularity concepts for broader nonsmooth problems is vast, relying on set-valued generalizations of the classical inverse function theorem and transversality concepts. We refer the reader to the monographs of Dontchev–Rockafellar [57] and Ioffe [34] for details, as well the paper of Bolte et al. [6].

Let S denote the set of stationary points of f. One of the weakest regularity assumptions is that the slope $|\nabla f|(x)$ linearly bounds the distance $\operatorname{dist}(x; S)$ for all x near x^*. Indeed, this property, which we call the *slope error-bound*, always entails a classical quadratic growth condition away from S (see [19,24]), and is equivalent to it whenever f is a convex function on $\mathbf{R}^n$ (see [1,36]).

Assuming such regularity, a typical convergence analysis strategy, explored for example by Luo–Tseng [39], aims to deduce that the step-sizes $d(x_{k+1}, x_k)$ linearly bound the distance $\operatorname{dist}(x_k; S)$. The latter is called the *step-size error-bound property*. We now show that slope error-bound always implies the step-size error-bound, under the mild and natural assumption that the models f_{x_k} deviate form f by a quadratic error in the distance.

Theorem 3.5 (Slope and step-size error-bounds) *Let S be an arbitrary set and fix a point $x^* \in S$ satisfying the condition*

– **(Slope error-bound)** $\operatorname{dist}(x; S) \leq L \cdot |\nabla f|(x)$ *for all* $x \in \mathbf{B}_\gamma(x^*)$.

Consider a closed function $f_x \colon \mathcal{X} \to \overline{\mathbf{R}}$ and suppose that for some $\eta > 0$ the inequality

$$|f_x(y) - f(y)| \leq \frac{\eta}{2} d^2(y, x) \qquad \text{holds for all } y \in \mathcal{X}.$$

Then letting x^+ be any minimizer of f_x, the following holds:

– **(Step-size error-bound)**

$$\operatorname{dist}(x, S) \leq (3L\eta + 2) \cdot d(x^+, x) \quad \text{when} \quad x, x^+ \in \mathbf{B}_{\gamma/3}(x^*).$$

Proof Suppose that the points x and x^+ lie in $\mathbf{B}_{\gamma/3}(x^*)$. Let $\hat{x}$ be the point guaranteed to exist by Corollary 3.2. We deduce

$$d(\hat{x}, x^*) \leq d(\hat{x}, x^+) + d(x^+, x^*) \leq d(x^+, x) + d(x^+, x^*) < \gamma.$$

Thus $\hat{x}$ lies in $\mathbf{B}_\gamma(x^*)$ and we obtain

$$\begin{aligned} L \cdot |\nabla f|(\hat{x}) \geq \operatorname{dist}\left(\hat{x}; S\right) &\geq \operatorname{dist}(x; S) - d(x^+, \hat{x}) - d(x^+, x) \\ &\geq \operatorname{dist}(x; S) - 2d(x^+, x). \end{aligned}$$

Taking into account the inequality $|\nabla f|(\hat{x}) \leq 3\eta \cdot d(x^+, x)$, we conclude

$$\operatorname{dist}(x; S) \leq (3L\eta + 2) \cdot d(x^+, x),$$

as claimed. □

Remark 3.6 (Slope and subdifferential error-bounds) It is instructive to put the slope error-bound property in perspective for those more familiar with subdifferentials. To this end, suppose that f is defined on $\mathbf{R}^n$ and consider the *subdifferential error-bound* condition

$$\operatorname{dist}(x; S) \leq L \cdot \operatorname{dist}(0; \hat{\partial} f(x)) \quad \text{for all} \quad x \in \mathbf{B}_\gamma(x^*). \tag{3.3}$$

Clearly in light of the inequality (2.1), the slope error-bound implies the subdifferential error-bound (3.3). Indeed, the slope and subdifferential error-bounds are equivalent. To see this, suppose (3.3) holds and consider an arbitrary point $x \in \mathbf{B}_\gamma(x^*)$. Appealing to the equality (2.2), we obtain sequences x_i and $v_i \in \hat{\partial} f(x_i)$ satisfying $x_i \to x$ and $\|v_i\| \to \overline{|\nabla f|}(x)$. Inequality (3.3) then implies $\operatorname{dist}(x_i; S) \leq L \cdot \|v_i\|$ for each sufficiently large index i. Letting i tend to infinity yields the inequality, $\operatorname{dist}(x; S) \leq L \cdot \overline{|\nabla f|}(x) \leq L \cdot |\nabla f|(x)$, and therefore the slope error-bound is valid.

Lately, a different condition now called the *Kurdyka–Łojasiewicz inequality* [5,37] with exponent 1/2 has been often used to study linear rates of convergence, beginning with Polyak [51]. The manuscripts [2,8] are influential recent examples. We finish the section with the observation that the Kurdyka–Łojasiewicz inequality always implies the slope error-bound relative to a sublevel set S; that is, the KŁ inequality is no more general than the slope error-bound. A different argument for (semi) convex functions based on subgradient flow appears in [7, Theorem 5]. In Proposition 3.8 we will also observe that the converse implication holds for all prox-regular functions. Henceforth, we will use the sublevel set notation $[f \le b] := \{x : f(x) \le b\}$ and similarly $[a < f < b] := \{x : a < f(x) < b\}$.

Theorem 3.7 (KŁ-inequality implies the slope error-bound)
Suppose that there is a nonempty open set $\mathcal{U}$ in $\mathcal{X}$ such that the inequalities

$$(f(x) - f^*)^\theta \le \alpha \cdot |\nabla f(x)| \quad \text{hold for all} \quad x \in \mathcal{U} \cap [f^* < f < r],$$

where $\theta \in (0, 1)$, $\alpha > 0$, f^*, and $r > f^*$ are real numbers. Then there exists a nonempty open set $\widehat{\mathcal{U}}$ and a real number $\hat{r}$ so that the inequalities

$$d(x; [f \le f^*]) \le \frac{\alpha^{\theta^{-1}}}{1-\theta} \cdot |\nabla f|^{\frac{1-\theta}{\theta}}(x) \quad \text{hold for all} \quad x \in \widehat{\mathcal{U}} \cap [f^* < f < \hat{r}].$$

In the case $\mathcal{U} = \mathcal{X}$, we can ensure $\widehat{\mathcal{U}} = \mathcal{X}$ and $\hat{r} = r$.

Proof Define the function $g(x) = (\max\{0, f(x) - f^*\})^{1-\theta}$. Note the inequality $|\nabla g|(x) \ge \frac{1-\theta}{\alpha}$ for all $x \in \mathcal{U} \cap [f^* < f < r]$. Let $R > 0$ be strictly smaller than the largest radius of a ball contained in $\mathcal{U}$ and define $\varepsilon := \min\left\{r - f^*, \frac{(1-\theta)R}{\alpha}\right\}$. Define the nonempty set $\widehat{\mathcal{U}} := \{x \in \mathcal{U} : \mathbf{B}_R(x) \subseteq \mathcal{U}\}$ and fix a point $x \in \widehat{\mathcal{U}} \cap [f^* < f < f^* + \varepsilon]$.

