A nearly linearly convergent first-order method for nonsmooth functions with quadratic growth

Damek Davis* Liwei Jiang†

Abstract

Classical results show that gradient descent converges linearly to minimizers of smooth strongly convex functions. A natural question is whether there exists a locally nearly linearly convergent method for nonsmooth functions with quadratic growth. This work designs such a method for a wide class of nonsmooth and nonconvex locally Lipschitz functions, including max-of-smooth, Shapiro's decomposable class, and generic semialgebraic functions. The algorithm is parameter-free and derives from Goldstein's conceptual subgradient method.

1 Introduction

Slow sublinear convergence of first-order methods in nonsmooth optimization is often illustrated with the following simple strongly convex function:

$$f(x) = \max_{1\leq i\leq m} x_i + \frac{1}{2}\|x\|^2 \qquad \text{for some } m \leq d \text{ and all } x \in \mathbb{R}^d. \tag{1.1}$$

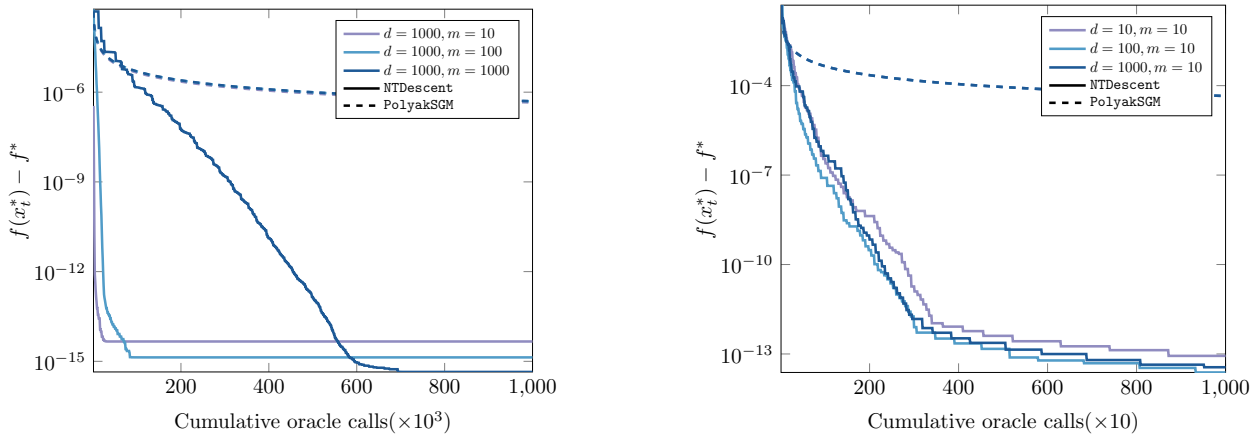
For example, consider the subgradient method applied to f, which generates iterates x_k. Since f is strongly convex, classical results dictate that $f(x_k) - \inf f = O(k^{-1})$. On the other hand, under proper initialization and an adversarial first-order oracle, there is a matching lower bound for the first m iterations: $f(x_k) - \inf f \geq (2k)^{-1}$ for all $k \leq m$; see [8, 36]. Beyond the subgradient method, the lower bound also holds for any algorithm whose kth iterate lies within the linear span of the initial iterate and past $k-1$ computed subgradients. Thus, one must make more than m first-order *oracle calls* to f, i.e., function and subgradient evaluations, before possibly seeing improved convergence behavior.

While such methods make little progress when $k \leq m$, this behavior may or may not continue for $k \gg m$. On one extreme, the subgradient method, continues to converge slowly even when equipped with the popular Polyak stepsize (`PolyakSGM`) [38]; see Figure 1. On the opposite extreme, more sophisticated algorithms such as the center of gravity method

*School of Operations Research and Information Engineering, Cornell University. Ithaca, NY 14850, USA; people.orie.cornell.edu/dsd95/. Research of Davis supported by an Alfred P. Sloan research fellowship and NSF DMS award 2047637.

†School of Operations Research and Information Engineering, Cornell University. Ithaca, NY 14850, USA; orie.cornell.edu/research/grad-students/liwei-jiang

or the ellipsoid method converge linearly, but their complexity scales with the dimension of the problem, a necessary consequence of the linear rate of convergence; see the discussion in [8, Chapter 2].



(a) (b)

Figure 1: Comparison of `NTDescent` with `PolyakSGM` on (1.1). Left: we fix d and vary m; Right: we fix m and vary d. For both algorithms the value $f(x_t^*)$ denotes the best function seen after t oracle evaluations.

A natural question is whether there exists a first-order method whose behavior lies in between these two extremes, at least for nonsmooth functions f satisfying regularity conditions at local minimizers. Regularity conditions often take the form of growth – linear or quadratic – away from minimizers. Well-known results show that subgradient methods converge linearly on nonsmooth functions with linear (also called *sharp*) growth [38]. On the other hand, in smooth convex optimization quadratic growth entails linear convergence of gradient methods. However, to the best of our knowledge, no parallel result for nonsmooth functions with quadratic growth exists. Thus, in this work, we ask

> is there a locally nearly linearly convergent method for nonsmooth functions with quadratic growth whose rate of convergence and region of rapid local convergence solely depends on f?

Let us explain the qualifiers "nearly" and the "solely depends on the function." First, the qualifier "nearly" signifies that the method locally achieves a function gap of size ε using at most, say, $O(C_f \log^3(1/\varepsilon))$ first-order oracle evaluations of f, where C_f depends on f. Second, the qualifier "solely depends on the function," signifies that C_f and the size of the region of local convergence do not depend on the dimension of the problem, but instead depend only on the function f through intrinsic quantities, such as Lipschitz and quadratic growth constants.

In this work, we positively answer the above question for a class nonsmooth optimization problems with quadratic growth. The method we develop is called *Normal Tangent Descent* (`NTDescent`). We formally describe `NTDescent` in Section 1.4. For now we illustrate the performance of `NTDescent` on f from (1.1) in Figure 1. In both plots, we see `NTDescent`

improves on the performance of `PolyakSGM`, measured in terms of oracle calls. This is a fair basis of comparison since both `PolyakSGM` and `NTDescent` perform a similar amount of computation per oracle call. Figure 1b also shows that the performance of `NTDescent` is dimension independent. We highlight that this performance was achieved without any tuning of parameters for `NTDescent`. Indeed, our main theoretical guarantees for `NTDescent` (Theorem 1.1) do not require the user to set any parameters.

The problem class on which `NTDescent` succeeds consists of locally Lipschitz nonsmooth functions with quadratic growth and a certain *smooth substructure* at local minimizers. Importantly, we do not assume the problems under consideration are convex, though convexity entails improved guarantees. Two example classes with such smooth substructure include (i) "generic" semialgebraic functions and (ii) properly C^p decomposable loss functions satisfying strict complementarity and quadratic growth conditions [40]. A semialgebraic function is one whose graph is the finite union of intersections of polynomial inequalities. Semialgebraic functions (more generally *tame* [25] functions) model most problems of interest in applications. If f is semialgebraic, for a full Lebesgue measure set of $w \in \mathbb{R}^d$, we will show that the tilted function $f_w \colon x \mapsto f(x) + w^\top x$ has quadratic growth and the desired smooth substructure at each local minimizer, explaining the qualifier "generic." On the other hand, a properly C^p decomposable function is one that decomposes near local minimizers as a composition of a positively homogeneous convex function with a smooth mapping that maps the minimizer to the origin. Decomposable functions appear often in practice, e.g., in eigenvalue and data fitting problems. An important subclass of decomposable functions consists of so-called "max-of-smooth" functions, which are the maximum of finitely many smooth functions that satisfy certain regularity conditions at minimizers, e.g., f in (1.1).

The precise smooth structure used in this work was recently identified in [15], where it was shown to be available in generic semialgebraic and decomposable problems. Since it is available in many problems of interest, throughout this introduction we call this the combination of quadratic growth and smooth substructure *typical structure* and call functions possessing this combined structure *typical*. We present the formal structure in Section 3. At the heart of this structure is a distinguished smooth manifold $\mathcal{M}$ – called the *active manifold* – containing a local minimizer of interest. We formally define the active manifold concept in Definition 1.2, but at a high-level the two crucial characteristics are that (i) along the manifold, the function f is smooth and (ii) normal to the manifold, the function grows sharply. For example, Figure 2 depicts the nonsmooth function $f(u, v) = u^2 + |v|$ for which the u-axis plays the role of $\mathcal{M}$. In Section 1.3.1 we will examine this function and explain how we use its typical structure in `NTDescent`. This example also has the smooth substructure developed in several seminal works in the optimization literature, including those found in work on identifiable surfaces [44], partly smooth manifolds [30], $\mathcal{VU}$-structures [28, 34], and minimal identifiable sets [20]. However, crucial to the analysis of `NTDescent` are two further properties introduced in [15], called *strong* (a)*-regularity* and $(b_\leq)$*-regularity*. Strong (a)-regularity roughly states that the function is smooth in tangent directions to manifold up to an error term which is linear in the distance to the manifold. On the other hand, $(b_\leq)$-regularity is a one-sided uniform semismoothness [32] property that holds automatically when f is (weakly) convex. Both properties hold for the two variable example in Figure 2 and for the function in (1.1), where the active manifold is the subspace in which the first m variables take on the same value: $\mathcal{M} = \{x \in \mathbb{R}^d \colon x_1 = x_2 = \dots x_m\}$.

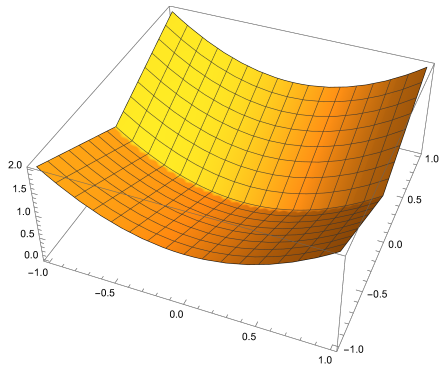


Figure 2: The function $f(u, v) = u^2 + |v|$ has typical structure.

Before turning to the description of `NTDescent`, we point out that similar smooth substructure has been utilized for first-order methods in nonsmooth optimization, most famously for functions with $\mathcal{VU}$-structure [28,34] and more recently for max-of-smooth functions.[1] For $\mathcal{VU}$ functions, so-called "bundle-methods," [27, 42] which possess an inner-outer loop structure, have been shown to converge superlinearly with respect to the number of outer-loop steps [33]; see also the survey [37]. These methods have excellent empirical performance, but a complete account of their inner-loop complexity remains elusive. On the other hand, in a recent breakthrough, Han and Lewis proposed a first-order method – Survey Descent – that converges linearly on certain strongly convex max-of-smooth objectives, stepping beyond the classical smooth setting [24]. The method shows favorable performance beyond the max-of-smooth class, e.g., on certain eigenvalue optimization problems, but no theoretical justification for this success is available. We discuss Survey Descent in more detail in Section 7.1. We now motivate `NTDescent`.

1.1 Motivation: Goldstein's conceptual subgradient method

To motivate `NTDescent` and the role of smooth substructure, let us set the stage: consider the nonsmooth optimization problem

$$\text{minimize}_{x \in \mathbb{R}^d} \; f(x),$$

where $f \colon \mathbb{R}^d \to \mathbb{R}$ is a locally Lipschitz function, which is not necessarily convex. The algorithm developed in this work assumes *first-order oracle access* to f [8, 35, 36]. In particular, at every $x \in \mathbb{R}^d$ we must be able to evaluate $f(x)$ and retrieve an element of the *Clarke subdifferential* $\partial f(x)$. Informally, the Clarke subdifferential is comprised of convex combinations of limits of gradients taken at nearby points; a formal definition appears in Section 1.7. The Clarke subdifferential reduces to the familiar objects in classical settings. For example, when f is C^1, the Clarke subdifferential reduces to the singleton mapping $\{\nabla f\}$. In addition, when f is convex, the Clarke subdifferential reduces to the subdifferential in the sense of convex analysis.

[1] Though they also benefit from smooth substructure, *proximal-methods* do not fall within the oracle model of first-order methods considered in this work. Thus, we omit them from our discussion.

The starting point of this work is the classical conceptual subgradient method of Goldstein [23]. The core object in this method is the Goldstein subdifferential:

$$\partial_\sigma f(x) := \operatorname{conv}\left(\bigcup_{y\in\overline{B}_\sigma(x)} \partial f(y)\right) \qquad \text{for all } x\in\mathbb{R}^d \text{ and } \sigma>0. \tag{1.2}$$

This subdifferential is simply the convex hull of all Clarke subgradients of f taken at points inside the ball of radius σ. Its importance arises from the following descent property proved in [23]: fix $\sigma>0$ and $x\in\mathbb{R}^d$ and let w denote the minimal norm element of $\partial_\sigma f(x)$. Then

$$f\left(x-\sigma\frac{w}{\|w\|}\right)\le f(x)-\sigma\|w\| \qquad \text{if } w\neq 0. \tag{1.3}$$

This property motivates Goldstein's conceptual subgradient method, which simply iterates:

$$x_{k+1}=x_k-\sigma\frac{w_k}{\|w_k\|} \qquad \text{where} \qquad w_k=\operatorname*{argmin}_{w\in\partial_\sigma f(x_k)}\|w\|. \tag{1.4}$$

This algorithm is remarkable since it is provably a descent method for any Lipschitz function and even converges at a sublinear rate. Indeed, a quick appeal to (1.3) yields

$$\min_{k=0,\dots,K-1}\|w_k\|\le\varepsilon \qquad \text{holds when} \qquad K=O\left(\frac{f(x_0)-\min f}{\sigma\varepsilon}\right).$$

While this exact variant of the Goldstein method is not necessarily implementable, recent work has devised approximate versions of the method that have similar sublinear convergence properties [17, 47].

The algorithm introduced in this work approximately implements the method (1.4). The goal of this work is to prove that the method is locally nearly linearly convergent on typical nonsmooth functions. To develop such a method, we must resolve two issues for this problem class. First, we must develop rapidly convergent algorithms that approximately compute the minimal norm element of the Goldstein subdifferential. Second, we must devise an appropriate regularity property that ensures the proposed method converges nearly linearly. We discuss both of these properties in turn, beginning with the regularity property, which we call the *Goldstein-Kurdyka-Łojasiewicz* (GKL) inequality.

1.2 Linear convergence via the GKL inequality

The GKL inequality is motivated by the following simple observation: if the bound

$$\sigma\|w_k\|\ge\eta(f(x_k)-\min f)$$

holds for some $\eta>0$ and all $k>0$, then the Goldstein method (1.4) converges linearly to a minimizer of f. A potential issue with this inequality is that the vector w_k is zero whenever σ is larger than the distance of x_k to the nearest critical point of f; thus the algorithm may stall whenever x_k is near enough to a minimizer. This suggests a simple relaxation of the property that allows σ to depend on x_k.

Indeed, in this work, we will provide conditions under which the following bound holds near a local minimizer $\bar{x}$ of f: there exists a constant $\eta > 0$ and a function $\sigma\colon \mathbb{R}^d \to \mathbb{R}_+$ such that for all x near $\bar{x}$, we have

$$\sigma(x)\text{dist}(0, \partial_{\sigma(x)} f(x)) \geq \eta(f(x) - f(\bar{x})). \tag{1.5}$$

We call this property the GKL inequality due to its similarity to the Kurdyka-Łojasiewicz (KL) inequality [6]. The KL inequality requires that a suitable nonlinear reparameterization $\psi\colon \mathbb{R} \to \mathbb{R}$ of the function gap is bounded by the minimal norm Clarke subgradient for all x near $\bar{x}$:

$$\text{dist}(0, \partial f(x)) \geq \psi(f(x) - f(\bar{x})).$$

In recent years, the KL inequality has played a key role in the convergence of proximal methods for nonsmooth optimization and in continuous time analogues of the subgradient method; see e.g., [2,3,5,6,45]. In contrast to the proximal and continuous-time settings, we do not know whether the KL inequality alone allows one to design a locally linear convergent discrete time subgradient method.

A well-known property of the KL inequality is its prevalence: it is valid for any lower-semicontinuous semialgebraic function f. We will show that the GKL inequality is also prevalent in the sense that it holds for the aforementioned problems with typical structure. In this way, the conceptual method (1.4) with varying $\sigma_k := \sigma(x_k)$ will locally converge linearly on such problems. The reader may wonder whether we can or must find the precise value $\sigma(x_k)$. We will show that for typical problems, an appropriate σ_k may be found through a line search procedure.