Observe now for any point $u \in [f^* < f < f^* + \varepsilon]$ with $d(x, u) \le R$, the inclusion $u \in \mathcal{U} \cap [f^* < f < r]$ holds, and hence $|\nabla g|(u) \ge \frac{1-\theta}{\alpha}$. Appealing to [20, Lemma 2.5] (or [33, Chapter 1, Basic Lemma]), we deduce the estimate

$$d(x; [f \le f^*]) \le \frac{\alpha}{1-\theta} \cdot g(x) = \frac{\alpha}{1-\theta} \cdot (f(x) - f^*)^{1-\theta} \le \frac{\alpha^{\theta^{-1}}}{1-\theta} \cdot (|\nabla f|(x))^{\frac{1-\theta}{\theta}}.$$

The proof is complete. □

The converse of Theorem 3.7 holds for "prox-regular functions" on $\mathbf{R}^n$, and in particular for "lower-C^2 functions". The latter are functions f on $\mathbf{R}^n$ such that around each point there is a neighborhood $\mathcal{U}$ and a real $l > 0$ such that $f + \frac{l}{2}\|\cdot\|^2$ is convex on $\mathcal{U}$.

Proposition 3.8 (Slope error-bound implies KŁ-inequality) *Consider a closed function $f : \mathbf{R}^n \to \overline{\mathbf{R}}$. Fix a real number f^* and a nonempty set $S \subseteq [f \le f^*]$. Suppose that there is a set $\mathcal{U}$, and constants L, l, ϵ, and $r > f^*$ such that the inequalities*

$$f(y) \geq f(x) + \langle v, y - x\rangle - \frac{l}{2}\|y - x\|^2,$$
$$\operatorname{dist}(x; S) \leq L \cdot \operatorname{dist}(0; \partial f(x)),$$

hold for all $x \in \mathcal{U} \cap [f^* < f < r]$, $y \in \mathcal{X}$, *and* $v \in \partial f(x) \cap \mathbf{B}_\epsilon(0)$. *Then the inequalities*

$$\sqrt{f(x) - f^*} \leq \sqrt{L + lL^2/2} \cdot \operatorname{dist}(0; \partial f(x)),$$

hold for all $x \in \mathcal{U} \cap [f^* < f < \hat{r}]$ *where we set* $\hat{r} := \min\{r, (L + lL^2/2)\epsilon^2\}$.

Proof Consider a point $x \in \mathcal{U} \cap [f^* < f < \hat{r}]$. Suppose first $\operatorname{dist}(0; \partial f(x)) \geq \epsilon$. Then we deduce $\sqrt{f(x) - f^*} \leq \sqrt{\hat{r}} \leq \sqrt{L + lL^2/2} \cdot \epsilon \leq \sqrt{L + lL^2/2} \cdot \operatorname{dist}(0; \partial f(x))$, as claimed. Hence we may suppose there exists a subgradient $v \in \partial f(x) \cap \mathbf{B}_\epsilon(0)$. We deduce

$$\begin{aligned} f^* \geq f(y) &\geq f(x) + \langle v, y - x\rangle - \frac{l}{2}\|y - x\|^2 \\ &\geq f(x) - \|v\| \cdot \|y - x\| - \frac{l}{2}\|y - x\|^2. \end{aligned}$$

Choosing v, y such that $\|v\|$ and $\|y - x\|$ attain $\operatorname{dist}(0; \partial f(x))$ and $\operatorname{dist}(x; S)$, respectively, we deduce $f(x) - f^* \leq \left(L + \frac{lL^2}{2}\right) \cdot \operatorname{dist}^2(0; \partial f(x))$. The result follows. □

4 Illustration: convex composite minimization

In this section, we briefly illustrate the results of the previous section in the context of composite minimization, and recall some consequences already derived in [23] from preliminary versions of the material presented in the current paper. This section will not be used in the rest of the paper, and so the reader can safely skip it if needed.

The notation and much of discussion follows that set out in [23]. Consider the minimization problem

$$\min_x \; f(x) := g(x) + h(c(x)), \tag{4.1}$$

where $g\colon \mathbf{R}^n \to \overline{\mathbf{R}}$ is a closed convex function, $h\colon \mathbf{R}^m \to \mathbf{R}$ is a finite-valued l-Lipschitz convex function, and $c\colon \mathbf{R}^n \to \mathbf{R}^m$ is a C^1-smooth map with the Jacobian $\nabla c(\cdot)$ that is β-Lipschitz continuous. Define the model function

$$f_x(y) := g(y) + h\Big(c(x) + \nabla c(x)(y - x)\Big) + \frac{l\beta}{2}\|y - x\|^2.$$

One can readily verify the inequality

$$0 \leq f_x(y) - f(y) \leq \frac{l\beta}{2}\|y - x\|^2 \qquad \text{for all } x, y \in \mathbf{R}^n.$$

In particular, the models f_x are "Taylor-like". The prox-linear algorithm iterates the steps

$$x_{k+1} = \operatorname*{argmin}_y \; f_{x_k}(y). \tag{4.2}$$

The following is a rudimentary convergence guarantee of the scheme [23, Section 5]:

$$\sum_{i=1}^{k} \|x_{i+1} - x_i\|^2 \leq \frac{2(f(x_1) - f^*)}{l\beta}, \tag{4.3}$$

where f^* is the limit of the decreasing sequence $\{f(x_k)\}$. In particular, the step-sizes $\|x_{i+1} - x_i\|$ tend to zero. Moreover, one can readily verify that for any limit point x^* of the iterate sequence x_k, equality $f^* = f(x^*)$ holds. Consequently, by Corollary 3.3, the point x^* is stationary for f:

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

We note that stationarity of the limit point x^* is well-known and can be proved by other means; see for example the discussion in [13]. From (4.3), one also concludes the rate

$$\min_{i=1,\ldots,k} \|x_{i+1} - x_i\|^2 \leq \frac{2(f(x_1) - f^*)}{l\beta \cdot k}.$$

What is the relationship of this rate to near-stationary of the iterate x_k? Corollary 3.2 shows that after $\frac{2l\beta(f(x_1)-f^*)}{25\cdot\epsilon^2}$ iterations, one is guaranteed to find an iterate x_k such that there exists a point $\hat{x}$ satisfying

$$\left.\begin{array}{c} 5l\beta \cdot \|\hat{x} - x_{k+1}\| \\ 5\sqrt{2l\beta} \cdot \sqrt{(f(\hat{x}) - f(x_{k+1}))^+} \\ \operatorname{dist}(0; \partial f(\hat{x})) \end{array}\right\} \leq \epsilon.$$

Let us now move on to linear rates of convergence. Fix a limit point x^* of the iterate sequence x_k and let S be the set of stationary points of f. Then Theorem 3.5 shows that the regularity condition

– **(Slope error-bound)**

$$\operatorname{dist}(x; S) \leq \frac{1}{\alpha} \cdot \operatorname{dist}(0; \partial f(x)) \quad \text{for all} \quad x \in \mathbf{B}_\gamma(x^*).$$

implies

– **(Step-size error-bound)**

$$\operatorname{dist}(x_k, S) \leq (3l\beta/\alpha + 2) \cdot \|x_{k+1} - x_k\| \quad \text{when} \quad x_k, x_{k+1} \in \mathbf{B}_{\gamma/3}(x^*).$$

Additionally, in the next section (Corollary 5.7) we will show that the slope error-bound also implies

– **(Model error-bound)**

$$\operatorname{dist}(x_k; S) \leq \left(\alpha^{-1}\sqrt{12l\beta} + \frac{2}{\sqrt{3l\beta}}\right) \cdot \sqrt{f(x_k) - \inf f_{x_k}},$$

whenever $f(x_k) - \inf f_{x_k} < \frac{3l\beta\gamma^2}{16}$ and x_k lies in $\mathbf{B}_{\gamma/2}(x^*)$.