1.3 Approximately implementing Goldstein's method

The GKL inequality ensures that the conceptual Goldstein method converges linearly, provided the stepsize σ is chosen adaptively. To move beyond the conceptual setting, we must develop strategies for approximating the minimal norm element of $\partial_\sigma f(x)$ for $\sigma > 0$ and $x \in \mathbb{R}^d$. Let us suppose we have such a method and denote it by $\texttt{MinNorm}(x, \sigma)$. Then the method of this work simply iterates:

$$x_{k+1} = x_k - \sigma_k \frac{w_k}{\|w_k\|} \qquad \text{and} \qquad w_k = \texttt{MinNorm}(x_k, \sigma_k) \tag{1.6}$$

for an appropriate sequence $\sigma_k > 0$. We discuss and develop two different implementations of $\texttt{MinNorm}(x, \sigma)$ in this work. Given $x \in \mathbb{R}^d$ and $\sigma > 0$, both methods iteratively construct a sequence of Clarke subgradients $g_0, \ldots, g_{T-1}$ taken at points in the ball $\overline{B}_\sigma(x)$ and then output a "small" convex combination $w \in \text{conv}\{g_0, \ldots, g_{T-1}\}$, which satisfies the descent condition

$$f\left(x - \sigma \frac{w}{\|w\|}\right) \leq f(x) - \frac{\sigma}{8}\|w\|.$$

The oracle complexity of $\texttt{MinNorm}(x, \sigma)$ is then T function/subgradient evaluations, and we hope to ensure that T is relatively small, say, constant or at most

$$T = O\left(\log\left(\Delta_{x,\sigma}^{-1}\right)\right) \qquad \text{where } \Delta_{x,\sigma} := \text{dist}(0, \partial_\sigma f(x)).$$

Provided that T is on this order, that f satisfies the GKL inequality, and that σ_k is chosen appropriately, the iterate x_k will satisfy $f(x_k) - f(\bar{x}) \leq \varepsilon$ after at most $O(\log^2(1/\varepsilon))$ iterations, a nearly linear rate of convergence. This complexity ignores the cost of choosing an appropriate stepsize σ_k, but we will show that in typical problems we can find appropriate σ_k with at most $O(\log(1/\varepsilon))$ function/subgradient evaluations.

We are aware of two `MinNorm` type methods in the literature, but their complexity is either too large or is useful only in low dimensions problems. For example, the works [17, 47] introduced such a method for general locally Lipschitz functions. However, the complexity of the method is $T = O(1/\Delta_{x,\sigma})$ – too large for our purposes. On the other hand, the work [17] also introduced a method tailored to low-dimensional weakly convex functions, a broad class of nonconvex functions that includes all compositions of Lipschitz convex functions with smooth mappings. However, the method is based on cutting plane techniques, so its complexity scales linearly with dimension: $T = O(d \log(1/\Delta_{x,\sigma}))$.

This work develops faster `MinNorm` type methods for the aforementioned typical problems. The `MinNorm` type algorithm we develop uses the existence of the active manifold to quickly reveal an approximate minimal norm Goldstein subgradient. Let us briefly illustrate this structure and `MinNorm` type algorithms with a simple example.

1.3.1 Finding small subgradients in a simple example with "typical" structure

Consider the following simple function of two variables $f(u, v) = u^2 + |v|$, which has a unique minimizer at $\bar{x} = (0, 0)$. Here, the u-axis plays the role of the manifold $\mathcal{M}$, along which f is smooth and grows quadratically and off of which f grows sharply; see Figure 2. The manifold $\mathcal{M}$ naturally induces a decomposition of f into smooth $f_{\mathcal{U}}(u, v) = u^2$ and nonsmooth $f_{\mathcal{V}}(u, v) = |v|$ components. This decomposition has two key properties. First, the gradients are orthogonal: $\nabla f_{\mathcal{U}}(u, v)$ is tangent to $\mathcal{M}$, while $\nabla f_{\mathcal{V}}(u, v)$ is normal $\mathcal{M}$ when $v \neq 0$. Second, the gradients provide a description of the minimal norm Goldstein subgradient $w_\sigma \in \partial_\sigma f(u, v)$ at a point $x = (u, v)$ near $\bar{x}$. The description depends on whether (x, σ) are in one of two regimes, which we call the *normal* and *tangent* regimes, respectively:

$$w_\sigma \approx \begin{cases} \nabla f_{\mathcal{V}}(u, v) & \text{if } (x, \sigma) \text{ are in the normal regime;} \\ \nabla f_{\mathcal{U}}(u, v) & \text{if } (x, \sigma) \text{ are in the tangent regime.} \end{cases} \tag{1.7}$$

A precise description of these regimes is not relevant at the moment. However, we mention that given any x sufficiently near $\bar{x}$, there exists some $\sigma > 0$ such that (x, σ) is in either the normal or tangent regime. Moreover, we can locate this σ via a line search. Let us now describe how this decomposition enables rapid estimation of w_σ.

In this work, we estimate w_σ with two separate `MinNorm` type methods, depending on whether (x, σ) is in the normal or tangent regime. In the normal regime, we use the `MinNorm` type method of [17], which we show terminates in finitely many steps due to the lower bound $\Delta_{x,\sigma} = \Omega(1)$. In the tangent regime, we introduce a new `MinNorm` type method which estimates $\nabla f_{\mathcal{U}}$ through a certain symmetry induced by u-axis. To motivate this symmetry, recall that we may always identify $\nabla f_{\mathcal{U}}$ by a reflection across the u-axis:

$$\nabla f_{\mathcal{U}}(u, v) = \frac{1}{2}\nabla f(u, v) + \frac{1}{2}\nabla f(u, -v) \qquad \text{for all } u, v \in \mathbb{R} \text{ with } v \neq 0.$$

We cannot hope for such a perfect symmetry in general problems. Instead, a central insight of this work is that a similar approximate symmetry exists in problems with typical structure. To illustrate, consider Figure 3. This figure depicts a point x in the tangent regime together with the result of a normalized gradient step:

$$x_+ := x - \sigma \frac{\nabla f(x)}{\|\nabla f(x)\|}.$$

As can be seen from the figure, x_+ is an approximate reflection of x across the u-axis, which "flips the sign" of the nonsmooth component of ∇f: $\nabla f_{\mathcal{V}}(x) = -\nabla f_{\mathcal{V}}(x_+)$. Thus, in this setting, one may "cancel out" the nonsmooth component by a simple averaging:

$$\nabla f_{\mathcal{U}}(x) \approx \frac{1}{2}\nabla f(x) + \frac{1}{2}\nabla f(x_+).$$

While seemingly crude, we will show this strategy generalizes to typical functions.

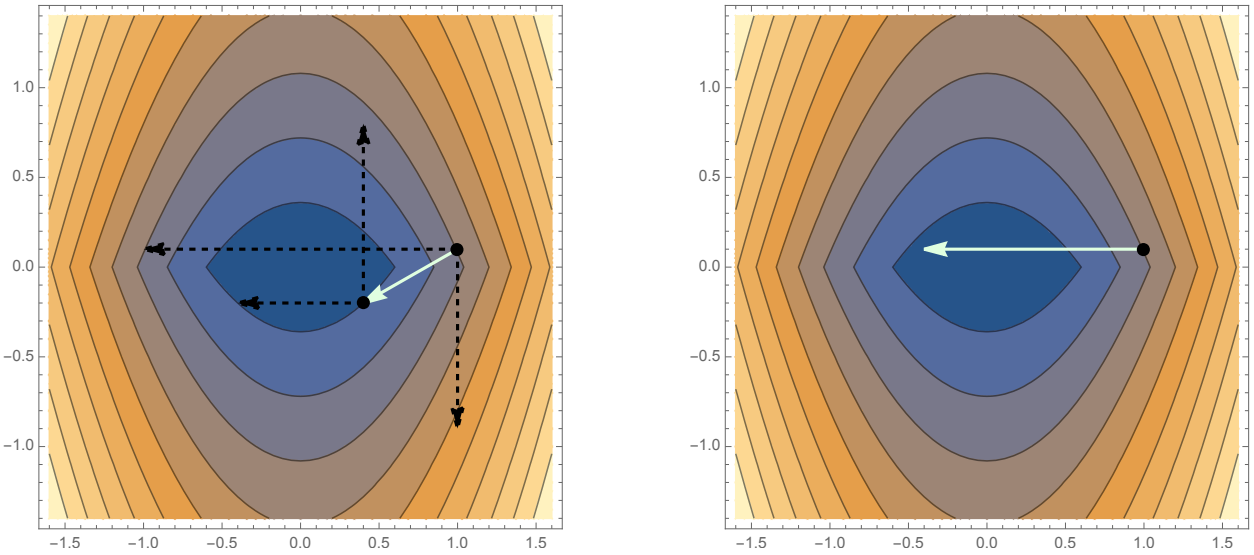


Figure 3: Contour plots for $f(u, v) = u^2 + |v|$. Left: The point $x = (1, .1)$ together with the approximate reflection $x_+ = x - .3\frac{\nabla f(x)}{\|\nabla f(x)\|}$ across the u axis. The solid light green arrow is parallel to the negative gradient direction $-\nabla f(x)$. The dashed arrows denote the orthogonal decomposition of $-\nabla f(x)$, respectively $-\nabla f(x_+)$, into the vectors $-\nabla f_{\mathcal{U}}(x)$ and $-\nabla f_{\mathcal{V}}(x)$, respectively $-\nabla f_{\mathcal{U}}(x_+)$ and $-\nabla f_{\mathcal{V}}(x_+)$. From the plot, we see $\nabla f_{\mathcal{V}}(x) = -\nabla f_{\mathcal{V}}(x_+)$. Right: The point x with estimate $-\frac{1}{2}(\nabla f(x) + \nabla f(x_+))$ of the vector $-\nabla f_{\mathcal{U}}(x)$.

1.3.2 Two `MinNorm` methods: `NDescent` and `TDescent`

To generalize the strategy of the previous section, we will prove that the minimal norm Goldstein subgradients of typical problems similarly split into tangent and normal components just as in (1.7). Then, we introduce two `MinNorm` type methods for "normal" and "tangent" steps. For normal steps, we use a small modification of the `MinNorm` type method of [17]. We call this method *Normal Descent* (`NDescent`) and describe it in Algorithm 1. As in the

simple example above, we will show that `NDescent` must terminate with sufficient descent in finitely many steps, provided σ is lies within an appropriate range.

Algorithm 1 `NDescent`(x, g, σ, T)

Set $g_0 = g$ and $t = 0$.
while $T - 1 \geq t$, $\|g_t\| > 0$, and $\frac{\sigma}{8}\|g_t\|_2 \geq f(x) - f\left(x - \sigma\frac{g_t}{\|g_t\|}\right)$ **do**
    Choose any r satisfying $0 < r < \sigma\|g_t\|$.
    Sample ζ_t uniformly from $\mathbb{B}_r(g_t)$.
    Choose y_t uniformly at random in the segment $\left[x, x - \sigma\frac{\zeta_t}{\|\zeta_t\|}\right]$.
    Choose $\hat{g}_t \in \partial f(y_t)$.
    $g_{t+1} = \operatorname{argmin}_{z \in [g_t, \hat{g}_t]} \|z\|_2$.
    $t = t + 1$.
end while
return g_t.

On the other hand, for tangent steps, we develop a new `MinNorm` type method, which likewise relies on an approximate reflection property. We call this method *Tangent Descent* (`TDescent`) and present it in Algorithm 2. Given an input point x, stepsize $\sigma > 0$, and initial subgradient $g_0 \in \partial f(x)$, `TDescent` repeats the following steps

$$\text{Choose: } \hat{g}_k \in \partial f\left(x - \sigma\frac{g_k}{\|g_k\|}\right);$$
$$\text{Update: } g_{k+1} = \operatorname*{argmin}_{g \in [g_k, \hat{g}_k]} \|g\|,$$

until it achieves descent $f(x - \sigma\frac{g_k}{\|g_k\|}) \leq f(x) - \frac{\sigma}{8}\|g_k\|$ or runs over budget. Here the interval $[g_k, \hat{g}_k]$ denotes the line-segment with endpoints g_k and $\hat{g}_k$. The motivation for this method is that for typical problems the step $x - \sigma\frac{g_k}{\|g_k\|}$ is locally an approximate reflection across $\mathcal{M}$ that "flips" the normal component of the Goldstein gradient. Indeed, let $y := P_{\mathcal{M}}(x)$ denote the projection of x onto $\mathcal{M}$ and let $N := N_{\mathcal{M}}(y)$ denote the normal space to $\mathcal{M}$ at y. Then we will prove that for all k, we have

$$\langle P_N g_k, \hat{g}_k \rangle \leq -C\|P_N g_k\| + O(\|y - \bar{x}\|^2),$$

for some $C > 0$, provided σ lies within an appropriate range. This inequality ensures that each step of the `TDescent` geometrically decreases the "normal component" of g_k, until we arrive at an approximate minimal norm Goldstein subgradient.

Algorithm 2 `TDescent`(x, g, σ, T)

Set $g_0 = g$ and $t = 0$.
while $T - 1 \geq t$, $\|g_t\| > 0$, and $\frac{\sigma}{8}\|g_t\|_2 \geq f(x) - f\left(x - \sigma\frac{g_t}{\|g_t\|}\right)$ **do**
    Choose $\hat{g}_t \in \partial f(x - \sigma\frac{g_t}{\|g_t\|})$.
    $g_{t+1} = \operatorname{argmin}_{z \in [g_t, \hat{g}_t]} \|z\|$.
    $t = t + 1$.
end while
return g_t.

1.4 The NTDescent algorithm

We call the main algorithm of this work *Normal Tangent Descent* (NTDescent) and present it in Algorithm 4. At a high-level the method is an approximate implementation of Goldstein's conceptual subgradient method as in (1.6), using NDescent and TDescent as MinNorm type methods. As input it takes three parameters: an initial point x; a sequence of grid-sizes $\{G_k\}$ for the line search on σ; and a sequence of budgets $\{T_k\}$ for the MinNorm type methods NDescent and TDescent. Later we will show that the user may simply set $T_k = G_k = k + 1$ for all $k \geq 0$.

Algorithm 3 linesearch(x, g, s, G, T)

1: **Set** $v_0 = g$.
2: **for** $i = 0, \ldots, G-1$ **do**
3: $\quad \sigma_i = 2^{-(G-i)}$.
4: $\quad u_i = \texttt{TDescent}(x_k, \sigma_i, T, v_i)$.
5: $\quad v_{i+1} = \texttt{NDescent}(x_k, \sigma_i, T, u_i)$.
6: **end for**
7: $\tilde{x} := \operatorname{argmin}\{f(x') \colon x' \in \{x\} \cup \{x - \sigma_i \frac{v_{i+1}}{\|v_{i+1}\|} \colon \sigma_i \leq \frac{\|v_{i+1}\|}{s}, i = 0, \ldots, G-1\}\}$.
8: **return** $\tilde{x}$.

Algorithm 4 NTDescent$(x, g, \{G_k\}, \{T_k\})$

Require: $g \neq 0$
1: **Set** $x_0 = x$ and $g_0 = g$.
2: **for** $k = 0, 1, \ldots$ **do**
3: $\quad x_{k+1} = \texttt{linesearch}(x_k, g_k, \|g_0\|, G_k, T_k)$.
4: $\quad$ Choose $g_{k+1} \in \partial f(x_{k+1})$.
5: **end for**

The workhorse of NTDescent is the line search procedure in Algorithm 3 (linesearch). Let us briefly comment on the structure of this method. Lines 2 through 6 of Algorithm 3 implement a line search on σ. Line 7 chooses the Goldstein subgradient that provides the most descent, while enforcing the trust-region constraint $\sigma_i \leq \frac{\|v_{i+1}\|}{s}$. Line 7 also ensures the NTDescent is a descent method. Within the line search procedure, we evaluate TDescent and NDescent a total of G times each. Not all of the calls to TDescent and NDescent will succeed with descent within the allotted budget T, but we will show that for typical problems, at least one will generate sufficient descent provided x_k is close enough to a local minimizer and T is sufficiently large. The line search allows the possibility that σ is as large as $1/2$, which might force x_{k+1} to leave the region surrounding the minimizer $\bar{x}$. This concern is what motivates the somewhat unusual structure of the line search method wherein the MinNorm-type methods are nested. Indeed, on the one hand the nesting ensures the norms of the Goldstein subgradients $\|v_{i+1}\|$ are decaying as σ_i increases. On the other hand, the trust region constraint ensures that σ_i is not chosen too large. Finally, we mention that the right-hand-side of the trust region constraint is divided by $s = \|g_0\|$ to ensure that the algorithm is invariant under rescaling of f.

1.5 Main convergence guarantees for NTDescent

The main contribution of this work is a local, nearly linear convergence rate for NTDescent. The local rate holds under a key structural assumption – Assumption A – which formalizes the concept of typical structure and mirrors the structure of the simple function considered in Section 1.3.1. While we formally describe Assumption A in Section 3, for now, we mention that it holds for max-of-smooth and properly C^p decomposable functions, provided the local minimizer $\bar{x}$ is in fact a strong local minimizer that satisfies a strict complementarity condition; this class that includes the max-of-smooth setting considered in [24]. Assumption A also holds for generic linear tilts of semialgebraic functions: if f is semialgebraic, then for a full Lebesgue measure set of $w \in \mathbb{R}^d$, Assumption A holds at every local minimizer $\bar{x}$ of the tilted function $f_w \colon x \mapsto f(x) + w^\top x$. We now present the theorem.

Theorem 1.1 (Main convergence theorem)**.** *Let $f \colon \mathbb{R}^d \to \mathbb{R}$ be satisfy Assumption A at a local minimizer $\bar{x} \in \mathbb{R}^d$. Fix budget $\{T_k\}$ and gridsize $\{G_k\}$ sequences satisfying*

$$\min\{T_k, G_k\} \geq k+1 \qquad \text{for all } k \geq 0.$$

Choose an initial point $x_0 \in \mathbb{R}^d$ and subgradient $g_0 \in \partial f(x_0)$ such that $g_0 \neq 0$. Consider iterates $\{x_k\}$ generated by NTDescent$(x_0, g_0, \{T_k\}, \{G_k\})$*. For any $q, k_0, C > 0$, let $E_{k_0,q,C}$ be the event:*

$$f(x_k) - f(\bar{x}) \leq \max\{(f(x_{k_0}) - f(\bar{x}))q^{k-k_0}, Cq^k\} \text{ for all } k \geq k_0.$$

Then there exists $q \in (0,1)$, $C, C' > 0$, and a neighborhood U of $\bar{x}$ depending solely on f such that for any failure probability $p \in (0,1)$ and all $k_0 \geq C' \max\{\log(1/p), 1\}$, we have

$$P(E_{k_0,q,C} \mid x_{k_0} \in U) \geq 1 - p.$$

Moreover, if f is convex, we have

$$P(E_{k_0,q,C}) \geq 1 - p.$$

The theorem, which is justified in Theorems 6.3 and 6.5, bounds the function gap and distance by a quantity that geometrically decays in k. Let us examine the local complexity. Recall that each outer iteration of NTDescent requires at most $2T_kG_k$ first-order oracle evaluations. Thus, if $T_k = G_k = k+1$ for all $k \geq 0$, the total number of oracle evaluations of K steps of NTDescent is at most $O(K^3)$. In other words, the local complexity of achieving an ε optimal solution is $O(\log^3(1/\varepsilon))$ for all sufficiently small $\varepsilon > 0$. Therefore the theorem establishes a local nearly linear rate of convergence for NTDescent.

1.6 Outline

The outline of this paper is as follows. In Section 1.7 we present notation and basic constructions. This section describes a key structure – the active manifold – and cannot be skipped. In Section 2, we present the sublinear convergence guarantees, which will be useful in the convex setting. This section also introduces key properties of the NDescent method, which will be used later in the work. In Section 3, we introduce our main structural assumption – Assumption A – and show that it is satisfied for the generic semialgebraic and decomposable

problem classes. In Section 4, we show that Assumption A implies the GKL inequality. In Section 5 we show that the `TDescent` and `NDescent` methods terminate rapidly under appropriate conditions. In Section 6, we use the GKL inequality and Assumption A to prove that `NTDescent` locally nearly linearly converges. Finally, in Section 7 we provide a brief numerical illustration.

1.7 Notation and basic constructions

We use standard convex analysis notation as set out in the monograph [39]. Throughout, $\mathbb{R}^d$ denotes a d-dimensional Euclidean space with the inner product $\langle \cdot, \cdot \rangle$ and the induced norm $\|x\| = \sqrt{\langle x, x \rangle}$. We denote the open ball of radius $\varepsilon > 0$ around a point $x \in \mathbb{R}^d$ by the symbol $B_\varepsilon(x)$. We use the symbol $\overline{B}$ to denote the closed unit ball at the origin. For any set $\mathcal{X} \subseteq \mathbb{R}^d$, the *distance function* and the *projection map* are defined by

$$\operatorname{dist}(x, \mathcal{X}) := \inf_{y \in \mathcal{X}} \|y - x\| \qquad \text{and} \qquad P_{\mathcal{X}}(x) := \operatorname*{argmin}_{y \in \mathcal{X}} \|y - x\|,$$

respectively. For any set $\mathcal{X} \subseteq \mathbb{R}^d$, all $\bar{x} \in \mathcal{X}$, all $x \in \mathbb{R}^d$, and all $y \in P_{\mathcal{X}}(x)$, we have

$$\|y - \bar{x}\| \leq 2\|x - \bar{x}\|.$$

We call a function $h \colon \mathbb{R}^d \to \mathbb{R}$ *sublinear* if its epigraph is a closed convex cone, and in that case we define

$$\operatorname{Lin}(h) := \{x \in \mathbb{R}^d \colon h(x) = -h(-x)\}$$

to be its *lineality space.* Given a mapping $F \colon \mathbb{R}^d \to \mathbb{R}^m$ and a point $\bar{x} \in \mathbb{R}^d$, we define

$$\operatorname{lip}_F(\bar{x}) := \limsup_{\substack{x, x' \to \bar{x} \\ x \neq x'}} \frac{\|F(x) - F(x')\|}{\|x - x'\|}.$$

Given a mapping $F \colon \mathbb{R}^d \to \mathbb{R}^{m \times n}$ into the space of $m \times n$ matrices and a point $x \in \mathbb{R}^d$ then we define

$$\operatorname{lip}_F^{\mathrm{op}}(\bar{x}) := \limsup_{\substack{x, x' \to \bar{x} \\ x \neq x'}} \frac{\|F(x) - F(x')\|_{\mathrm{op}}}{\|x - x'\|},$$

where $\| \cdot \|_{\mathrm{op}}$ denotes the operator norm defined on $\mathbb{R}^{m \times n}$.

Semialgebraicity. We call a set $\mathcal{X} \subseteq \mathbb{R}^d$ *semialgebraic* if it is the union of finitely many sets defined by finitely many polynomial inequalities. Likewise, we call a function $f \colon \mathbb{R}^d \to \mathbb{R}$ semialgebraic if its graph $\operatorname{gph}(f) = \{(x, f(x)) \colon x \in \mathbb{R}^d\}$ is semialgebraic.

Subdifferentials. Consider a locally Lipschitz function $f \colon \mathbb{R}^d \to \mathbb{R}^d$ and a point $x \in \mathbb{R}^d$. The Clarke subdifferential is the convex hull of limits of gradients evaluated at nearby points of differentiability:

$$\partial f(x) = \operatorname{conv} \left\{ \lim_{i \to \infty} \nabla f(x_i) \colon x_i \xrightarrow{\Omega} x \right\},$$

where $\Omega \subseteq \mathbb{R}^d$ is the set of points at which f is differentiable (recall Radamacher's theorem). If f is L-Lipschitz on a neighborhood U, then for all $x \in U$ and $v \in \partial f(x)$, we have $\|v\| \leq L$. A point $\bar{x}$ satisfying $0 \in \partial f(x)$ is said to be critical for f. The Goldstein subdifferential, which appears in (1.2), will be a central object throughout. An important fact is that $\partial_\sigma f(x)$ is a closed convex set for any $x \in \mathbb{R}^d$ and $\sigma > 0$.

Manifolds. We will need a few basic results about smooth manifolds, which can be found in the references [7, 26]. A set $\mathcal{M} \subseteq \mathbb{R}^d$ is called a C^p-smooth manifold around $\bar{x}$ (with $p \geq 1$) if there exists a natural number m, an open neighborhood U of $\bar{x}$, and a C^p smooth mapping $F\colon U \to \mathbb{R}^m$ such that the Jacobian $\nabla F(x)$ is surjective and $\mathcal{M} \cap U = F^{-1}(0)$. The tangent and normal spaces to $\mathcal{M}$ at $x \in \mathcal{M}$ near $\bar{x}$ are defined to be $T_{\mathcal{M}}(x) = \ker(\nabla F(x))$ and $N_{\mathcal{M}}(x) = T_{\mathcal{M}}(x)^\perp = \operatorname{range}(\nabla F(x)^*)$, respectively. If $\mathcal{M}$ is a C^2-smooth manifold around a point $\bar{x}$, then there exists $C > 0$ such that $y - x \in T_{\mathcal{M}}(x) + C\|y - x\|^2 \overline{B}$ for all $x, y \in \mathcal{M}$ near $\bar{x}$. We also have that $x - P_{\mathcal{M}}(x) \in N_{\mathcal{M}}(P_{\mathcal{M}}(x))$ for all x near $\bar{x}$. Moreover, the projection mapping $P_{\mathcal{M}}\colon \mathbb{R}^d \to \mathbb{R}$ is C^{p-1} smooth on a neighborhood of $\bar{x}$ and satisfies $\nabla P_{\mathcal{M}}(x) = P_{T_{\mathcal{M}}(x)}$ for all $x \in \mathcal{M}$ near $\bar{x}$.

Covariant gradients and smooth extension Let $\mathcal{M} \subset \mathbb{R}^d$ be a C^p-manifold around a point x for some $p \geq 1$. Then a function $f\colon \mathcal{M} \to \mathbb{R}$ is called C^q-smooth (with $q \geq 1$) around a point $x \in \mathcal{M}$ if there exists a C^q function $\hat{f}\colon U \to \mathbb{R}$ defined on an open neighborhood U of x and that agrees with f on $U \cap \mathcal{M}$. In that case, the projection of $\nabla \hat{f}(x)$ onto $T_{\mathcal{M}}(x)$ is independent of the choice of $\hat{f}$. We call this projection the *covariant gradient of f at x* and denote it by

$$\nabla_{\mathcal{M}} f(x) := P_{T_{\mathcal{M}}(x)}(\nabla \hat{f}(x)).$$

For example, the smooth extension

$$f_{\mathcal{M}} := f \circ P_{\mathcal{M}}$$

of f is $C^{\min\{p-1,q\}}$ smooth on a neighborhood of x and agrees with f along $\mathcal{M}$. Thus, we will use the identification: $\nabla_{\mathcal{M}} f(x) := \nabla f_{\mathcal{M}}(x)$.

Active Manifolds. In this work, we will assume the local minimizers of interest lie on an *active manifold.* Informally, an active manifold is a smooth manifold along which the function varies smoothly and off of which the function varies sharply. We adopt the formal model of activity explicitly used in [20]. Related models exist, e.g., identifiable surfaces [44], partly smooth manifolds [30], $\mathcal{VU}$-structures [28, 34], and $g \circ F$ decomposable functions [41].

Definition 1.2 (Active manifold)**.** Consider a function $f\colon \mathbb{R}^d \to \mathbb{R}$ and fix a set $\mathcal{M} \subseteq \mathbb{R}^d$ containing a point $\bar{x}$ satisfying $0 \in \partial f(\bar{x})$. Then $\mathcal{M}$ is called an *active C^p-manifold around* $\bar{x}$ if there exists a neighborhood U of $\bar{x}$ such that the following are true:

- **(smoothness)** The set $\mathcal{M}$ is a C^p-smooth manifold near $\bar{x}$ and the restriction of f to $\mathcal{M}$ is C^p-smooth near $\bar{x}$.
- **(sharpness)** The lower bound holds:

$$\inf\{\|v\| : v \in \partial f(x),\ x \in U \setminus \mathcal{M}\} > 0.$$

We now turn to sublinear convergence guarantees.

2 Global sublinear convergence of `NTDescent`

The main goal of this work is to show that `NTDescent` locally converges nearly linearly for "typical" nonsmooth optimization problems. A natural question is whether `NTDescent` also possess global nonasymptotic convergence guarantees. In this section, we prove two such guarantees: First, for arbitrary Lipschitz functions, we analyze the rate at which $\operatorname{dist}(0, \partial_{\sigma_i} f(x_k))$ tends to zero. Second, for convex Lipschitz functions, we analyze the rate at which $f(x_k)$ tends to $\inf f$. Before stating the result, we recall three key Lemmas, which underlie the proof. The first lemma shows that the vectors u_i and v_i generated by `linesearch` are Goldstein subgradients of decreasing norm.

Lemma 2.1 (Properties of `linesearch`)**.** *Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a locally Lipschitz function. Fix $x \in \mathbb{R}^d$, subgradient $g \in \partial f(x)$, budget T, and gridsize G. Let u_i and v_i be generated by* `linesearch`(x, g, G, T)*. Then*

$$u_i, v_{i+1} \in \partial_{\sigma_i} f(x) \qquad \text{and} \qquad \|v_{i+1}\| \le \|u_i\| \le \|v_i\| \tag{2.1}$$

for all $i = 0, \dots, G-1$.

Proof. The proof follows by induction. We prove the base case only, since the induction is straightforward. First note that the inclusion $v_0 \in \partial f(x)$ implies that $u_0 \in \partial_{\sigma_0} f(x)$, since `TDescent` constructs u_0 as a convex combinations of subgradients evaluated in the ball $\overline{B}_{\sigma_0}(\bar{x})$. Likewise, due to the argmin operation on line 4 of Algorithm 2, the subgradients generated by `TDescent` are decreasing in norm. Consequently, we have $\|u_0\| \le \|v_0\|$. A similar argument shows that $v_1 \in \partial_{\sigma_0} f(x)$ and $\|v_1\| \le \|u_0\|$. This completes the proof. □

The next lemma shows that when f is convex, the minimal norm Goldstein subgradient may be used to bound the function values. We place the proof in Appendix A, since it follows from a standard argument.

Lemma 2.2 (Subgradient inequality)**.** *Suppose that $f\colon \mathbb{R}^d \to \mathbb{R}$ is a continuous convex function. Let $x, y \in \mathbb{R}^d$. Let L denote a Lipschitz constant for f on the ball $B_{2\sigma}(x)$. Then*

$$f(x) - f(y) \le \|x - y\| \operatorname{dist}(0, \partial_\sigma f(x)) + 2\sigma L.$$

The final lemma provides conditions under which `NDescent` terminates with descent with high probability. The result is closely related to [13, Corollary 2.6].

Lemma 2.3 (`NDescent` loop terminates with contraction)**.** *Let f be a locally Lipschitz function. Fix initial point $x \in \mathbb{R}^d$, radius $\sigma > 0$, subgradient $g \in \partial_\sigma f(x)$, and failure probability $p \in (0, 1)$. Furthermore, let L be a Lipschitz constant of f on the ball $B_{2\sigma}(x)$. Suppose that $\sigma \le \frac{\operatorname{dist}(0, \partial_\sigma f(x))}{\sqrt{128} L}$ and that T satisfies:*

$$T \ge \left\lceil \frac{64 L^2}{\operatorname{dist}^2(0, \partial_\sigma f(x))} \right\rceil \lceil 2 \log(1/p) \rceil,$$

Define $g_+ := \texttt{NDescent}(x, g, \sigma, T)$. *Then* $\|g_+\| \neq 0$ *and the point* $x_+ := x - \sigma \frac{g_+}{\|g_+\|}$ *satisfies*

$$f(x_+) \leq f(x) - \frac{\sigma \operatorname{dist}(0, \partial_\sigma f(x))}{8} \qquad \text{with probability at least } 1 - p.$$

Proof. First note that $g_+ \in \partial_\sigma f(x)$, so $\|g_+\| \geq \operatorname{dist}(0, \partial_\sigma f(x)) > 0$. Now, recall that `NDescent` is precisely [13, Algorithm 1] with a more restrictive bound on the perturbation radius r. Indeed, in [13, Algorithm 1], r must satisfy

$$r < \|g_t\| \sqrt{1 - \left(1 - \frac{\|g_t\|^2}{128 L^2}\right)^2}$$

for all $t \geq 0$. We now show that the constraint $r \leq \sigma \|g_t\|$ implies the above bound. To that end, define the univariate function $h\colon a \mapsto \sqrt{1 - (1 - \frac{a^2}{128 L^2})^2}$. Then h is increasing in a for $a \leq L$. Moreover, for $a \in [0, L]$, we have $h(a) \geq \frac{a}{\sqrt{128} L}$. Consequently, since

$$\operatorname{dist}(0, \partial_\sigma f(x)) \leq \|g_t\| \leq L$$

for all $t \leq T$, we have

$$r < \sigma \|g_t\| \leq \frac{\operatorname{dist}(0, \partial_\sigma f(x)) \|g_t\|}{\sqrt{128} L} \leq h(\operatorname{dist}(0, \partial_\sigma f(x))) \|g_t\| \leq h(\|g_t\|) \|g_t\|.$$

Thus the proof is a direct application of [13, Corollary 2.6]. □

Given the above three lemmas, we are now ready to state and prove our main sublinear convergence guarantee.