It was, in fact, proved in [23, Theorem 5.10] that the slope and step-size error bounds are equivalent up to a change of constants. Moreover, as advertised in the introduction, the above implications were used in [23, Theorem 5.5] to show that if the slope error-bound holds then the function values converge linearly:

$$f(x_{k+1}) - f^* \leq q(f(x_k) - f^*) \qquad \text{for all large } k,$$

where

$$q \approx 1 - \left(\frac{\alpha}{l\beta}\right)^2.$$

The rate improves to $q \approx 1 - \frac{\alpha}{L\beta}$ under a stronger regularity condition, called tilt-stability [23, Theorem 6.3], while a local quadratic rate of convergence is assured under a sharpness property [23, Theorem 7.2]. The arguments of the better local rates again crucially employ a comparison of step-lengths and subgradients at near-by points.

Our underlying assumption is that the models f_{x_k} are easy to minimize, by an interior point method for example. This assumption may not be realistic in some large-scale applications. Instead, one must solve the subproblems (4.2) inexactly by a first-order method. The recent manuscript [25] investigates efficiency estimates of such methods in the first-order oracle model.

Remark 4.1 (Primal vs. primal–dual termination criteria) We propose a general, simple, intuitive stopping criterion that is both independent of the explicit presentation of the objective function and involves only the primal iterates. For particular cases, such as the compositional problems considered in this section, KKT-residual-based stopping criteria are often available. Such termination criteria, however, can be narrow in scope. For example, KKT based conditions do not allow one to compare such algorithms to methods that ignore the composite structure completely.

To lillustrate, suppose for simplicity $g = 0$ and consider applying a subgradient method to the problem (4.1), namely

$$x_{k+1} = x_k - \alpha_k v_k \qquad \text{with} \quad v_k \in \partial f(x_k).$$

Here α_k has to be appropriately chosen (roughly in the order of $1/\sqrt{T}$, where T is the total number of iterations) The recent paper [16] shows that the convergence rate of this method can be succinctly summarized exactly in terms described here. The iterates become ϵ-close to an ϵ-stationary point after $O(1/\epsilon^4)$ iterations. The analogous iteration complexity for the prox-linear method is $O(1/\epsilon^2)$, and even if the subproblems are solved inexactly by first-order methods, the number of matrix-vector multiplications required is at most $O(1/\epsilon^3)$ [25]. The composite structure is irrelevant for the subgradient method, and therefore measuring progress using the KKT residual is unnatural in this context. In contrast, using primal only guarantees that we advocate here allow for a more fair and direct comparison.

There are other reasons to focus on primal-only guarantees. (1) Proximity to nearly stationary points is an intuitive principle, while the impact of the KKT residual being small on the primal iterate—one we primarily care about—is more opaque. (2) Complexity guarantees for solving the primal–dual pair can in general be much worse than for approximately solving the primal problem only. That is, the dual iterate may lag behind the primal. Therefore the impact of the small stepsize on the primal iterates alone appears meaningful. (3) The error bound property of the step-size is in general much weaker than the error bound for the KKT system, as it involves stability only in the primal. It is this property that underlies rapid convergence of primal-only methods.

5 Inexact extensions and model decrease as termination criteria

Often, it may be impossible to obtain an exact minimizer x^+ of a model function f_x. What can one say then when x^+ minimizes the model function f_x only approximately? By "approximately", one can mean a number of concepts. Two most natural candidates are that x^+ is ϵ-*optimal*, meaning $f_x(x^+) \leq \inf f_x + \epsilon$, or that x^+ is ϵ-*stationary*, meaning $|\nabla f_x|(x^+) \leq \epsilon$. In both cases, all the results of Sect. 3 generalize quickly by bootstrapping Theorem 3.1; under a mild condition, both of the two notions above imply that x^+ is a minimizer of a slightly perturbed function, to which the key Theorem 3.1 can be directly applied.

5.1 Near-optimality for the subproblems

We begin with ϵ-optimality, and discuss ϵ-stationarity in Sect. 5.3. The following is an inexact analogue of Theorem 3.1. Though the statement may appear cumbersome at first glance, it simplifies dramatically in the most important case where ω is a quadratic; this case is recorded in Corollary 5.2.

Theorem 5.1 (Perturbation result under approximate optimality) *Consider a closed function $f_x : \mathcal{X} \to \overline{\mathbf{R}}$ such that the inequality*

$$|f_x(y) - f(y)| \leq \omega(d(x, y)) \quad \textit{holds for all } y \in \mathcal{X},$$

where ω is some growth function. Let x^+ be a point satisfying $f_x(x^+) \leq \inf f_x + \epsilon$. Then for any constant $\rho > 0$, there exist two points z and $\hat{x}$ satisfying the following.

1. **(point proximity)** *The inequalities*

$$d(x^+, z) \le \frac{\epsilon}{\rho} \quad \text{and} \quad d(z, \hat{x}) \le 2 \cdot \frac{\omega(d(z,x))}{\omega'(d(z,x))} \quad \text{hold,}$$

   under the convention $\frac{0}{0} = 0$,
2. **(value proximity)** $f(\hat{x}) \le f(x^+) + 2\omega(d(z,x)) + \omega(d(x^+,x))$,
3. **(near-stationarity)** $|\nabla f|(\hat{x}) \le \rho + \omega'(d(z,x)) + \omega'(d(\hat{x},x))$.