Theorem 2.4 (Sublinear convergence)**.** *Let* $f\colon \mathbb{R}^d \to \mathbb{R}$ *be a locally Lipschitz function. Fix initial point* $x_0 \in \mathbb{R}^d$ *and subgradient* $g_0 \in \partial f(x)$. *Assume that* $g_0 \neq 0$. *Let* L *be any Lipschitz constant of* f *over the widened sublevel set*

$$S := \{x + u\colon f(x) \leq f(x_0) \text{ and } u \in \overline{B}(x)\}.$$

Fix a budget sequence $\{T_k\}$, *gridsize sequence* $\{G_k\}$, *and failure probability* $p \in (0, 1)$. *Let* $\{x_k\}$ *be generated by* $\texttt{NTDescent}(x, g, \{T_k\}, \{G_k\})$. *Then for all* $K > 0$, *the following holds with probability at least* $1 - p$: *Define* $G := \min_{K \leq k \leq 2K-1} G_k$ *and* $T := \min_{K \leq k \leq 2K-1} T_k$. *Then for all* $i \leq G$, *we have*

$$\min_{K \leq k \leq 2K-1} \operatorname{dist}(0, \partial_{\sigma_i} f(x_k)) \leq \max\left\{ \frac{8(f(x_K) - \inf f)}{\sigma_i K}, \frac{16 L \sqrt{2 \log(KG/p)}}{\sqrt{T}}, \sqrt{128} L \sigma_i \right\}.$$

Now suppose that f *is convex and denote* $D := \max_{K \leq k \leq 2K-1} \operatorname{dist}(x_k, \operatorname{argmin} f)$. *Then*

$$f(x_{2K-1}) - f^* \leq \min_{i \leq G} \left\{ D \max\left\{ \frac{8(f(x_K) - \inf f)}{\sigma_i K}, \frac{16 L \sqrt{2 \log(KG/p)}}{\sqrt{T}}, \sqrt{128} L \sigma_i \right\} + 2 L \sigma_i \right\}.$$

Proof. Fix $i \le G$ and define

$$\epsilon_i := \max\left\{\frac{16L\sqrt{2\log(KG/p)}}{\sqrt{T}}, \sqrt{128}L\sigma_i\right\}.$$

For every $K \le k \le 2K-1$, define

$$x_{k,i} := x_k - \sigma_i \frac{v_{i+1}}{\|v_{i+1}\|} \qquad \text{where } v_{i+1} := \texttt{NDescent}(x_k, u_i, \sigma_i, T_k),$$

where u_i and v_i appear in the definition of $\texttt{linesearch}(x_k, g_k, G_k, T_k)$; see Algorithm 3. Note that $v_{i+1} \in \partial_{\sigma_i} f(x_k)$ by Lemma 2.1. Thus, in the event $\{\mathrm{dist}(0, \partial_{\sigma_i} f(x_k)) \ge \epsilon_i\}$, we have

1. $x_{k,i}$ is well-defined since $v_{i+1} \neq 0$;
2. the trust region constraint $\sigma_i \le \frac{\|v_{i+1}\|}{s}$ is satisfied, since

$$\frac{\|v_{i+1}\|}{s} \ge \frac{\mathrm{dist}(0, \partial_{\sigma_i} f(x_k))}{s} \ge \frac{\sqrt{128}L\sigma_i}{s} \ge \sigma_i,$$

   where the final inequality follows from the bound $s = \|g_0\| \le L$, a consequence of the inclusion $B(x_0) \subseteq S$ and the Lipschitz continuity of f on S.

Finally, for every $K \le k \le 2K-1$, define

$$A_{k,i} := \left\{f(x_{k,i}) - f(x_k) \ge -\frac{\sigma_i \mathrm{dist}(0, \partial_{\sigma_i} f(x_k))}{8}\right\} \cap \{\mathrm{dist}(0, \partial_{\sigma_i} f(x_k)) \ge \epsilon_i\}.$$

Now we apply Lemma 2.3.

To that end, observe that since $f(x_k)$ is nonincreasing and $\sigma_i \le 1/2$, every iterate x_k satisfies $B_{2\sigma_i}(x_k) \subseteq S$. Consequently, L is a Lipschitz constant of f on $B_{2\sigma_i}(x_k)$. Therefore, by Lemma 2.3, for every $K \le k \le 2K-1$, we have

$$P(A_{k,i}) \le P(A_{k,i} \mid \mathrm{dist}(0, \partial_{\sigma_i} f(x_k)) \ge \epsilon_{k,i}) \le \frac{p}{GK}. \tag{2.2}$$

Thus, by a union bound, with probability at least $1 - \frac{p}{G}$, at least one of the following must hold at every index $K \le k \le 2K-1$:

$$f(x_{k,i}) - f(x_k) \le -\frac{\sigma_i \mathrm{dist}(0, \partial_{\sigma_i} f(x_k))}{8} \qquad \text{or} \qquad \mathrm{dist}(0, \partial_{\sigma_i} f(x_k)) \le \epsilon_i.$$

If $\mathrm{dist}(0, \partial_{\sigma_i} f(x_k)) \le \epsilon_i$ for some k satisfying $K \le k \le 2K-1$, then the result follows. On the other hand, suppose that for all $K \le k \le 2K-1$, we have $\mathrm{dist}(0, \partial_{\sigma_i} f(x_k)) > \epsilon_i$; in particular, we have $\mathrm{dist}(0, \partial_{\sigma_i} f(x_k)) > \sqrt{128}L\sigma_i$. Therefore, with probability at least $1 - \frac{p}{G}$, we must have

$$f(x_{k+1}) \le f(x_{k,i}) \le f(x_k) - \frac{\sigma_i \mathrm{dist}(0, \partial_{\sigma_i} f(x_k))}{8}, \qquad \text{for all } K \le k \le 2K-1.$$

where the first inequality follows since the trust region constraint is satisfied for $x_{k,i}$. Iterating this inequality, we have with probability at least $1 - \frac{p}{G}$, the bound

$$\min_{K \le k \le 2K-1} \operatorname{dist}(0, \partial_{\sigma_i} f(x_k)) \le \frac{1}{K} \sum_{k=K}^{2K-1} \operatorname{dist}(0, \partial_{\sigma_i} f(x_k)) \le \frac{8(f(x_K) - f(x_{2K}))}{\sigma_i K}.$$

This proves the result for i. Taking a union bound over i then yields the bound for minimal norm Goldstein subgradient for all $i \le G$.

To prove the function value bound, fix an $i \le G$ and let k_i be the index which attains the minimum. Then

$$f(x_{2K-1}) - \inf f \le f(x_{k_i}) - \inf f \le \operatorname{dist}(x_{k_i}, \mathcal{X}_*) \min_{K \le k \le 2K-1} \operatorname{dist}(0, \partial_{\sigma_i} f(x_k)) + 2\sigma_i L,$$

where the first inequality follows since $f(x_k)$ is nonincreasing and the second inequality follows from Lemma 2.2. The proof then follows immediately. □

The theorem provides bounds on the minimal norm Goldstein subgradient within any window of indices $K \le k \le 2K - 1$. Let us briefly investigate the setting $T_k = k + 1$ for all $k \ge 0$. In this case, the theorem implies that with probability at least $1 - p$, we have

$$\min_{K \le k \le 2K-1} \operatorname{dist}(0, \partial_{\sigma_i} f(x_k)) \le \max \left\{ \frac{8(f(x_K) - \inf f)}{\sigma_i K}, \frac{32L\sqrt{2\log(GK/p)}}{\sqrt{2K}}, \sqrt{128} L \sigma_i \right\}$$

for all $i \le G$. Let us now suppose G is large enough that there exists $i \le G$ satisfying $(1/2)K^{-1/2} \le \sigma_i \le K^{-1/2}$, e.g., we may assume $G_k = \Omega(\log(k^{1/2}))$ for all $k > 0$. Then, we find that at most $O(KG)$ first-order oracle evaluations are needed to choose find a point with x_k satisfying

$$\operatorname{dist}(0, \partial_{K^{-1/2}} f(x_k)) = \tilde{O}(K^{-1/2}),$$

where $\tilde{O}$ hides logarithmic terms in G, K and p. Let's consider two settings for G_k.

1. **Setting 1:** $G_k = O(\log(k^{1/2}))$**.** In this case, `NTDescent` finds a point x_k satisfying $\operatorname{dist}(0, \partial_\varepsilon f(x_k)) \le \varepsilon$ using at most $\tilde{O}(\varepsilon^{-4})$ first-order oracle evaluations.

2. **Setting 2:** $G_k = k + 1$**.** In this case, `NTDescent` finds a point x_k satisfying $\operatorname{dist}(0, \partial_\varepsilon f(x_k)) \le \varepsilon$ using at most $\tilde{O}(\varepsilon^{-6})$ first-order oracle evaluations.

The complexity of Setting 1 is smaller than the complexity of Setting 2. Nevertheless, when we establish our local rapid convergence guarantees, we will work in setting 2, which has more favorable local convergence properties. Before moving on, we note that the above guarantees likewise apply in the convex setting, namely `NTDescent` finds a point x_k with $f(x_k) - f^* \le \varepsilon$ using at most $\tilde{O}(\varepsilon^{-4})$, respectively $\tilde{O}(\varepsilon^{-6})$, first-order oracle evaluations in Setting 1, respectively Setting 2.

This concludes our sublinear convergence guarantees for `NTDescent`. In the following section, we describe the key structural assumptions needed to ensure that `NTDescent` locally rapidly converges.

3 Main assumption, examples, and consequences

In this section, we introduce our key structural assumption – Assumption A. In Section 3.1 we show that Assumption A holds for generic semialgebraic functions and certain properly C^p decomposable functions. Then in Section 3.2, we extract several key consequences of Assumption A. These consequences will be instrumental in proving the GKL inequality and rapid convergence of `NTDescent`. We now turn to the assumption.

Assumption A. Function $f\colon \mathbb{R}^d \to \mathbb{R}$ is locally Lipschitz with local minimizer $\bar{x} \in \mathbb{R}^d$.

(A1) **(Quadratic Growth)** There exists $\gamma > 0$ such that

$$f(x) - f(\bar{x}) \geq \frac{\gamma}{2}\|x - \bar{x}\|^2 \qquad \text{for all } x \text{ near } \bar{x}.$$

(A2) **(Active Manifold)** Function f admits a C^4-smooth active manifold $\mathcal{M}$ around $\bar{x}$.

(A3) **(Strong-(a) regularity)** There exists $C_{(a)} > 0$ such that

$$\|P_{T_{\mathcal{M}}(y)}(v - \nabla_{\mathcal{M}} f(y))\| \leq C_{(a)}\|x - y\| \qquad \text{for all } x \in \mathbb{R}^d,\ v \in \partial f(x), \text{ and } y \in \mathcal{M} \text{ near } \bar{x}.$$

(A4) **($(b_{\leq})$-regularity)** The following inequality holds

$$f(y) \geq f(x) + \langle v, y - x\rangle + o(\|y - x\|) \qquad \text{as } y \xrightarrow{\mathcal{M}} \bar{x} \text{ and } x \to \bar{x} \text{ with } v \in \partial f(x),$$

where $o(\cdot)$ is any univariate function satisfying $\lim_{t\to 0} o(t)/t = 0$.

Some comments are in order. Assumption (A1) is a classical regularity condition that ensures local linear convergence of gradient methods for smooth convex optimization. Assumptions (A2), (A3), and (A4) describe the interaction of f and a distinguished smooth manifold $\mathcal{M}$. Assumption (A2) requires $\mathcal{M}$ to be an active manifold for f around $\bar{x}$ in the sense of Definition 1.2. In particular, along the manifold $\mathcal{M}$, the function f is C^4 smooth with covariant gradient $\nabla_{\mathcal{M}} f$; see Section 1.7 for a definition. Assumption (A3) shows that in tangent directions the covariant gradient along the manifold approximates the subgradients of f up to a linear error. This property recently appeared in [4,15], where it was used to study saddle avoidance properties of the subgradient method for nonsmooth optimization. Finally, Assumption (A4) is a restricted lower smoothness property, showing that linear models of f off the manifold are underapproximators of f on the manifold up to first-order. Note that the property is automatic if f is weakly convex, meaning the mapping $x \mapsto f(x) + \frac{\rho}{2}\|x\|^2$ is convex for some $\rho \geq 0$. The weakly convex class is broad and contains all compositions of convex functions with smooth mappings that have Lipschitz Jacobians; see the survey [12] for an introduction. We mention that the name "$(b_{\leq})$-regularity" is motivated by "uniform semismoothness" property of [15], which was called the "(b)-regularity property."

In the following section, we provide several examples of Assumption A.

3.1 Examples of Assumption A

In this section we show that the aforementioned problems satisfy Assumption A. The most important example is the class of generic semialgebraic functions. The following theorem is essentially contained in [15, 19], but we provide a proof for completeness.

Theorem 3.1 (Generic semialgebraic functions)**.** *Consider a locally Lipschitz semialgebraic function $f\colon \mathbb{R}^d \to \mathbb{R}$. Then for a full Lebesgue measure set of $w \in \mathbb{R}^d$, the tilted function $f_w\colon x \mapsto f(x) + w^\top x$ satisfies Assumption A at every local minimizer.*

Proof. The proof is a consequence of [15, Corollary 2.7.6] and [19, Theorem 4.16(ii)]. The result [19, Theorem 4.16] shows that for a full Lebesgue measure set of $w \in \mathbb{R}^d$, the following hold: every local minimizer $\bar{x}$ of f_w lies on a C^4 active manifold $\mathcal{M}$, verifying (A2); and the quadratic growth condition (A1) holds at $\bar{x}$. Next, [15, Corollary 2.7.6] shows that f_w also satisfies the strong (a) property (A3) along $\mathcal{M}$; applying [15, Theorem 2.2.4], we deduce that f_w also satisfies the $(b_{\leq})$-regularity property (A4) along $\mathcal{M}$ at $\bar{x}$. □

Turning to our second class, we introduce so-called *properly C^p decomposable* functions, originally proposed and analyzed in [40]. At a high-level, the class consists of functions that are locally the composition of a sublinear function with a smooth mapping, which together satisfy a transversality condition.

Definition 3.2 (Decomposable functions)**.** A function $f\colon \mathbb{R}^d \to \mathbb{R}$ is called *properly C^p decomposable at $\bar{x}$ as $h \circ c$* if near $\bar{x}$ it can be written as

$$f(x) = f(\bar{x}) + h(c(x))$$

for some C^p-smooth mapping $c\colon \mathbb{R}^d \to \mathbb{R}^m$ satisfying $c(\bar{x}) = 0$ and some proper, closed sublinear function $h\colon \mathbb{R}^m \to \mathbb{R}$ satisfying the transversality condition:

$$\operatorname{lin}(h) + \operatorname{range}(\nabla c(\bar{x})) = \mathbb{R}^m.$$

The following theorem theorem shows shows that decomposable functions satisfy Assumption A near local minimizers if they also satisfy a strict complementarity condition and a quadratic growth bound. The proof is a consequence of results found in works [15,20,30,40].

Theorem 3.3 (Properly decomposable functions)**.** *Consider a locally Lipschitz function $f\colon \mathbb{R}^d \to \mathbb{R}$. Let $\bar{x}$ be a local minimizer of f and suppose that f is properly C^4 decomposable at $\bar{x}$. Furthermore, suppose that*

1. **(Strict Complementarity)** *We have that $0 \in \operatorname{ri} \partial f(\bar{x})$.*
2. **(Quadratic growth)** *There exists $\gamma > 0$ such that*

$$f(x) - f(\bar{x}) \geq \frac{\gamma}{2}\|x - \bar{x}\|^2 \qquad \text{for all } x \text{ near } \bar{x}.$$

Then f satisfies Assumption A at $\bar{x}$.

Proof. To set the notation for the proof, recall that since f is properly C^4 decomposable, there exists functions h and c satisfying the conditions of Definition 3.2. The discussion in [40, p. 683-4] then shows that the set

$$\mathcal{M} := c^{-1}(\mathrm{lin}(h))$$

is a so-called C^4 *partly smooth manifold* for f around $\bar{x}$ in the sense of Lewis [30]. Thus, according to [20, Proposition 10.12], partial smoothness and strict complementarity ensure that f admits a C^4 smooth active manifold $\mathcal{M}$ around $\bar{x}$, verifying (A2). In addition, [15, Corollary 2.6.3] ensures that f satisfies the (A3) and (A4) properties along $\mathcal{M}$. □

A popular class of decomposable objectives arise from pointwise maxima of smooth functions that satisfy an affine independence property. For example, this class was considered in the work of Han and Lewis [24]. As an immediate corollary of Theorem 3.3, we show that such functions satisfy Assumption A.

Corollary 3.4 (Max-of-smooth functions)**.** *Consider a locally Lipschitz function f and a family of C^4 smooth functions $f_i \colon \mathbb{R}^d \to \mathbb{R}$ indexed by a finite set $i \in I$. Fix a local minimizer $\bar{x}$ of f and suppose the set $\{\nabla f_i(\bar{x})\}_{i\in I}$ is affinely independent. Suppose furthermore that f is locally expressible as*

$$f(x) := \max_{i\in I} f_i(x) \qquad \text{for all } x \text{ near } \bar{x}.$$

Then provided the strict complementarity and quadratic growth conditions of Theorem 3.3 hold, the function f satisfies Assumption A at $\bar{x}$.

Proof. To prove the result, note that the affine independence property is simply a restatement of the transversality condition of Definition 3.2 for the smooth mapping $x \mapsto (f_i(x))_{i\in I}$ and the sublinear function $y \mapsto \max_{i\in I} y_i$. □

We now turn our attention to the key consequences of Assumption A.

3.2 Key consequences of Assumption A

The following proposition summarizes the key consequences of Assumption A. The proof of the result is straightforward, but technical, so we place it in Appendix B.

Proposition 3.5 (Consequences of Assumption A)**.** *Suppose f satisfies Assumption A at $\bar{x}$. Then there exists $\delta_{\mathrm{A}} > 0$ such that on the ball $B_{2\delta_{\mathrm{A}}}(\bar{x})$, the projection operator $P_{\mathcal{M}}$ is C^3 with Lipschitz Jacobian and the smooth extension $f_{\mathcal{M}} := f \circ P_{\mathcal{M}}$ is C^3 with Lipschitz gradient. Moreover, following bounds hold:*

1. **(Quadratic growth)** *The quadratic growth bound* (A1) *holds throughout $\overline{B}_{2\delta_{\mathrm{A}}}(\bar{x})$.*
2. **(Smoothness of $P_{\mathcal{M}}$)** *For all $x \in B_{\delta_{\mathrm{A}}}(\bar{x})$ and $x' \in B_{2\delta_{\mathrm{A}}}(\bar{x})$, we have*

$$\|P_{\mathcal{M}}(x') - P_{\mathcal{M}}(x) - P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(x' - x)\| \leq C_{\mathcal{M}}(\mathrm{dist}^2(x, \mathcal{M}) + \|x - x'\|^2), \tag{3.1}$$

where $C_{\mathcal{M}} := 2\mathrm{lip}^{\mathrm{op}}_{\nabla P_{\mathcal{M}}}(\bar{x})$.

3. **(Bounds on $\nabla_{\mathcal{M}} f$)** *For all $x \in B_{\delta_{\mathrm{A}}}(\bar{x})$, we have*

$$\frac{\gamma}{2}\|P_{\mathcal{M}}(x) - \bar{x}\| \leq \|\nabla_{\mathcal{M}} f(P_{\mathcal{M}}(x))\| \leq \beta\|P_{\mathcal{M}}(x) - \bar{x}\|, \tag{3.2}$$

where $\beta := 2\mathrm{lip}_{\nabla f_{\mathcal{M}}}(\bar{x})$.

4. **(Consequence of strong (a))** *For all $x \in B_{\delta_{\mathrm{A}}}(\bar{x})$ and $\sigma \leq \delta_{\mathrm{A}}$, we have*

$$\|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(g - \nabla_{\mathcal{M}} f(y))\| \leq C_{(a)}(\mathrm{dist}(x, \mathcal{M}) + \sigma); \tag{3.3}$$

$$\sup_{g \in \partial_\sigma f(x)} \|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))} g\| \leq C_{(a)}(\mathrm{dist}(x, \mathcal{M}) + \sigma) + \beta\|y - \bar{x}\|; \tag{3.4}$$

$$\sup_{g, g' \in \partial_\sigma f(x)} \|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(g - g')\| \leq 2C_{(a)}(\mathrm{dist}(x, \mathcal{M}) + \sigma). \tag{3.5}$$

5. **(Aiming)** *For all $x \in B_{\delta_{\mathrm{A}}}(\bar{x})$ and all $v \in \partial f(x)$, we have*

$$\langle v, x - P_{\mathcal{M}}(x)\rangle \geq \mu\,\mathrm{dist}(x, \mathcal{M}), \tag{3.6}$$

where $\mu := \frac{1}{4}\liminf_{x' \overset{\mathcal{M}^c}{\to} \bar{x}} \mathrm{dist}(0, \partial f(x'))$.

6. **(Subgradient bound)** *For all $x \in B_{\delta_{\mathrm{A}}}(\bar{x})$ and $\sigma \leq \delta_{\mathrm{A}}$, we have*

$$\sup_{g \in \partial_\sigma f(x)} \|g\| \leq L,$$

where $L := 2\mathrm{lip}_f(\bar{x})$

7. **(Function gap)** *For all $x \in B_{\delta_{\mathrm{A}}}(\bar{x})$, we have*

$$f(x) - f(\bar{x}) \leq L\mathrm{dist}(x, \mathcal{M}) + \frac{\beta}{2}\|P_{\mathcal{M}}(x) - \bar{x}\|^2. \tag{3.7}$$

Let us briefly comment on the result. Item 2 provides a crucial smoothness property of the projection operator of $\mathcal{M}$. Item 3 shows that the gradient of the smooth extension $f_{\mathcal{M}}$ is proportional to the distance of the projection y to $\bar{x}$. Item 4 shows how the Goldstein subgradients inherit the strong (a) property (A3) of Assumption A. Indeed, Equation (3.4) shows that Goldstein subgradients are "small" in tangent directions and Equation (3.5) shows Goldstein subgradients vary in an approximate Lipschitz fashion in tangent directions. Item 5 shows that the subgradients of f off of the manifold have a constant level of correlation with $x - P_{\mathcal{M}}(x)$, i.e., the direction $-v$ "aims" towards the manifold. Note that $\mu > 0$ due to Assumption (A2). The proof of Item 5 is based on Assumptions (A2) and (A4); a similar result appears in [14, Theorem D.2]. Item 6 provides a bound on the Goldstein subgradients of f near $\bar{x}$; we will appeal to this bound many times throughout the analysis without referencing this proposition. Finally, Item 7 decomposes the function gap into a sum of two terms: the distance to the manifold and the squared distance of the projection to the solution. The proof relies on smoothness of f along the manifold. Note that the trivial upper bound $L\|x - \bar{x}\|$ for the gap can be weaker than (3.7).

This concludes our discussion of Assumption A. The following three sections establish further consequences: the GKL inequality (Section 4); rapid local convergence of `NDescent` and `TDescent` (Section 5); and rapid local convergence of `NTDescent` (Section 6). In all three sections, we use the notation and results introduced in Proposition 3.5.

4 Verifying the GKL inequality under Assumption A

In this section we establish the GKL inequality (1.5) for functions satisfying Assumption A. Throughout the section, we assume that that Assumption A is in force. We also use notation set out in Proposition 3.5.

We present the formal statement and the GKL inequality in Theorem 4.3, which appears at the end of this section. The proof is a consequence of the two lemmata. In the first lemma, we prove a constant size lower bound for $\partial_\sigma f(x)$, whenever σ is sufficiently small. The proof of this bound relies on the active manifold assumption (A2) and the aiming inequality (3.6). A consequence of the argument is that all elements of $\partial_\sigma f(x)$ are correlated with the normal direction $x - P_{\mathcal{M}}(x) \in N_{\mathcal{M}}(P_{\mathcal{M}}(x))$. Later in Proposition 5.1 we will also show that Algorithm 1 (`NDescent`) terminates rapidly when σ is in the regime, motivating the name Normal Descent. We now turn to the lemma.

Lemma 4.1 (Lower bound on Goldstein subgradients)**.** *Define*

$$D_1 := \frac{1}{2(1+3L/\mu)} \qquad \text{and} \qquad D_2 := \frac{\mu}{2}.$$

Then for all x near $\bar{x}$ and $0 < \sigma \leq D_1 \mathrm{dist}(x, \mathcal{M})$, we have

$$\mathrm{dist}(0, \partial_\sigma f(x)) \geq D_2.$$

Proof. Assume $x \in B_\delta(\bar{x})$ and $\sigma \leq \delta$ where $\delta \leq \delta_{\mathrm{A}}/4$ and δ_{A} is chosen as in Proposition 3.5. Let $x' \in \overline{B}_\sigma(x) \subseteq B_{\delta_{\mathrm{A}}}(\bar{x})$ and observe that

$$\langle v, x' - P_{\mathcal{M}}(x') \rangle \geq \mu \mathrm{dist}(x', \mathcal{M})$$

for all $v \in \partial f(x')$. Observe that we may upper bound the inner product:

$$\begin{aligned}
&\langle v, x' - P_{\mathcal{M}}(x') \rangle \\
&\leq \langle v, x - P_{\mathcal{M}}(x) \rangle + \|v\| \|x' - x\| + \|v\| \|P_{\mathcal{M}}(x) - P_{\mathcal{M}}(x')\| \\
&\leq \langle v, x - P_{\mathcal{M}}(x) \rangle + L D_1 \mathrm{dist}(x, \mathcal{M}) + L \|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(x - x')\| + L C_{\mathcal{M}}(\sigma^2 + \mathrm{dist}^2(x, \mathcal{M})) \\
&\leq \langle v, x - P_{\mathcal{M}}(x) \rangle + 2 L D_1 \mathrm{dist}(x, \mathcal{M}) + L C_{\mathcal{M}}(\sigma^2 + \mathrm{dist}^2(x, \mathcal{M})),
\end{aligned}$$

where the second inequality follows from the bound $\|x - x'\| \leq \sigma$, the bound $\|v\| \leq L$, and Item 2 of Proposition 3.5. Therefore, shrinking δ if necessary, we have

$$\langle v, x' - P_{\mathcal{M}}(x') \rangle \leq \langle v, x - P_{\mathcal{M}}(x) \rangle + 3 L D_1 \mathrm{dist}(x, \mathcal{M}).$$

Consequently we have

$$\begin{aligned}
\langle v, x - P_{\mathcal{M}}(x) \rangle &\geq \mu \mathrm{dist}(x', \mathcal{M}) - 3 L D_1 \mathrm{dist}(x, \mathcal{M}) \\
&\geq \mu \mathrm{dist}(x, \mathcal{M}) - \mu\sigma - 3 L D_1 \mathrm{dist}(x, \mathcal{M}) \\
&\geq \mu(1 - D_1(1 + 3L/\mu)) \mathrm{dist}(x, \mathcal{M}) \\
&= D_2 \mathrm{dist}(x, \mathcal{M}). \qquad (4.1)
\end{aligned}$$

Now fix $g \in \partial_\sigma f(x)$. By definition of $\partial_\sigma f(x)$, there exists a family of coefficients $\lambda_i \in [0,1]$, points $x_i \in \overline{B}_\sigma(x) \subseteq \overline{B}_\delta(\bar{x})$, and subgradients $g_i \in \partial f(x_i)$ indexed by a finite set $i \in I$ such that $\sum_{i\in I} \lambda_i = 1$ and $g = \sum_{i\in I} \lambda_i g_i$. Thus, by (4.1), we have

$$\langle g, x - P_{\mathcal{M}}(x)\rangle = \sum_{i\in I} \lambda_i \langle g_i, x - P_{\mathcal{M}}(x)\rangle \geq D_2 \mathrm{dist}(x, \mathcal{M})$$

Therefore, we have

$$\|g\| \geq \frac{\langle g, x - P_{\mathcal{M}}(x)\rangle}{\mathrm{dist}(x,\mathcal{M})} \geq D_2,$$

as desired. □

In the second lemma, we provide a lower bound for $\mathrm{dist}(0, \partial_\sigma f(x))$ on the order of $\|P_{\mathcal{M}}(x) - \bar{x}\|$, provided σ is on the order of $\|P_{\mathcal{M}}(x) - \bar{x}\|$. The proof of this bound relies on quadratic growth and strong (a)-regularity. A consequence of the argument is that the minimal norm element of $\partial_\sigma f(x)$ is close to the tangent vector $\nabla_{\mathcal{M}} f(P_{\mathcal{M}}(x)) \in T_{\mathcal{M}}(P_{\mathcal{M}}(x))$. Later in Proposition 5.5 we will also show that Algorithm 2 (`TDescent`) terminates rapidly when σ is in the regime, motivating the name Tangent Descent. We now turn to the lemma.

Lemma 4.2 (Lower bound on Goldstein subgradients)**.** *Define*

$$C_1 := \frac{\gamma}{4} \qquad \text{and} \qquad C_2 := \frac{\gamma}{8C_{(a)}}.$$

Then for all x near $\bar{x}$ satisfying

$$\max\{\mathrm{dist}(x,\mathcal{M}), \sigma\} \leq C_2 \|P_{\mathcal{M}}(x) - \bar{x}\|,$$

we have

$$\|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(g)\| \geq C_1 \|P_{\mathcal{M}}(x) - \bar{x}\| \qquad \text{for all } g \in \partial_\sigma f(x).$$

Proof. Assume $x \in B_{\delta_{\mathrm{A}}}(\bar{x})$ and $\sigma \leq \delta_{\mathrm{A}}$ where δ_{A} is chosen as in Proposition 3.5. Define $y = P_{\mathcal{M}}(x)$. Then by (3.3), for all $g \in \partial_\sigma f(x)$, we have

$$\|P_{T_{\mathcal{M}}(y)}(g - \nabla_{\mathcal{M}} f(y))\| \leq C_{(a)}(\mathrm{dist}(x,\mathcal{M}) + \sigma) \leq \frac{\gamma}{4}\|y - \bar{x}\|.$$

In addition, by (3.2), we have $\|\nabla_{\mathcal{M}} f(y)\| \geq \frac{\gamma}{2}\|y - \bar{x}\|$. Therefore, for all $g \in \partial_\sigma f(x)$, we have

$$\|P_{T_{\mathcal{M}}(y)}(g)\| \geq \|\nabla_{\mathcal{M}} f(y)\| - C_{(a)}(\mathrm{dist}(x,\mathcal{M}) + \sigma) \geq \frac{\gamma}{4}\|y - \bar{x}\|,$$

as desired. □

Given these lemmata, we are now ready to establish the GKL inequality (1.5). The following result, verifies the bound

$$\sigma \mathrm{dist}(0, \partial_\sigma f(x)) \geq \eta(f(x) - f(\bar{x})),$$

for some $\eta > 0$ provided x is sufficiently near $\bar{x}$ and σ lies within one of two regimes, described in Item 1 and Item 2 of Theorem 4.3. Item 1 and Item 2 roughly correspond to the regimes

considered in Lemma 4.1 and Lemma 4.2, respectively. Comparing with the statement of the GKL inequality in (1.5), we see that GKL inequality of Theorem 4.3 does not require knowledge of an explicit function $\sigma(x)$. Instead, we need only find some σ proportional to $D_1\text{dist}(x,\mathcal{M})$ or $C_2\|P_{\mathcal{M}}(x)-\bar{x}\|$ up to a factor of, say, 2. Later in Proposition 6.1 we show that this flexibility allows us find an appropriate σ through the `linesearch` procedure.