Proof By Theorem 2.2, for any $\rho > 0$ there exists a point z satisfying $f_x(z) \le f_x(x^+)$, $d(z, x^+) \le \frac{\epsilon}{\rho}$, and so that z is the unique minimizer of the function $y \mapsto f_x(y) + \rho \cdot d(y,z)$. Define the functions $\widetilde{f}(y) := f(y) + \rho \cdot d(y,z)$ and $\widetilde{f}_x(y) := f_x(y) + \rho \cdot d(y,z)$. Notice the inequality

$$|\widetilde{f}_x(y) - \widetilde{f}(y)| \le \omega(d(x,y)) \quad \text{for all } y.$$

Thus applying Theorem 3.1, we deduce that there exists a point $\hat{x}$ satisfying $d(z,\hat{x}) \le 2 \cdot \frac{\omega(d(z,x))}{\omega'(d(z,x))}$, $\widetilde{f}(\hat{x}) \le \widetilde{f}(z) + \omega(d(z,x))$, and $|\nabla \widetilde{f}|(\hat{x}) \le \omega'(d(z,x)) + \omega'(d(\hat{x},x))$. The point proximity claim is immediate. The value proximity follows from the inequality

$$\begin{aligned} f(\hat{x}) \le \widetilde{f}(\hat{x}) &\le f(z) + \omega(d(z,x)) \le f_x(z) + 2\omega(d(z,x)) \le f_x(x^+) + 2\omega(d(z,x)) \\ &\le f(x^+) + 2\omega(d(z,x)) + \omega(d(x^+,x)). \end{aligned}$$

Finally, the inequalities

$$|\nabla f|(\hat{x}) \le \rho + |\nabla \widetilde{f}|(\hat{x}) \le \rho + \omega'(d(z,x)) + \omega'(d(\hat{x},x))$$

imply the near-stationarity claim. □

Specializing to when ω is a quadratic yields the following.

Corollary 5.2 (Perturbation under quadratic error) *Consider a closed function* $f_x : \mathcal{X} \to \overline{\mathbf{R}}$ *and suppose that with some real* $\eta > 0$ *the inequality*

$$|f_x(y) - f(y)| \le \frac{\eta}{2} d^2(x,y) \quad \textit{holds for all } y \in \mathcal{X}.$$

Let x^+ *be a point satisfying* $f_x(x^+) \le \inf f_x + \epsilon$. *Then there exists a point* $\hat{x}$ *satisfying the following.*

1. **(point proximity)** $d(x^+, \hat{x}) \le \sqrt{\frac{4\epsilon}{3\eta}} + d(x^+, x)$,
2. **(value proximity)** $f(\hat{x}) \le f(x^+) + \eta \left(\sqrt{\frac{\epsilon}{3\eta}} + d(x^+,x) \right)^2 + \frac{\eta}{2} d^2(x^+,x)$,
3. **(near-stationarity)** $|\nabla f|(\hat{x}) \le \sqrt{12\eta\epsilon} + 3\eta \cdot d(x^+,x)$.

Proof Consider the two point $\hat{x}$ and z guaranteed to exist by Theorem 5.1. Observe the inequalities

$$d(z, x) \leq d(z, x^+) + d(x^+, x) \leq \frac{\epsilon}{\rho} + d(x^+, x),$$

and

$$d(x^+, \hat{x}) \leq d(x^+, z) + d(z, \hat{x}) \leq \frac{\epsilon}{\rho} + d(z, x) \leq 2\frac{\epsilon}{\rho} + d(x^+, x).$$

Hence we obtain

$$f(\hat{x}) \leq f(x^+) + \eta\left(\frac{\epsilon}{\rho} + d(x^+, x)\right)^2 + \frac{\eta}{2}d^2(x^+, x),$$

and

$$|\nabla f|(\hat{x}) \leq \rho + \eta\left(\frac{\epsilon}{\rho} + d(x^+, x)\right) + \eta \cdot d(\hat{x}, x) \leq \left(\rho + \frac{3\eta\epsilon}{\rho}\right) + 3\eta \cdot d(x^+, x).$$

Minimizing the right-hand-side of the last inequality in $\rho > 0$ yields the choice $\rho = \sqrt{3\eta\epsilon}$. The result follows. □

An immediate consequence of Theorem 5.1 is a subsequence converge result analogous to Corollary 3.3.

Corollary 5.3 (Subsequence convergence under near-optimality) *Consider a sequence of points x_k and closed functions $f_{x_k} \colon \mathcal{X} \to \overline{\mathbf{R}}$ satisfying $d(x_{k+1}, x_k) \to 0$ and $f(x_{k+1}) \leq \inf f_{x_k} + \epsilon_k$ for some sequence $\epsilon_k \to 0$. Suppose moreover that the inequality*

$$|f_{x_k}(y) - f(y)| \leq \omega(d(y, x_k)) \qquad \textit{holds for all indices } k \textit{ and points } y \in \mathcal{X},$$

where ω is a proper growth function. If $(x^*, f(x^*))$ is a limit point of the sequence $(x_k, f(x_k))$, then x^* is stationary for f.

Proof The proof is virtually identical to the proof of Corollary 3.3, except that Theorem 5.1 replaces Theorem 3.1 with $\rho_k = \sqrt{\epsilon_k}$. We leave the details to the reader. □

5.2 Model decrease as a stopping criterion

The underlying premise of our work so far is that the step-size $d(x_{k+1}, x_k)$ can be reliably used to terminate the model-based algorithm in the sense of Theorem 3.1. We now prove that the same can be said for termination criteria based on the model decrease $\Delta_x := f(x_k) - \inf f_{x_k}$. Indeed, this follows quickly by setting $x^+ := x$, $\epsilon := \sqrt{\Delta_x}$, and ρ a multiple of $\sqrt{\Delta_x}$ in Theorem 5.1.

Corollary 5.4 (Perturbation result for model decrease) *Consider a closed function* $f_x\colon \mathcal{X} \to \overline{\mathbf{R}}$ *such that the inequality*

$$|f_x(y) - f(y)| \le \omega(d(x,y)) \quad \text{holds for all } y \in \mathcal{X},$$

where ω *is some growth function. Define the model decrease* $\Delta_x := f(x) - \inf f_x$. *Then for any constant* $c > 0$, *there exist two points* z *and* $\hat{x}$ *satisfying the following.*

1. **(point proximity)** *The inequalities*

$$d(x,z) \le c^{-1}\sqrt{\Delta_x} \quad \text{and} \quad d(z,\hat{x}) \le 2 \cdot \frac{\omega(d(z,x))}{\omega'(d(z,x))} \quad \text{hold},$$

   under the convention $\frac{0}{0} = 0$,
2. **(value proximity)** $f(\hat{x}) \le f(x) + 2\omega(d(z,x))$,
3. **(near-stationarity)** $|\nabla f|(\hat{x}) \le c\sqrt{\Delta_x} + \omega'(d(z,x)) + \omega'(d(\hat{x},x))$.

Proof Simply set $x^+ := x$, $\epsilon := \sqrt{\Delta_x}$, and $\rho = c\sqrt{\Delta_x}$ in Theorem 5.1. □

To better internalize the estimates, let us look at the case when ω is a quadratic.

Corollary 5.5 (Perturbation for model decrease with quadratic error) *Consider a closed function* $f_x\colon \mathcal{X} \to \overline{\mathbf{R}}$ *and suppose that with some real* $\eta > 0$ *the inequality*

$$|f_x(y) - f(y)| \le \frac{\eta}{2} d^2(x,y) \quad \text{holds for all } y \in \mathcal{X}.$$

Define the model decrease

$$\Delta_x := f(x) - \inf_y f_x(y).$$

Then there exists a point $\hat{x}$ *satisfying*

1. **(point proximity)** $d(\hat{x},x) \le \sqrt{\frac{4}{3\eta}} \cdot \sqrt{\Delta_x}$,
2. **(value proximity)** $f(\hat{x}) \le f(x) + \frac{1}{3} \cdot \Delta_x$,
3. **(near-stationarity)** $|\nabla f|(\hat{x}) \le \sqrt{12\eta} \cdot \sqrt{\Delta_x}$.