Theorem 4.3 (GKL Inequality)**.** *Suppose that function f satisfies Assumption A at $\bar{x}\in\mathbb{R}^d$. For any constants $a_1\in(0,D_1]$ and $a_2\in(0,C_2]$, there exists $\delta>0$ such that*

$$\sigma\text{dist}(0,\partial_\sigma f(x))\geq\min\left\{\frac{\gamma a_2}{40\max\{20La_2^2,\beta\}},\frac{\mu a_1}{20\max\{L/2,\beta/a_2^2\}}\right\}(f(x)-f(\bar{x})),$$

whenever $x\in B_\delta(\bar{x})$ and $\sigma>0$ satisfy Item 1 or Item 2:

1. *(a)* $\frac{a_1}{10}\text{dist}(x,\mathcal{M})\leq\sigma\leq a_1\text{dist}(x,\mathcal{M})$;
   (b) $a_2^2\|P_{\mathcal{M}}(x)-\bar{x}\|^2\leq\text{dist}(x,\mathcal{M})$.
2. *(a)* $\frac{a_2}{10}\|P_{\mathcal{M}}(x)-\bar{x}\|\leq\sigma\leq a_2\|P_{\mathcal{M}}(x)-\bar{x}\|$;
   (b) $\frac{\text{dist}(x,\mathcal{M})}{\sigma}\leq 10a_2\|P_{\mathcal{M}}(x)-\bar{x}\|$.

Moreover, for any $x\in B_\delta(\bar{x})\backslash\{\bar{x}\}$, there exists $\sigma>0$ such that Item 1 or Item 2 is satisfied.

Proof. Suppose that $\delta\leq\min\{\delta_{\text{A}},1/(10C_2)\}$ is small enough that the conclusions of Lemma 4.1 and Lemma 4.2 hold. We first show we that either Item 1 or Item 2 is satisfied for any $x\in B_\delta(\bar{x})$ and some $\sigma>0$. Note that if $x\in\mathcal{M}$, Item 2 is trivially satisfied for $\sigma=a_2\|P_{\mathcal{M}}(x)-\bar{x}\|$. Thus, we focus on the case where $\text{dist}(x,\mathcal{M})>0$ and Item 1 cannot be satisfied for any $\sigma>0$. In this case, we have

$$\text{dist}(x,\mathcal{M})\leq a_2^2\|P_{\mathcal{M}}(x)-\bar{x}\|^2=10a_2\sigma\|P_{\mathcal{M}}(x)-\bar{x}\|^2$$

with $\sigma=a_2\|P_{\mathcal{M}}(x)-\bar{x}\|/10$. Thus, Item 2 is satisfied.

Now we prove the GKL bound is satisfied whenever Item 1 or Item 2 holds. Let us suppose that Item 1 holds for some $x\in B_\delta(\bar{x})$ and $\sigma>0$. From (3.7), we have the bound:

$$\begin{aligned}\text{dist}(x,\mathcal{M})&\geq\frac{1}{2}\left(\text{dist}(x,\mathcal{M})+a_2^2\|P_{\mathcal{M}}(x)-\bar{x}\|^2\right)\\&\geq\frac{1}{\max\{L/2,\beta/a_2^2\}}\left(L\text{dist}(x,\mathcal{M})+\frac{\beta}{2}\|P_{\mathcal{M}}(x)-\bar{x}\|^2\right)\\&\geq\frac{1}{\max\{L/2,\beta/a_2^2\}}(f(x)-f(\bar{x})).\end{aligned}$$

Now observe that the assumptions of Lemma 4.1 are satisfied since $a_1\leq D_1$ and x and σ satisfy Item 1. Therefore, we have

$$\sigma\text{dist}(0,\partial_\sigma f(x))\geq\frac{\sigma\mu}{2}\geq\frac{\mu a_1}{20}\text{dist}(x,\mathcal{M})\geq\frac{\mu a_1}{20\max\{L/2,\beta/a_2^2\}}(f(x)-f(\bar{x})),$$

as desired.

Next, let us suppose that Item 2 holds for some $x \in B_\delta(\bar{x})$ and $\sigma > 0$. From (3.7), we have the bound:

$$\begin{aligned}\|P_{\mathcal{M}}(x) - \bar{x}\|^2 &\geq \frac{1}{2}\left(\frac{\operatorname{dist}(x,\mathcal{M})\|P_{\mathcal{M}}(x) - \bar{x}\|}{10a_2\sigma} + \|P_{\mathcal{M}}(x) - \bar{x}\|^2\right) \\ &\geq \frac{1}{2}\left(\frac{\operatorname{dist}(x,\mathcal{M})}{10a_2^2} + \|P_{\mathcal{M}}(x) - \bar{x}\|^2\right) \\ &\geq \frac{1}{\max\{20La_2^2, \beta\}}\left(L\operatorname{dist}(x,\mathcal{M}) + \frac{\beta}{2}\|P_{\mathcal{M}}(x) - \bar{x}\|^2\right) \\ &\geq \frac{1}{\max\{20La_2^2, \beta\}}(f(x) - f(\bar{x})).\end{aligned}$$

Now observe that since $a_2 \leq C_2$ and x and σ satisfy Item 2, we have

$$\sigma \leq C_2\|P_{\mathcal{M}}(x) - \bar{x}\| \leq C_2\delta \leq 1/10.$$

Consequently, we have

$$\operatorname{dist}(x,\mathcal{M}) \leq 10\sigma C_2\|P_{\mathcal{M}}(x) - \bar{x}\| \leq C_2\|P_{\mathcal{M}}(x) - \bar{x}\|.$$

Therefore, $\max\{\operatorname{dist}(x,\mathcal{M}), \sigma\} \leq C_2\|P_{\mathcal{M}}(x) - \bar{x}\|$, so the conditions of Lemma 4.2 are satisfied. Thus, we have

$$\sigma\operatorname{dist}(0, \partial_\sigma f(x)) \geq \sigma\|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(g)\| \geq \frac{\sigma\gamma}{4}\|P_{\mathcal{M}}(x) - \bar{x}\| \geq \frac{\gamma a_2}{40\max\{20La_2^2, \beta\}}(f(x) - f(\bar{x})),$$

where the last inequality follows from $\sigma \geq \frac{a_2}{10}\|P_{\mathcal{M}}(x) - \bar{x}\|$. This completes the proof. □

This concludes the proof of the GKL inequality under Assumption A. In Section 6, we will use the GKL inequality to establish rapid local convergence of `NTDescent`. Before proving that, the following section analyzes `TDescent` and `NDescent` methods.

5 Rapid termination of `NDescent` and `TDescent` under Assumption A

In this section, we analyze the `NDescent` and `TDescent` method, showing that both methods rapidly terminate with descent in appropriate regimes. Throughout the section, we assume that that Assumption A is in force. We also use the results and notation of Proposition 3.5, Lemma 4.1, and Lemma 4.2.

The main results of this section are Propositions 5.1 and 5.5, which analyze `NDescent` and `TDescent`, respectively. Proposition 5.1 shows that `NDescent` terminates with descent in a constant number of iterations within the regime considered in Item 1 of Theorem 4.3. Proposition 5.5 shows that `TDescent` either terminates with descent in $O(\log^{-1}(f(x) - f(\bar{x})))$ iterations or $f(x) - f(\bar{x})$ is already exponentially small in T within the regime considered in Item 2 of Theorem 4.3. These lemmata will be the basis of our main convergence theorem – Theorem 6.3 – appearing in Section 6.

5.1 Analysis of `NDescent`

The following proposition shows that `NDescent` locally terminates in finitely many iterations whenever σ is sufficiently small. The result is a simple consequence of Lemmas 2.3 and 4.1.

Proposition 5.1 (`NDescent` loop terminates with descent)**.** *For all x near $\bar{x}$, radii $\sigma > 0$ with $\sigma \leq D_1 \mathrm{dist}(x, \mathcal{M})$, subgradients $g \in \partial_\sigma f(x)$, failure probabilities $p \in (0,1)$ and budgets $T > 0$ satisfying*

$$T \geq \left\lceil \frac{64L^2}{D_2} \right\rceil \lceil 2\log(1/p) \rceil,$$

the point $x_+ := \texttt{NDescent}(x, g, \sigma, T)$ satisfies

$$f(x_+) \leq f(x) - \frac{\sigma \mathrm{dist}(0, \partial_\sigma f(x))}{8} \qquad \textit{with probability at least } 1 - p.$$

Proof. Choose x near enough to $\bar{x}$ that $\sigma \leq \min\left\{\frac{D_2}{L\sqrt{128}}, \delta_{\mathrm{A}}\right\}$ and the conclusion of Lemma 4.1 holds. Then we have the lower bound $\mathrm{dist}(0, \partial_\sigma f(x)) \geq D_2$. Consequently, the result follows immediately from Lemma 2.3. □

We now turn to analysis of the `TDescent` step.

5.2 Analysis of `TDescent`

The main goal of this section is to prove Proposition 5.5, which shows that `TDescent` terminates rapidly. The proof of the Proposition relies on three technical lemmata that analyze the structure of Goldstein subgradients when σ is sufficiently small and x is sufficiently near $\bar{x}$: Lemma 5.2 states that elements of Goldstein subdifferential with small normal component are descent directions. Lemma 5.3 shows that normalized subgradient steps approximately reflect points across the active manifold. Lemma 5.4 uses the approximate reflection property to show that `TDescent` geometrically decreases the normal component of the input subgradient, ensuring that we rapidly find a descent direction.

We now turn to the Lemmata. The first lemma shows that Goldstein subgradients with small normal component are descent directions.

Lemma 5.2 (Descent with small normal part)**.** *There exists $C_3, C_4 > 0$ such that for all x near $\bar{x}$, $\sigma > 0$, and $g \in \partial_\sigma f(x) \backslash \{0\}$ satisfying*

1. $\max\{\mathrm{dist}(x, \mathcal{M}), \sigma\} \leq C_3 \|P_{\mathcal{M}}(x) - \bar{x}\|$;
2. $\|P_{N_{\mathcal{M}}(P_{\mathcal{M}}(x))}(g)\| \leq C_4 \|P_{\mathcal{M}}(x) - \bar{x}\|^2$,

we have

$$f\left(x - \sigma \frac{g}{\|g\|}\right) \leq f(x) - \frac{\sigma \|g\|}{8}.$$

Proof. We define the constants

$$C_3 := \min\left\{C_2, \frac{C_1}{8C_{(a)}}\right\} \qquad \text{and} \qquad C_4 := \frac{C_1^2}{8L}.$$

We fix $x \in B_\delta(\bar{x})$ for some $\delta \le \delta_{\mathrm{A}}$ and define $y := P_{\mathcal{M}}(x)$, $T := T_{\mathcal{M}}(y)$, and $N := N_{\mathcal{M}}(y)$. We assume δ is small enough that $\sigma \le \delta_{\mathrm{A}}$ and $\|y - \bar{x}\| \le \frac{C_1}{\sqrt{8}C_4}$.

The starting point of the proof is Lebourg's mean value Theorem [11, Theorem 2.4], which ensures that there exists $v \in \partial_\sigma f(x)$ such that

$$f\left(x - \sigma\frac{g}{\|g\|}\right) - f(x) = \left\langle v, -\sigma\frac{g}{\|g\|}\right\rangle = -\frac{\sigma}{\|g\|}\left\langle v, P_T(g)\right\rangle - \frac{\sigma}{\|g\|}\left\langle v, P_N(g)\right\rangle.$$

In what follows, we will show that the first term satisfies $\langle v, P_T(g)\rangle \ge \frac{1}{4}\|g\|^2$, while the second term satisfies $|\langle v, P_N(g)\rangle| \le \frac{1}{8}\|g\|^2$, yielding the result.

Indeed, beginning with $|\langle v, P_N(g)\rangle|$, we have

$$|\langle v, P_N(g)\rangle| \le L\|P_N(g)\| \le C_4 L\|y - \bar{x}\|^2 \le \frac{C_4 L}{C_1^2}\|g\|^2 = \frac{1}{8}\|g\|^2,$$

where the first inequality follows from Lipschitz continuity of f; the second and fourth inequalities follow by assumption; and the third inequality follows by Lemma 4.2. Next we prove a lower bound on $\langle v, P_T(g)\rangle$. Since $v \in \partial_\sigma f(x)$, Equation (3.5) implies that

$$\|P_T(v - g)\| \le 2C_{(a)}(\mathrm{dist}(x, \mathcal{M}) + \sigma).$$

Consequently, we have the bound

$$\begin{aligned}
\langle v, P_T(g)\rangle = \langle P_T(v), P_T(g)\rangle &= \frac{1}{2}\left(\|P_T(v)\|^2 + \|P_T(g)\|^2 - \|P_T(v-g)\|^2\right) \\
&\ge \frac{1}{2}\|P_T(g)\|^2 - 2C_{(a)}^2(\mathrm{dist}^2(x, \mathcal{M}) + \sigma^2) \\
&\ge \frac{1}{2}\|g\|^2 - \|P_N(g)\|^2 - 4C_{(a)}^2 C_3^2\|y - \bar{x}\|^2 \\
&\ge \frac{1}{2}\|g\|^2 - C_4^2\|y - \bar{x}\|^4 - \frac{4C_{(a)}^2 C_3^2}{C_1^2}\|g\|^2 \\
&\ge \frac{1}{2}\|g\|^2 - \frac{C_4^2}{C_1^2}\|y - \bar{x}\|^2\|g\|^2 - \frac{4C_{(a)}^2 C_3^2}{C_1^2}\|g\|^2 \\
&\ge \frac{1}{4}\|g\|^2,
\end{aligned}$$

where the third and fourth inequalities follow from Lemma 4.2. This completes the proof. □

The next Lemma proves the approximate reflection property that was described in the introduction. The lemma roughly shows that normalized subgradient steps approximately "flip the sign" of the normal component of the subgradient; see Section 1.3.1 for more intuition.

Lemma 5.3 (Approximate reflection across manifold)**.** *There exists $C_5, C_6 > 0$ such that for all x near $\bar{x}$, $\sigma > 0$, and $g \in \partial_\sigma f(x)\backslash\{0\}$ satisfying*

$$\max\left\{\frac{\mathrm{dist}(x, \mathcal{M})}{\sigma}, \sigma\right\} \le C_6\|P_{\mathcal{M}}(x) - \bar{x}\|,$$

we have

$$\langle P_{N_{\mathcal{M}}(P_{\mathcal{M}}(x))}(\hat{g}), g\rangle \leq -C_5\|P_{N_{\mathcal{M}}(P_{\mathcal{M}}(x))}g\| + \frac{C_4C_5}{2}\|P_{\mathcal{M}}(x)-\bar{x}\|^2 \qquad \textit{for all } \hat{g} \in \partial f\left(x - \sigma\frac{g}{\|g\|}\right).$$

Proof. We fix $x \in B_\delta(\bar{x})$ for some $\delta \leq \delta_{\mathrm{A}}/2$ and define $y := P_{\mathcal{M}}(x)$, $T := T_{\mathcal{M}}(y)$, and $N := N_{\mathcal{M}}(y)$. We assume δ is small enough that $\sigma \leq \delta_{\mathrm{A}}/2$. We define constants $C_5 := \frac{\mu}{2}$ and

$$C_6 := \min\left\{\frac{\beta}{C_{(a)}(1+\delta_{\mathrm{A}})}, \frac{C_4C_5}{4(1+(1+\delta_{\mathrm{A}})C_{\mathcal{M}})(\mu+L)\beta)}\right\}.$$

We also assume δ is small enough that

$$\|y-\bar{x}\| \leq \frac{\mu}{2(\mu+L)C_6(1+(1+\delta)C_{\mathcal{M}})}. \tag{5.1}$$

We now turn to the proof.

Define $u := \frac{g}{\|g\|}$. Note that since $x \in B_{\delta_{\mathrm{A}}/2}(\bar{x})$ and $\sigma \leq \delta_{\mathrm{A}}/2$, we have $x - \sigma u \in B_{\delta_{\mathrm{A}}}(\bar{x})$. Therefore, by the aiming inequality (3.6), we have

$$\underbrace{\langle \hat{g}, x-\sigma u - P_{\mathcal{M}}(x-\sigma u)\rangle}_{=:A} \geq \underbrace{\mu\,\|x-\sigma u - P_{\mathcal{M}}(x-\sigma u)\|}_{=:B}.$$

We aim to simplify this inequality with (3.1). To that end, first note that

$$\begin{aligned}\|x-\sigma u - P_{\mathcal{M}}(x-\sigma u) - (x - P_{\mathcal{M}}(x) - \sigma P_N(u))\| &= \|P_{\mathcal{M}}(x-\sigma u) - P_{\mathcal{M}}(x) + \sigma P_T(u)\| \\ &\leq C_{\mathcal{M}}(\mathrm{dist}^2(x,\mathcal{M}) + \sigma^2). \end{aligned}\tag{5.2}$$

Consequently, we have

$$\begin{aligned} B &\geq \mu\|x - P_{\mathcal{M}}(x) - \sigma P_N(u)\| - \mu C_{\mathcal{M}}(\mathrm{dist}^2(x,\mathcal{M}) + \sigma^2) \\ &\geq \sigma\mu\|P_N(u)\| - \mu\mathrm{dist}(x,\mathcal{M}) - \mu C_{\mathcal{M}}(\mathrm{dist}^2(x,\mathcal{M}) + \sigma^2). \end{aligned}$$

Likewise, we have

$$\begin{aligned}\langle \hat{g}, \sigma P_N(u)\rangle &= -A + \langle \hat{g}, x - \sigma u - P_{\mathcal{M}}(x-\sigma u) + \sigma P_N u\rangle \\ &\leq -A + \|\hat{g}\|\mathrm{dist}(x,\mathcal{M}) + \|\hat{g}\|C_{\mathcal{M}}(\mathrm{dist}^2(x,\mathcal{M}) + \sigma^2) \\ &\leq -B + L\mathrm{dist}(x,\mathcal{M}) + LC_{\mathcal{M}}(\mathrm{dist}^2(x,\mathcal{M}) + \sigma^2) \\ &\leq -\sigma\mu\|P_N(u)\| + (\mu+L)\mathrm{dist}(x,\mathcal{M}) + (\mu+L)C_{\mathcal{M}}(\mathrm{dist}^2(x,\mathcal{M}) + \sigma^2), \end{aligned}\tag{5.3}$$

where the first inequality follows from triangle inequality and (5.2). Note that so far we have only used the assumption $x \in B_{\delta_{\mathrm{A}}/2}(\bar{x})$ and $\sigma < \delta_{\mathrm{A}}/2$. Next, multiply both sides of the above inequality by $\|g\|$ and bound the right-hand-side:

$$\begin{aligned}\langle P_N\hat{g}, g\rangle &\leq -\mu\|P_N g\| + \frac{(\mu+L)\|g\|\mathrm{dist}(x,\mathcal{M})}{\sigma} + \frac{(\mu+L)\|g\|C_{\mathcal{M}}(\mathrm{dist}^2(x,\mathcal{M}) + \sigma^2)}{\sigma} \\ &\leq -\mu\|P_N g\| + (\mu+L)C_6\|g\|\|y-\bar{x}\| + (\mu+L)C_6C_{\mathcal{M}}\|g\|\|y-\bar{x}\|(1+\delta) \\ &\leq -\mu\|P_N g\| + (1+(1+\delta)C_{\mathcal{M}})(\mu+L)C_6\|g\|\|y-\bar{x}\| \end{aligned}$$

$$\leq -\mu\|P_N g\| + (1 + (1+\delta)C_{\mathcal{M}})(\mu + L)C_6(\|P_T(g)\| + \|P_N(g)\|)\|y - \bar{x}\|$$
$$\leq -\frac{\mu}{2}\|P_N g\| + \frac{C_4 C_5}{4\beta}\|P_T(g)\|\|y - \bar{x}\|,$$

where the second inequality follows from the Lemma assumption and the bound $\operatorname{dist}(x, \mathcal{M}) \leq \|x - \bar{x}\| \leq \delta$; the third inequality follows from combining terms; the fourth inequality follows from the decomposition $g = P_N(g) + P_T(g)$; and the fifth inequality follows from (5.1) and the definition of C_6. The proof will be complete if we can show that

$$\|P_T(g)\| \leq 2\beta\|y - \bar{x}\|.$$

To that end, we have

$$\|P_T(g)\| \leq (C_{(a)}(\operatorname{dist}(x, \mathcal{M}) + \sigma) + \beta\|y - \bar{x}\|) \leq (C_6 C_{(a)}(1 + \delta_{\mathrm{A}}) + \beta)\|y - \bar{x}\| \leq 2\beta\|y - \bar{x}\|,$$

where the first inequality follows from (3.4); the second inequality follows from the lemma assumptions and the bound $\operatorname{dist}(x, \mathcal{M}) \leq C_6\sigma\|y - \bar{x}\| \leq C_6(\delta_{\mathrm{A}}/2)\|y - \bar{x}\|$; and the third inequality follows from the bounds on C_6. This completes the proof. □

The following lemma shows that every step of `TDescent` geometrically decreases the normal component of the subgradient, up to a tolerance of $O(\|P_{\mathcal{M}}(x) - \bar{x}\|^2)$.

Lemma 5.4 (Normal component shrinks geometrically)**.** *There exists constant $C_7 > 0$ such that for all x near $\bar{x}$, $\sigma > 0$, and $\hat{g} \in \partial f(x - \sigma\frac{g}{\|g\|})\backslash\{0\}$ satisfying*

1. $\|P_{N_{\mathcal{M}}(P_{\mathcal{M}}(x))} g\| \geq C_4\|P_{\mathcal{M}}(x) - \bar{x}\|^2$;
2. $\max\left\{\frac{\operatorname{dist}(x,\mathcal{M})}{\sigma}, \sigma\right\} \leq C_7\|P_{\mathcal{M}}(x) - \bar{x}\|$,

the vector $g' = \operatorname{argmin}_{h \in [g, \hat{g}]} \|h\|$ satisfies:

$$\|P_{N_{\mathcal{M}}(P_{\mathcal{M}}(x))}(g')\|^2 \leq \left(1 - \frac{3C_5^2}{64L^2}\right)\|P_{N_{\mathcal{M}}(P_{\mathcal{M}}(x))} g\|^2.$$

Proof. We fix $x \in B_\delta(\bar{x})$ for some $\delta \leq \delta_{\mathrm{A}}$ and define $y := P_{\mathcal{M}}(x)$, $T := T_{\mathcal{M}}(y)$, and $N := N_{\mathcal{M}}(y)$. We assume δ is small enough that $\sigma \leq \delta_{\mathrm{A}}$. We define

$$C_7 := \min\left\{\frac{\beta}{2C_{(a)}}, \frac{C_4 C_5}{32 C_{(a)}\beta}, C_6, C_3\right\}.$$

We also assume δ is small enough that $\|y - \bar{x}\| \leq 1/C_7$. Consequently,

$$\operatorname{dist}(x, \mathcal{M}) \leq C_7\sigma\|y - \bar{x}\| \leq C_7\|y - \bar{x}\|,$$

since $\sigma \leq C_7\|y - \bar{x}\| \leq 1$. We now turn to the proof.

Consider the optimal weight $\lambda' := \operatorname{argmin}_{\lambda \in [0,1]} \|g + \lambda(\hat{g} - g)\|$. By definition we have $g' = g + \lambda'(\hat{g} - g)$. Moreover, a quick calculation shows that

$$\lambda' = \max\left\{\min\left\{-\frac{\langle g, \hat{g} - g\rangle}{\|\hat{g} - g\|^2}, 1\right\}, 0\right\}.$$

We claim that the following bound holds on λ':

$$-\frac{\langle P_N(g), \hat{g}-g\rangle}{8L^2} \leq \lambda' \leq -\frac{3\langle P_N(g), \hat{g}-g\rangle}{2\|P_N(\hat{g}-g)\|^2}. \tag{5.4}$$

Note that (5.4) is an immediate consequence of the following bound:

$$0 \leq -\frac{1}{2}\langle P_N(g), \hat{g}-g\rangle \leq -\langle g, \hat{g}-g\rangle \leq -\frac{3}{2}\langle P_N(g), \hat{g}-g\rangle. \tag{5.5}$$

Indeed, if (5.5) holds, then $\lambda' = \min\left\{-\frac{\langle g,\hat{g}-g\rangle}{\|\hat{g}-g\|^2}, 1\right\}$. Thus, we obtain the upper bound

$$\lambda' \leq -\frac{\langle g, \hat{g}-g\rangle}{\|\hat{g}-g\|^2} \leq -\frac{3}{2}\frac{\langle P_N(g), \hat{g}-g\rangle}{\|g-\hat{g}\|^2} \leq -\frac{3}{2}\frac{\langle P_N(g), \hat{g}-g\rangle}{\|P_N(g-\hat{g})\|^2}.$$

Likewise, we obtain the lower bound

$$\lambda' = \min\left\{-\frac{\langle g, \hat{g}-g\rangle}{\|\hat{g}-g\|^2}, 1\right\} \geq \min\left\{-\frac{\langle g, \hat{g}-g\rangle}{4L^2}, 1\right\} = -\frac{\langle g, \hat{g}-g\rangle}{4L^2} \geq -\frac{\langle P_N(g), \hat{g}-g\rangle}{8L^2},$$

where the first inequality follows from the bound $\|\hat{g}-g\|^2 \leq 2(\|\hat{g}\|^2 + \|g\|^2) \leq 4L^2$. Thus, we now prove (5.5).

To that end, note that (5.5) is equivalent to the following bound:

$$|\langle P_T(g), \hat{g}-g\rangle| \leq \frac{-\langle P_N(g), \hat{g}-g\rangle}{2}. \tag{5.6}$$

Therefore, we first bound $|\langle P_T(g), \hat{g}-g\rangle|$:

$$\begin{aligned}
|\langle P_T(g), \hat{g}-g\rangle| &\leq \|P_T(g)\|\|P_T(\hat{g}-g)\| \\
&\leq 2C_{(a)}(\mathrm{dist}(x,\mathcal{M})+\sigma)(C_{(a)}(\mathrm{dist}(x,\mathcal{M})+\sigma)+\beta\|y-\bar{x}\|) \\
&\leq 4C_{(a)}C_7(2C_{(a)}C_7+\beta)\|y-\bar{x}\|^2 \\
&\leq \frac{C_4C_5}{4}\|y-\bar{x}\|^2,
\end{aligned}$$

where the second inequality follows from (3.4) and (3.5); the third inequality follows from the lemma assumption; and the fourth inequality follows from the definition of C_7. To complete the proof of (5.6), we show that $\frac{C_4C_5}{4}\|y-\bar{x}\|^2 \leq -\frac{1}{2}\langle P_N(g), \hat{g}-g\rangle$:

$$\begin{aligned}
-\langle P_N(g), \hat{g}-g\rangle &\geq -\langle P_N(g), \hat{g}\rangle \\
&\geq C_5\|P_N(g)\| - \frac{C_4C_5}{2}\|y-\bar{x}\|^2 \\
&\geq \frac{C_4C_5}{2}\|y-\bar{x}\|^2,
\end{aligned} \tag{5.7}$$

where the second inequality follows from Lemma 5.3 (recall $C_7 \leq C_6$); and the third inequality follows from the bound $\frac{C_5}{2}\|P_N(g)\| \geq \frac{C_4C_5}{2}\|y-\bar{x}\|^2$. Thus, the equivalent bounds (5.6) and (5.5) hold. Consequently, Equation (5.4) holds.

Now we turn to the contraction argument. Now consider the function $r\colon \mathbb{R} \to \mathbb{R}$ satisfying

$$r(\lambda) = \|P_N(g)\|^2 + 2\lambda \langle P_N(g), \hat{g} - g\rangle + \lambda^2 \|P_N(\hat{g} - g)\|^2 \qquad \text{for all } \lambda \in \mathbb{R}.$$

Observe that

$$\|P_N(g')\|^2 = \|P_N(g)\|^2 + 2\lambda' \langle P_N(g), \hat{g} - g\rangle + (\lambda')^2 \|P_N(\hat{g} - g)\|^2 = r(\lambda').$$

Therefore, by convexity of r and (5.4), we have

$$\|P_N(g')\|^2 = r(\lambda') \le \max\left\{ r\left(\frac{-3\langle P_N(g), \hat{g} - g\rangle}{2\|P_N(\hat{g} - g)\|^2} \right), r\left(\frac{-\langle P_N(g), \hat{g} - g\rangle}{8L^2} \right) \right\}.$$

To complete the proof, we show each term in the "max" is bounded by $\left(1 - \frac{3C_5^2}{64L^2}\right) \|P_N(g)\|^2$.

To show this, we will use the following consequence of (5.7):

$$-\langle P_N(g), \hat{g} - g\rangle \ge C_5 \|P_N(g)\| - \frac{C_4 C_5}{2} \|y - \bar{x}\|^2 \ge \frac{C_5}{2} \|P_N(g)\|, \tag{5.8}$$

where the final inequality follows from the bound $\frac{C_4 C_5}{2}\|y - \bar{x}\|^2 \le \frac{C_5}{2}\|P_N(g)\|$. Indeed, first observe that

$$\begin{aligned} r\left(\frac{-3\langle P_N(g), \hat{g} - g\rangle}{2\|P_N(\hat{g} - g)\|^2} \right) &= \|P_N(g)\|^2 - \frac{3}{4} \frac{\langle P_N(g), \hat{g} - g\rangle^2}{\|P_N(\hat{g} - g)\|^2} \\ &\le \left(1 - \frac{3C_5^2}{16\|P_N(\hat{g} - g)\|^2}\right) \|P_N(g)\|^2 \\ &\le \left(1 - \frac{3C_5^2}{64L^2}\right) \|P_N(g)\|^2, \end{aligned}$$

where the second inequality from (5.8) and the third inequality follows from the bound $\|P_N(\hat{g} - g)\|^2 \le \|\hat{g} - g\|^2 \le 4L^2$. Likewise, observe that

$$\begin{aligned} r\left(\frac{-\langle P_N(g), \hat{g} - g\rangle}{8L^2} \right) &= \|P_N(g)\|^2 - \frac{\langle P_N(g), \hat{g} - g\rangle^2}{4L^2} + \frac{\langle P_N(g), \hat{g} - g\rangle^2 \|P_N(\hat{g} - g)\|^2}{64L^4} \\ &\le \|P_N(g)\|^2 - \frac{\langle P_N(g), \hat{g} - g\rangle^2}{4L^2} + \frac{\langle P_N(g), \hat{g} - g\rangle^2}{16L^2} \\ &\le \left(1 - \frac{3C_5^2}{64L^2}\right) \|P_N(g)\|^2, \end{aligned}$$

where the first inequality follows from the bound $\|P_N(\hat{g} - g)\|^2 \le \|\hat{g} - g\|^2 \le 4L^2$ and the third inequality follows from (5.8). Therefore, the proof is complete. □

The following proposition is the main result of this section. It shows that `TDescent` must either terminate with descent or $f(x) - f(\bar{x})$ is already exponentially small in T.

Proposition 5.5 (`TDescent` loop terminates with descent)**.** *Fix $T \in \mathbb{N}$. Then for all x near $\bar{x}$, $v \in \partial f(x)$, and $\sigma > 0$ satisfying*

$$\max\left\{ \frac{\operatorname{dist}(x, \mathcal{M})}{\sigma}, \sigma \right\} \le \min\{C_2, C_7\} \|y - \bar{x}\|,$$

at least one of the following holds:

1. *we have*
$$f(x) - f(\bar{x}) \leq \frac{(\min\{C_2^2, C_7^2\}L + \beta)L}{C_4}(1 - \frac{3\mu^2}{256L^2})^{T/2};$$

2. *the vector* $g := \texttt{TDescent}(x, v, \sigma, T)$ *satisfies* $\|g\| > 0$ *and*
$$f\left(x - \sigma\frac{g}{\|g\|}\right) \leq f(x) - \frac{\sigma \text{dist}(0, \partial_\sigma f(x))}{8}.$$

Proof. We fix $x \in B_\delta(\bar{x})$ for some $\delta \leq \delta_{\text{A}}$ and define $y := P_{\mathcal{M}}(x)$, and $N := N_{\mathcal{M}}(y)$. We assume δ is small enough that $\sigma \leq \delta_{\text{A}}$ and that the conclusions of Lemmas 4.2, 5.2, and 5.4.