Proof Simply set $x^+ := x$ and $\epsilon := \sqrt{\Delta_x}$ in Corollary 5.2. □

The subsequential convergence result in Corollary 5.3 assumes that the step-sizes $d(x_{k+1}, x_k)$ tend to zero. Now, it is easy to see that an analogous conclusion holds if instead the model decreases $f(x_k) - f_{x_k}(x_{k+1})$ tend to zero.

Corollary 5.6 (Subsequence convergence under approximate optimality II) *Consider a sequence of points* x_k *and closed functions* $f_{x_k}\colon \mathcal{X} \to \overline{\mathbf{R}}$ *satisfying* $f_{x_k}(x_{k+1}) \le \inf f_{x_k} + \epsilon_k$ *for some sequence* $\epsilon_k \to 0$. *Suppose that the inequality*

$$|f_{x_k}(y) - f(y)| \le \omega(d(y,x_k)) \quad \text{holds for all indices } k \text{ and points } y \in \mathcal{X},$$

where ω is a proper growth function. Suppose moreover that the model decreases $f(x_k) - f_{x_k}(x_{k+1})$ tend to zero. If $(x^*, f(x^*))$ is a limit point of the sequence $(x_k, f(x_k))$, then x^* is stationary for f.

Following the pattern of the previous sections, we next pass to error-bounds. The following result shows that the slope error-bound implies that, not only do the step-sizes $d(x_{k+1}, x_k)$ linearly bound the distance of x_k to the stationary-point set (Theorem 3.5), but so do the values $\sqrt{f(x_k) - \inf f_{x_k}}$.

Corollary 5.7 (Slope and model error-bounds) *Let S be an arbitrary set and fix a point $x^* \in S$ satisfying the condition:*

– **(Slope error-bound)** $\operatorname{dist}(x; S) \leq L \cdot |\nabla f|(x)$ *for all* $x \in \mathbf{B}_\gamma(x^*)$.

Consider a closed function $f_x \colon \mathcal{X} \to \overline{\mathbf{R}}$ and suppose that for some $\eta > 0$ the inequality

$$|f_x(y) - f(y)| \leq \frac{\eta}{2} d^2(y, x) \qquad \text{holds for all } y \in \mathcal{X}.$$

Then the following holds:

– **(Model error-bound)**

$$\operatorname{dist}(x; S) \leq \left(L\sqrt{12\eta} + \frac{2}{\sqrt{3\eta}} \right) \cdot \sqrt{f(x) - \inf f_x},$$

whenever $f(x) - \inf f_x < \frac{3\eta\gamma^2}{16}$ and x lies in $\mathbf{B}_{\gamma/2}(x^*)$.

Proof Suppose the inequality $f(x) - \inf f_x < \frac{3\eta\gamma^2}{16}$ holds and x lies in $\mathbf{B}_{\epsilon/2}(x^*)$. Define $\Delta_x := f(x) - \inf f_x$ and let $\hat{x}$ be the point guaranteed to exist by Corollary 5.5. We deduce

$$d(\hat{x}, x^*) \leq d(\hat{x}, x) + d(x, x^*) \leq \sqrt{\frac{4}{3\eta}} \cdot \sqrt{\Delta_x} + d(x, x^*) < \gamma.$$

Thus $\hat{x}$ lies in $\mathbf{B}_\gamma(x^*)$ and we obtain

$$L \cdot |\nabla f|(\hat{x}) \geq \operatorname{dist}\left(\hat{x}; S\right) \geq \operatorname{dist}(x; S) - d(x, \hat{x}) \geq \operatorname{dist}(x; S) - \sqrt{\frac{4}{3\eta}} \cdot \sqrt{\Delta_x}.$$

Taking into account the inequality $|\nabla f|(\hat{x}) \leq \sqrt{12\eta} \cdot \sqrt{\Delta_x}$, the result follows. □

Finally in the inexact regime, the slope error-bound (as in Theorem 3.5) implies an *inexact* error-bound condition.

Corollary 5.8 (Error-bounds under approximate optimality) *Let S be an arbitrary set and fix a point $x^* \in S$ satisfying the condition*

– **(Slope error-bound)** $\operatorname{dist}(x; S) \le L \cdot |\nabla f|(x) \quad$ *for all* $\quad x \in \mathbf{B}_\gamma(x^*)$.

Consider a closed function $f_x \colon \mathcal{X} \to \overline{\mathbf{R}}$ *and suppose that for some* $\eta > 0$ *the inequality*

$$|f_x(y) - f(y)| \le \frac{\eta}{2} d^2(y, x) \qquad \text{holds for all } y \in \mathcal{X}.$$

Define the constant $\mu := 2\sqrt{L(5L\eta + 4)}$. *Then letting* x^+ *be any point satisfying* $f_x(x^+) \le \inf f_x + \epsilon$, *the following two error-bounds hold:*

– **(Step-size error-bound)**

$$\operatorname{dist}(x; S) \le \mu\sqrt{\epsilon} + (7L\eta + 6) \cdot d(x^+, x)$$

whenever $\sqrt{\epsilon} < \gamma\mu/12L$, $d(x^+, x) < \gamma/9$, *and* x^+ *lies in* $\mathbf{B}_{\gamma/3}(x^*)$.

– **(Model error-bound)**

$$\operatorname{dist}(x; S) \le \left(L\sqrt{12\eta} + \frac{2}{\sqrt{3\eta}} \right) \sqrt{f(x) - f_x(x_+) + \epsilon}.$$

whenever $f(x) - \inf f_x < \frac{3\eta\gamma^2}{16}$ *and* x *lies in* $\mathbf{B}_{\gamma/2}(x^*)$.

Proof Consider two points x, x^+ satisfying $\sqrt{\epsilon} \le \gamma\mu/12L$, $d(x^+, x) < \gamma/9$, and $x^+ \in \mathbf{B}_{\gamma/3}(x^*)$. Let $\hat{x}, z$ be the points guaranteed to exist by Corollary 3.2 for some ρ; we will decide on the value of $\rho > 0$ momentarily. First, easy manipulations using the triangle inequality yield

$$\begin{aligned} &d(\hat{x}, z) \le d(z, x), && d(z, x^+) \le d(z, x) + d(x^+, x), \\ &d(z, x) \le \epsilon/\rho + d(x^+, x), && d(x^+, \hat{x}) \le 4\epsilon/\rho + 5d(x^+, x). \end{aligned}$$

Suppose for the moment $\hat{x}$ lies in $\mathbf{B}_\gamma(x^+)$; we will show after choosing ρ appropriately that this is the case. Then we obtain the inequality

$$\begin{aligned} L \cdot |\nabla f|(\hat{x}) \ge \operatorname{dist}\left(\hat{x}; S\right) &\ge \operatorname{dist}(x; S) - d(x^+, \hat{x}) - d(x^+, x) \\ &\ge \operatorname{dist}(x; S) - 4\epsilon/\rho_k - 6d(x^+, x). \end{aligned}$$

Taking into account the inequality

$$|\nabla f|(\hat{x}) \le \rho + \eta(d(z, x) + d(\hat{x}, x)) \le \rho + \eta(5\epsilon/\rho + 7d(x^+, x)),$$

we conclude

$$\operatorname{dist}(x; S) \le L\rho + \frac{5L\eta\epsilon}{\rho} + \frac{4\epsilon}{\rho} + (7L\eta + 6) \cdot d(x^+, x),$$

as claimed. Minimizing the right-hand-side in ρ yields $\rho := \sqrt{\frac{(5L\eta+4)\epsilon}{L}}$. With this choice, the inequality above becomes

$$\operatorname{dist}(x; S) \leq 2\sqrt{L(5L\eta + 4)\epsilon} + (7L\eta + 6) \cdot d(x^+, x).$$

Finally, let us verify that $\hat{x}$ indeed lies in $\mathbf{B}_\gamma(x^*)$. To see this, simply observe

$$\begin{aligned} d(\hat{x}, x^*) &\leq d(\hat{x}, z) + d(z, x^+) + d(x^+, x^*) \leq 2d(z, x) + d(x^+, x) + d(x^+, x^*) \\ &\leq 2\epsilon_k/\rho_k + 3d(x^+, x) + d(x^+, x^*) < \gamma. \end{aligned}$$

The result follows. The step-size error bound condition follows. The functional error-bound is immediate from Corollary 5.7. □

In particular, in the notation of Corollary 5.8, if one wishes the error $d(x^+, x)$ to linearly bound the distance $d(x; S)$, then one should ensure that the tolerance ϵ is on the order of $d^2(x^+, x)$.

5.3 Near-stationarity for the subproblems

In this section, we explore the setting where x^+ is only ϵ-stationary for f_x. To make progress in this regime, however, we must first assume a linear structure on the metric space. We suppose throughout that $\mathcal{X}$ is a Banach space, and denote its dual by $\mathcal{X}^*$. For any dual element $v \in \mathcal{X}^*$ and a point $x \in \mathcal{X}$, we use the notation $\langle v, x\rangle := v(x)$. Second, the property $|\nabla f_x|(x^+) \leq \epsilon$ alone appears to be too weak. Instead, we will require a type of uniformity in the slopes. In the simplest case, we will assume that x^+ is such that the function f_x majorizes the simple quadratic

$$f_x(x^+) + \langle v, \cdot - x^+\rangle - \eta\|\cdot - x^+\|^2$$

where $v \in \mathcal{X}^*$ is some dual element satisfying $\|v\| \leq \epsilon$. In the language of variational analysis, v is a *proximal subgradient* of f_x at x^+; see e.g. [15,56]. A quick computation immediately shows the inequality $|\nabla f_x|(x^+) \leq \epsilon$. Assuming that η is uniform throughout the iterative process will allow us to generalize the results of Sect. 3. Such uniformity is immediately implied by prox-regularity [50] for example—a broad and common setting for nonsmooth optimization.

Corollary 5.9 (Perturbation result under approximate stationarity) *Consider a closed function $f_x : \mathcal{X} \to \overline{\mathbf{R}}$ on a Banach space $\mathcal{X}$ such that the inequality*

$$|f_x(y) - f(y)| \leq \omega_1(d(x, y)) \quad \textit{holds for all } y \in \mathcal{X},$$

where ω_1 is some growth function. Suppose moreover that for some point $x^+ \in \mathcal{X}$, a dual element $v \in \mathcal{X}^*$, and a growth function ω_2, the inequality

$$f_x(y) \geq f_x(x^+) + \langle v, y - x^+\rangle - \omega_2(d(y, x^+)) \qquad \textit{holds for all } y \in \mathcal{X}.$$

Then there exists a point $\hat{x}$ satisfying

1. **(point proximity)** $d(x^+, \hat{x}) \leq 2 \cdot \frac{\omega_1(d(x^+,x))}{\omega_1'(d(x^+,x))}$,
2. **(value proximity)**

$$f(\hat{x}) \leq f(x^+) + \langle v, \hat{x} - x^+ \rangle + \omega_1(d(x^+, x)) - \omega_2(d(\hat{x}, x^+)),$$

3. **(near-stationarity)**

$$|\nabla f|(\hat{x}) \leq \|v\| + \omega_1'(d(x^+, x)) + \omega_1'(d(\hat{x}, x)) + \omega_2'(d(\hat{x}, x^+)).$$

Proof Define the functions $\widetilde{f}(y) := f(y) - \langle v, y - x^+ \rangle + \omega_2(d(y, x^+))$ and $\widetilde{f}_x(y) := f_x(y) - \langle v, y - x^+ \rangle + \omega_2(d(y, x^+))$. Note that x^+ minimizes $\widetilde{f}_x$ and that the inequality

$$|\widetilde{f}_x(y) - \widetilde{f}(y)| \leq \omega_1(d(x, y)) \quad \text{holds for all } y \in \mathcal{X}.$$

Applying Theorem 3.1, we obtain a point $\hat{x}$ satisfying the point proximity claim, along with the inequalities $\widetilde{f}(\hat{x}) \leq \widetilde{f}(x^+) + \omega_1(d(x^+, x))$ and $|\nabla \widetilde{f}|(\hat{x}) \leq \omega_1'(d(x^+, x)) + \omega_1'(d(\hat{x}, x))$. The value proximity claim follows directly from definitions, while the near-stationarity is immediate from the inequality, $|\nabla \widetilde{f}|(\hat{x}) \geq |\nabla f|(\hat{x}) - \|v\| - \omega_2'(d(\hat{x}, x^+))$. The result follows. □

As an immediate consequence, we obtain the subsequence convergence result.

Corollary 5.10 (Convergence under approximate optimality) *Consider a sequence of points x_k and closed functions $f_{x_k} : \mathcal{X} \to \overline{\mathbf{R}}$ satisfying*

$$|f_{x_k}(y) - f(y)| \leq \omega_1(d(y, x_k)) \qquad \text{for all indices } k \text{ and points } y \in \mathcal{X},$$

where ω_1 is some proper growth function. Suppose that the inequality

$$f_{x_k}(y) \geq f_{x_k}(x_{k+1}) + \langle v_{k+1}, y - x_{k+1} \rangle - \omega_2(d(y, x_{k+1}))$$

holds for all k and all $y \in \mathcal{X}$, where ω_2 is some proper growth function and $v_k \in \mathcal{X}^*$ are some dual elements. Assume moreover that $d(x_{k+1}, x_k)$ and $\|v_k\|$ tend to zero. If $(x^*, f(x^*))$ is a limit point of the sequence $(x_k, f(x_k))$, then x^* is stationary for f.

Finally, the following inexact error-bound result holds, akin to Theorem 3.5.

Corollary 5.11 (Error-bounds under approximate stationarity) *Let S be an arbitrary set and fix a point $x^* \in S$ satisfying the condition*

– **(Slope error-bound)** $\operatorname{dist}(x, S) \leq L \cdot |\nabla f|(x) \quad$ *for all* $\quad x \in \mathbf{B}_\gamma(x^*)$.

Consider a closed function $f_x : \mathcal{X} \to \overline{\mathbf{R}}$ and suppose that for some $\eta > 0$ the inequality

$$|f_x(y) - f(y)| \leq \frac{\eta}{2} \cdot \|y - x\|^2 \qquad \text{holds for all } y \in \mathcal{X}.$$

Fix a point x^+ and a dual element $v \in \mathcal{X}^*$ so that the inequality

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

Then the approximate error-bound holds:

– **(Step-size error-bound)**

$$\operatorname{dist}(x; S) \leq L\|v\| + (4\eta L + 2)\|x^+ - x\| \quad \textit{when } x, x^+ \in \mathbf{B}_{\gamma/3}(x^*).$$

Proof The proof is entirely analogous to that of Theorem 3.5. Consider two points $x, x^+ \in \mathbf{B}_{\gamma/3}(x^*)$. Let $\hat{x}$ be the point guaranteed to exist by Corollary 5.9. We deduce

$$d(\hat{x}, x^*) \leq d(\hat{x}, x^+) + d(x^+, x^*) \leq d(x^+, x) + d(x^+, x^*) < \gamma.$$

Thus $\hat{x}$ lies in $\mathbf{B}_\epsilon(x^*)$ and we deduce

$$\begin{aligned} L \cdot |\nabla f|(\hat{x}) \geq \operatorname{dist}\left(\hat{x}; S\right) &\geq \operatorname{dist}(x; S) - d(x^+, \hat{x}) - d(x^+, x) \\ &\geq \operatorname{dist}(x; S) - 2d(x^+, x). \end{aligned}$$

Taking into account the inequality $|\nabla f|(\hat{x}) \leq \|v\| + 4\eta\|x^+ - x\|$, we conclude

$$\operatorname{dist}(x; S) \leq L\|v\| + (4\eta L + 2) \cdot \|x^+ - x\|,$$

as claimed. □

Conclusion

In this paper, we considered a general class of nonsmooth minimization algorithms that use Taylor-like models. We showed that both the step-size and the decrease in the model's value can be used as reliable stopping criteria. We deduced subsequence convergence to stationary points, and error-bound conditions under natural regularity properties of the function. The results fully generalized to the regime where the models are minimized inexactly. Ekeland's variation principle (Theorem 2.2) underlies all of our current work. Despite the wide uses of the principle in variational analysis, its impact on convergence of basic algorithms, such as those covered here and in [21,23], is not as commonplace as it should be. We believe that this work takes an important step towards rectifying this disparity and the techniques presented here will pave the way for future algorithmic insight.

Acknowledgements We thank the two anonymous referees and the Associate Editor for their insightful comments, which have improved the exposition of this work.

References

1. Aragón Artacho, F.J., Geoffroy, M.H.: Characterization of metric regularity of subdifferentials. J. Convex Anal. **15**(2), 365–380 (2008)
2. Attouch, H., Bolte, J., Svaiter, B.F.: Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods. Math. Program. **137**(1–2, Ser. A), 91–129 (2013)
3. Bai, Y., Duchi, J., Mei, S.: Proximal algorithms for constrained composite optimization, with applications to solving low-rank SDPs (2019). Preprint arXiv:1903.00184
4. Beck, A., Teboulle, M.: A fast iterative shrinkage–thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)
5. Bolte, J., Daniilidis, A., Lewis, A.S., Shiota, M.: Clarke subgradients of stratifiable functions. SIAM J. Optim. **18**(2), 556–572 (2007)
6. Bolte, J., Daniilidis, A., Ley, O., Mazet, L.: Characterizations of łojasiewicz inequalities: subgradient flows, talweg, convexity. Trans. Am. Math. Soc. **362**(6), 3319–3363 (2010)
7. Bolte, J., Nguyen, T.P., Peypouquet, J., Suter, B.: From error bounds to the complexity of first-order descent methods for convex functions. Math. Program. **165**(2), 471–507 (2017)
8. Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math. Program. **146**(1–2, Ser. A), 459–494 (2014)
9. Burke, J.V.: Descent methods for composite nondifferentiable optimization problems. Math. Program. **33**(3), 260–279 (1985)
10. Burke, J.V., Ferris, M.C.: A Gauss–Newton method for convex composite optimization. Math. Program. **71**(2), 179–194 (1995)
11. Burke, J.V., Lewis, A.S., Overton, M.L.: A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. SIAM J. Optim. **15**(3), 751–779 (2005)
12. Byrd, R.H., Nocedal, J., Oztoprak, F.: An inexact successive quadratic approximation method for l-1 regularized optimization. Math. Program. **157**(2), 375–396 (2016)
13. Cartis, C., Gould, N.I.M., Toint, P.L.: On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming. SIAM J. Optim. **21**(4), 1721–1739 (2011)
14. Charisopoulos, V., Davis, D., Díaz, M., Drusvyatskiy, D.: Composite optimization for robust blind deconvolution (2019). arXiv preprint arXiv:1901.01624
15. Clarke, F.H., Ledyaev, Y., Stern, R.I., Wolenski, P.R.: Nonsmooth Analysis and Control Theory. Texts in Mathematics, vol. 178. Springer, New York (1998)
16. Davis, D., Drusvyatskiy, D.: Stochastic model-based minimization of weakly convex functions. SIAM J. Optim. **29**(1), 207–239 (2019)
17. De Giorgi, E., Marino, A., Tosques, M.: Problemi di evoluzione in spazi metrici e curve di massima pendenza. Atti Acad. Nat. Lincei Rend. Cl. Sci. Fiz. Mat. Natur. **68**, 180–187 (1980)
18. Drusvyatskiy, D.: Slope and geometry in variational mathematics. PhD thesis, Cornell University (2013)
19. Drusvyatskiy, D., Ioffe, A.D.: Quadratic growth and critical point stability of semi-algebraic functions. Math. Program. **153**(2, Ser. A), 635–653 (2015)
20. Drusvyatskiy, D., Ioffe, A.D., Lewis, A.S.: Curves of descent. SIAM J. Control Optim. **53**(1), 114–138 (2015)
21. Drusvyatskiy, D., Ioffe, A.D., Lewis, A.S.: Transversality and alternating projections for nonconvex sets. Found. Comput. Math. **15**(6), 1637–1651 (2015)
22. Drusvyatskiy, D., Ioffe, A.D., Lewis, A.S.: Nonsmooth optimization using Taylor-like models: error bounds, convergence, and termination criteria (2016). arXiv:1610.03446 (Ver. 1)
23. Drusvyatskiy, D., Lewis, A.S.: Error bounds, quadratic growth, and linear convergence of proximal methods. Math. Oper. Res. **43**(3), 919–948 (2018)
24. Drusvyatskiy, D., Mordukhovich, B.S., Nghia, T.T.A.: Second-order growth, tilt stability, and metric regularity of the subdifferential. J. Convex Anal. **21**(4), 1165–1192 (2014)
25. Drusvyatskiy, D., Paquette, C.: Efficiency of minimizing compositions of convex functions and smooth maps. Math. Prog. (2016). https://doi.org/10.1007/s1010, arXiv:1605.00125
26. Duchi, J.C., Ruan, F.: Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. Inf. Infer. J. IMA **8**(3), 471–529 (2018)
27. Ekeland, I.: On the variational principle. J. Math. Anal. Appl. **47**, 324–353 (1974)

28. Fletcher, R.: A model algorithm for composite nondifferentiable optimization problems. In: Sorensen, D.C., Wets, R.J.B. (eds.) Nondifferential and Variational Techniques in Optimization (Lexington, Ky., 1980). Mathematical Programming Studies, vol. 17, pp. 67–76. Springer, Berlin (1982)
29. Fletcher, R.: A model algorithm for composite nondifferentiable optimization problems. In: Sorensen, D.C., Wets, R.J.B. (eds.) Nondifferential and Variational Techniques in Optimization, pp. 67–76. Springer, Berlin (1982)
30. Geiping, J., Moeller, M.: Composite optimization by nonconvex majorization–minimization. SIAM J. Imaging Sci. **11**(4), 2494–2528 (2018)
31. Ghadimi, S., Lan, G.: Accelerated gradient methods for nonconvex nonlinear and stochastic programming. Math. Program. **156**(1–2, Ser. A), 59–99 (2016)
32. Goldstein, A.A.: Optimization of Lipschitz continuous functions. Math. Program. **13**(1), 14–22 (1977)
33. Ioffe, A.D.: Metric regularity and subdifferential calculus. Uspekhi Mat. Nauk **55**(3(333)), 103–162 (2000)
34. Ioffe, A.D.: Variational Analysis of Regular Mappings. Springer Monographs in Mathematics. Springer, Berlin (2017)
35. Jin, C., Netrapalli, P., Jordan, M.I.: Minmax optimization: stable limit points of gradient descent ascent are locally optimal (2019). arXiv preprint arXiv:1902.00618
36. Klatte, D., Kummer, B.: Nonsmooth Equations in Optimization: Regularity, Calculus, Methods and Applications. Nonconvex Optimization and Its Applications, vol. 60. Kluwer Academic Publishers, Dordrecht (2002)
37. Kurdyka, K.: On gradients of functions definable in o-minimal structures. Ann. Inst. Fourier (Grenoble) **48**(3), 769–783 (1998)
38. Lewis, A.S., Wright, S.J.: A proximal method for composite minimization. Math. Program. **158**, 1–46 (2015)
39. Luo, Z.-Q., Tseng, P.: 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
40. Martinet, B.: Régularisation d'inéquations variationnelles par approximations successives. Rev. Française Informat. Recherche Opérationnelle **4**(Ser. R–3), 154–158 (1970)
41. Martinet, B.: Détermination approchée d'un point fixe d'une application pseudo-contractante. Cas de l'application prox. C. R. Acad. Sci. Paris Sér. A-B **274**, A163–A165 (1972)
42. Nesterov, Y., Polyak, B.T.: Cubic regularization of Newton method and its global performance. Math. Program. **108**(1, Ser. A), 177–205 (2006)
43. Nesterov, Y.: A method for solving the convex programming problem with convergence rate $O(1/k^2)$. Dokl. Akad. Nauk SSSR **269**(3), 543–547 (1983)
44. Nesterov, Y.: Modified Gauss–Newton scheme with worst case guarantees for global performance. Optim. Methods Softw. **22**(3), 469–483 (2007)
45. Nesterov, Y.: Accelerating the cubic regularization of Newton's method on convex problems. Math. Program. **112**(1, Ser. B), 159–181 (2008)
46. Nesterov, Y.: Gradient methods for minimizing composite functions. Math. Program. **140**(1, Ser. B), 125–161 (2013)
47. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer Series in Operations Research and Financial Engineering, 2nd edn. Springer, New York (2006)
48. Noll, D., Prot, O., Rondepierre, A.: A proximity control algorithm to minimize nonsmooth and nonconvex functions. Pac. J. Optim. **4**(3), 571–604 (2008)
49. Ochs, P., Fadili, J., Brox, T.: Non-smooth non-convex Bregman minimization: unification and new algorithms. J. Optim. Theory Appl. **181**, 1–35 (2017)
50. Poliquin, R.A., Rockafellar, R.T.: Prox-regular functions in variational analysis. Trans. Am. Math. Soc. **348**, 1805–1838 (1996)
51. Polyak, B.T.: Gradient methods for the minimisation of functionals. USSR Comput. Math. Math. Phys. **3**(4), 864–878 (1963)
52. Powell, M.J.D.: General algorithms for discrete nonlinear approximation calculations. In: Chui, C.K., Schumaker, L.L., Ward, J.D. (eds.) Approximation Theory, IV (College Station, Tex., 1983), pp. 187–218. Academic Press, New York (1983)
53. Powell, M.J.D.: On the global convergence of trust region algorithms for unconstrained minimization. Math. Program. **29**(3), 297–303 (1984)
54. Rafique, H., Liu, M., Lin, Q., Yang, T.: Non-convex min–max optimization: provable algorithms and applications in machine learning (2018). arXiv preprint arXiv:1810.02060

55. Rockafellar, R.T.: Monotone operators and the proximal point algorithm. SIAM J. Control Optim. **14**(5), 877–898 (1976)
56. Rockafellar, R.T.: Proximal subgradients, marginal values, and augmented Lagrangians in nonconvex optimization. Math. Oper. Res. **6**(3), 424–436 (1981)
57. Rockafellar, R.T., Dontchev, A.L.: Implicit Functions and Solution Mappings. Monographs in Mathematics. Springer, Berlin (2009)
58. Scheinberg, K., Tang, X.: Practical Inexact Proximal Quasi-Newton Method with Global Complexity Analysis. Mathematical Programming, pp. 1–35. Springer, Berlin (2016)
59. Wild, S.M.: Solving Derivative-Free Nonlinear Least Squares Problems with POUNDERS. Argonne National Lab, Lemont (2014)
60. Wright, S.J.: Convergence of an inexact algorithm for composite nonsmooth optimization. IMA J. Numer. Anal. **10**(3), 299–321 (1990)
61. Yuan, Y.: On the superlinear convergence of a trust region algorithm for nonsmooth optimization. Math. Program. **31**(3), 269–285 (1985)
62. Zhang, R., Treiman, J.: Upper-Lipschitz multifunctions and inverse subdifferentials. Nonlinear Anal. **24**(2), 273–286 (1995)