Turning to the proof, note that Lemma 4.2 ensures that $\text{dist}(0, \partial_\sigma f(x)) > 0$. Thus, if $\texttt{TDescent}(x, v, \sigma, T)$ terminates at $t < T$, then Item 2 must hold. For the remainder of the proof, we suppose that $\texttt{TDescent}(x, v, \sigma, T)$ terminates at iteration $t = T$ and that Item 2 does not hold. In this case, Lemma 5.2 ensures that the iterates g_t of $\texttt{TDescent}(x, v, \sigma, T)$ satisfy $\|P_N(g_t)\| > C_4\|y - \bar{x}\|^2$ for all $0 \leq t \leq T$. Therefore, by Lemma 5.4, we have

$$\|P_N(g_{t+1})\|^2 \leq \left(1 - \frac{3C_5^2}{64L^2}\right)\|P_N(g_t)\|^2, \qquad \text{for all } 0 \leq t \leq T-1.$$

As a result,

$$\|y - \bar{x}\|^2 \leq \frac{\|P_N(g_T)\|}{C_4} \leq \frac{\|P_N(g_0)\|}{C_4}\left(1 - \frac{3C_5^2}{64L^2}\right)^{T/2} \leq \frac{L}{C_4}\left(1 - \frac{3C_5^2}{64L^2}\right)^{T/2}.$$

Consequently,

$$\begin{aligned} f(x) - f(\bar{x}) &\leq L\text{dist}(x, \mathcal{M}) + \beta\|y - \bar{x}\|^2 \\ &\leq (\min\{C_2^2, C_7^2\}L + \beta)\|y - \bar{x}\|^2 \\ &\leq \frac{(\min\{C_2^2, C_7^2\}L + \beta)L}{C_4}\left(1 - \frac{3C_5^2}{64L^2}\right)^{T/2}, \end{aligned}$$

where the first inequality follows from (3.7) and the second inequality follows from the lemma assumptions. The proof then follows from the identity $C_5 = \frac{\mu}{2}$. □

5.2.1 Some loose ends: eventually small subgradients

Before ending this section, we must establish one final technical result for `TDescent`. Namely, in Lemma 5.7, we show that for appropriate σ, `TDescent` eventually generates small subgradients on the order of $O(\|x - \bar{x}\|)$. This property is intuitive because $\text{dist}(0, \partial_\sigma f(x)) = 0$ whenever $\sigma \geq \|x - \bar{x}\|$. This property will help us ensure that the iterates of `NTDescent` (Algorithm 4) cannot leave sufficiently small neighborhoods of $\bar{x}$. Indeed, since the subgradients v_{i+1} generated by Algorithm 3 (`linesearch`) are decreasing in norm, we will show that the trust region constraint $\sigma_i \leq \frac{\|v_{i+1}\|}{s}$ in Line 7 of Algorithm 3 must eventually be violated for large i. This ensures large σ_i are never chosen.

To prove this claim, we first establish a refinement of the approximate reflection property in Lemma 5.3. Compared to Lemma 5.3, the following lemma deals with a different range of parameters.

Lemma 5.6 (Approximate reflection across manifold)**.** *Define* $C_8 := \frac{8(\mu+L)}{\mu}$*. Then there exists a constant* $\delta_{\text{Grid}} \in (0, \delta_{\text{A}}/2]$ *such that for all* x *near* $\bar{x}$*,* $\sigma > 0$*, and* $g \in \partial_\sigma f(x) \backslash \{0\}$ *satisfying*

$$C_8 \text{dist}(x, \mathcal{M}) \leq \sigma \leq \delta_{\text{Grid}},$$

we have

$$\langle \hat{g}, g \rangle \leq -C_5 \|g\| + 2C_5 \|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(g)\| \qquad \text{for all } \hat{g} \in \partial f\left(x - \sigma \frac{g}{\|g\|}\right).$$

Proof. We fix $x \in B_\delta(\bar{x})$ for some $\delta \leq \delta_{\text{A}}/2$ and define $y := P_{\mathcal{M}}(x)$, $T := T_{\mathcal{M}}(y)$, and $N := N_{\mathcal{M}}(y)$. We assume that δ and δ_{Grid} are small enough that

$$\max\{2(\mu + L)C_{\mathcal{M}}(\delta/C_8 + \delta_{\text{Grid}}), C_{(a)}(\delta + \delta_{\text{Grid}}) + 2\beta\delta\} \leq \frac{\mu}{4}. \tag{5.9}$$

Then arguing just as in the buildup to (5.3) in the proof of Lemma 5.3 (which only requires $x \in B_{\delta_{\text{A}}/2}(\bar{x})$ and $\sigma \leq \delta_{\text{Grid}} \leq \delta_{\text{A}}/2$), we have

$$\left\langle \hat{g}, \sigma \frac{P_N g}{\|g\|} \right\rangle \leq -\sigma\mu \frac{\|P_N g\|}{\|g\|} + (\mu + L)\text{dist}(x, \mathcal{M}) + (\mu + L)C_{\mathcal{M}}(\text{dist}^2(x, \mathcal{M}) + \sigma^2).$$

Rearranging, we find that

$$\begin{aligned}
\langle P_N \hat{g}, g \rangle &\leq -\mu \|P_N g\| + \frac{(\mu + L)\|g\|\text{dist}(x, \mathcal{M})}{\sigma} + \frac{(\mu + L)\|g\| C_{\mathcal{M}}(\text{dist}^2(x, \mathcal{M}) + \sigma^2)}{\sigma} \\
&\leq -\mu \|P_N g\| + \frac{\mu}{8}\|g\| + (\mu + L)C_{\mathcal{M}}(\text{dist}(x, \mathcal{M})/C_8 + \sigma) \cdot \|g\| \\
&\leq -\mu \|P_N g\| + \frac{\mu}{4}\|g\|,
\end{aligned}$$

where the second inequality follows by definition of C_8 and the third follows from (5.9) together with the bounds $\text{dist}(x, \mathcal{M}) \leq \delta$ and $\sigma \leq \delta_{\text{Grid}}$. Now observe that

$$\langle P_T \hat{g}, g \rangle \leq \|P_T \hat{g}\| \|g\| \leq (C_{(a)}(\text{dist}(x, \mathcal{M}) + \sigma) + \beta\|y - \bar{x}\|) \cdot \|g\| \leq \frac{\mu}{4}\|g\|,$$

where second inequality follows from (3.4) and the third inequality follows from (5.9) together with the following three bounds: (i) $\text{dist}(x, \mathcal{M}) \leq \|x - \bar{x}\| \leq \delta$; (ii) $\|y - \bar{x}\| \leq 2\|x - \bar{x}\| \leq 2\delta$; and (iii) $\sigma \leq \delta_{\text{Grid}}$. Therefore,

$$\langle \hat{g}, g \rangle = \langle P_N \hat{g}, g \rangle + \langle P_T \hat{g}, g \rangle \leq -\mu \|P_N g\| + \frac{\mu}{2}\|g\| \leq -\frac{\mu}{2}\|g\| + \mu \|P_T(g)\|,$$

as desired. □

Finally, we prove that that `TDescent` eventually generates small subgradients.

Lemma 5.7 (`TDescent` yields small subgradients)**.** *Fix* $T \in \mathbb{N}$*. Then for all* x *near* $\bar{x}$*, for all* $\sigma > 0$*, and* $g \in \partial_\sigma f(x) \backslash \{0\}$ *satisfying*

$$C_8 \text{dist}(x, \mathcal{M}) \leq \sigma \leq \delta_{\text{Grid}},$$

the vector $g' := \texttt{TDescent}(x, g, \sigma, T)$ *satisfies*

$$\|g'\| \leq \max\left\{ \left(1 - \frac{\mu^2}{64L^2}\right)^{T/2} \|g\|, 4C_{(a)}\sigma + 4(C_{(a)} + 2\beta)\|x - \bar{x}\|, \frac{8(f(x) - f(\bar{x}))}{\sigma} \right\}.$$

Proof. We fix $x \in B_\delta(\bar{x})$ for some $\delta < \delta_{\mathrm{A}}/2$ and define $y := P_{\mathcal{M}}(x)$ and $\mathcal{T} := T_{\mathcal{M}}(y)$. We assume δ is small enough that the conclusion of Lemma 5.6 holds for x. We also define $c := C_{(a)}(\mathrm{dist}(x, \mathcal{M}) + \sigma) + \beta\|y - \bar{x}\|$. Note that by Proposition 3.5, we have $\|P_{\mathcal{T}}(v)\| \leq c$ for all $v \in \partial_\sigma f(x)$.

Turning to the proof, note that the result is holds automatically if $g' = 0$. Thus, we first consider the case where `TDescent` terminates in descent, meaning

$$f(x_+) - f(x) \leq -\frac{\sigma\|g'\|}{8} \qquad \text{where } x_+ := x - \sigma\frac{g'}{\|g'\|}.$$

Since $\sigma \leq \delta_{\mathrm{Grid}} \leq \delta_{\mathrm{A}}/2$ and $x \in B_{\delta_{\mathrm{A}}/2}(\bar{x})$, it follows that $x_+ \in B_{\delta_{\mathrm{A}}}(\bar{x})$. Thus, by Item 1 of Proposition 3.5, we have

$$f(x_+) \geq f(\bar{x}) + \frac{\gamma}{2}\|x - \bar{x}\|^2 \geq f(\bar{x}).$$

Consequently, we have

$$f(\bar{x}) - f(x) \leq -\frac{\sigma\|g'\|}{8}.$$

Rearranging then gives the upper bound $\|g'\| \leq \frac{8(f(x)-f(\bar{x}))}{\sigma}$, as desired.

Let us now suppose that `TDescent` does not terminate with descent or with $g' = 0$. In this case, the iterates $g_0, \ldots, g_T$ of $\texttt{TDescent}(x, g, \sigma, T)$ exist and satisfy $g_t \in \partial_\sigma f(x)$ for all $t \leq T$. We consider two cases.

Case 1. Now suppose $\|g_t\| \leq 4c$ for some t satisfying $0 \leq t \leq T$. Since $\|g_t\|$ is a decreasing sequence, it follows that $\|g'\| = \|g_T\| \leq 4c$. Recalling that $\mathrm{dist}(x, \mathcal{M}) \leq \|x - \bar{x}\|$ and $\|y - \bar{x}\| \leq 2\|x - \bar{x}\|$ yields the bound:

$$\|g'\| \leq 4c \leq 4C_{(a)}\sigma + 4(C_{(a)} + 2\beta)\|x - \bar{x}\|,$$

as desired.

Case 2. Next suppose that for all $0 \leq t \leq T$ we have $4c < \|g_t\|$. In this case, Lemma 5.6 shows that for all $t \leq T$, we have

$$\langle \hat{g}_t, g_t \rangle \leq -\frac{\mu}{2}\|g_t\| + \mu\|P_{\mathcal{T}} g_t\| \leq -\frac{\mu}{2}\|g_t\| + \mu c \leq -\frac{\mu}{4}\|g_t\|. \tag{5.10}$$

We now use this bound to prove a one-step geometric improvement bound for $\|g_t\|^2$. To that end, fix any $t \leq T - 1$ and define the weight $\lambda := \frac{\mu\|g_t\|}{16L^2}$ and the vector $g_\lambda := g_t + \lambda(\hat{g}_t - g_t)$. Notice that $\lambda \in [0, 1]$, since

$$\lambda = \frac{\mu\|g_t\|}{16L^2} \leq \frac{\mu}{16L} \leq 1,$$

where the first equation follows since $g_t \in \partial_\sigma f(x)$ and the second follows since $L \geq \mu$. Thus

$$\begin{aligned}\|g_{t+1}\|^2 \leq \|g_\lambda\|^2 &= \|g_t\|^2 + 2\lambda\langle g_t, \hat{g}_t - g_t\rangle + \lambda^2\|\hat{g}_t - g_t\|^2 \\ &\leq \|g_t\|^2 + 2\lambda\langle g_t, \hat{g}_t\rangle - 2\lambda\|g_t\|^2 + 4L^2\lambda^2\end{aligned}$$

$$\leq \|g_t\|^2 - \frac{\lambda\mu}{2}\|g_t\| + 4L^2\lambda^2$$
$$= \left(1 - \frac{\mu^2}{64L^2}\right)\|g_t\|^2,$$

where the first inequality follows by definition of g_{t+1}; the second inequality follows from the fact that L is a local Lipschitz constant of f near $\bar{x}$; and the third inequality follows from (5.10). Thus, to complete the proof, simply unfold this recursion to get the bound

$$\|g'\| = \|g_T\| \leq \left(1 - \frac{\mu^2}{64L^2}\right)^{T/2}\|g_0\|^2,$$

as desired. □

6 Rapid local convergence of `NTDescent`

In this Section, we present our main convergence guarantees for the `NTDescent` method under Assumption A. The main results of the section are Theorem 6.3 and Theorem 6.5, which analyze the nonconvex and convex settings respectively. In the nonconvex setting, we prove that iterates of `NTDescent` locally nearly linearly converge, provided some iterate reaches a sufficiently small neighborhood of $\bar{x}$. In the convex setting, we strengthen this guarantee, showing that for any initial starting point x_0 and any failure probability p, there exists some index K_p after which `NTDescent` nearly converges linearly with probability at least $1 - p$. Both results are a consequence of the local one-step improvement bound of Proposition 6.1. This proposition shows that with high probability, the following hold locally for `linesearch`: its output is nearby its input; and the function gap geometrically decreases whenever it is larger than a quantity that is exponentially small in the inner loop budget and the gridsize. The former property is useful later for ensuring that that the iterates of `NTDescent` do not escape a local neighborhood of $\bar{x}$.

Throughout this section, we assume that following assumptions and notation are in force.

Assumptions and Notations. We assume that

1. the budget T_k and gridsize G_k satisfy $\min\{T_k, G_k\} \geq k + 1$ for all $k \geq 0$.
2. We fix an initial we an initial point $x_0 \in \mathbb{R}^d$ and $g_0 \in \partial f(x_0)$. We assume that x_0 is not a critical point, so $g_0 \neq 0$. We denote the initial subgradient norm by $s := \|g_0\|$.

We let $\{x_k\}$ denote the sequence of iterates $\texttt{NTDescent}(x_0, g_0, \{T_k\}, \{G_k\})$ applied to f. We assumption A is in force at a point $\bar{x}$. We fix a base neighborhood $B_\delta(\bar{x})$ with $\delta \leq \delta_A$ small enough that if $x \in B_\delta(\bar{x})$, then conclusions of Proposition 3.5, Theorem 4.3, and Lemmas 4.1, 4.2, 5.1, 5.5, and 5.7 hold. Below, we apply Theorem 4.3 with the following constants a_1, a_2:

$$a_1 = \min\{D_1, D_2/s\} \qquad \text{and} \qquad a_2 = \frac{\min\{C_1/s, C_2, C_7\}}{10}.$$

We furthermore assume that the following upper bound on δ:

$$\delta \leq \min\left\{1, \frac{1}{2(a_1+2a_2)}, \frac{\gamma}{32C_8^2 L}, \frac{\gamma \min\{\delta_{\mathrm{Grid}}/2, 1/4\}^2}{2C_8^2 L}\right\}.$$

Now consider the following three terms, which appear in our convergence rate analysis: for all $T, G > 0$, define

$$\epsilon_{1,T} := \max\left\{\frac{(\min\{C_2^2, C_7^2\}L + \beta)L}{C_4}\left(1 - \frac{3\mu^2}{256L^2}\right)^{T/2}, \left(1 - \frac{\mu^2}{64L^2}\right)^{T/2} L\right\};$$

$$\epsilon_{2,G} := \max\left\{\frac{L}{\min\{1, a_1\}} + \frac{\beta}{2\min\{1, a_1\}a_2^2}, 8C_{(a)}, s\right\} 2^{-G};$$

$$\rho := 1 - \frac{1}{8}\min\left\{\frac{\gamma a_2}{40\max\{2La_2^2, \beta\}}, \frac{\mu a_1}{20\max\{L/2, \beta/a_2^2\}}\right\}.$$

In the statement of the following propositions, the constant $\rho \in (0,1)$ plays the role of a local contraction factor, while the terms $\epsilon_{1,T}$ and $\epsilon_{2,G}$ are upper bounds for function gap of `NTDescent`.

We now turn to the one step improvement argument.

6.1 One step improvement

The following proposition presents our one step improvement bound.

Proposition 6.1 (One step improvement)**.** *Suppose that function f satisfies Assumption A at $\bar{x} \in \mathbb{R}^d$. The following holds for all x sufficiently near $\bar{x}$, subgradients $g \in \partial f(x)$ and gridsizes $G > \lceil \log_2(1/\delta_{\mathrm{Grid}}) \rceil$: Fix a failure probability $p \in (0,1)$ and budget T satisfying*

$$T \geq \left\lceil \frac{256L^2}{\mu^2} \right\rceil \lceil 2\log(1/p) \rceil.$$

Then with probability at least $1-p$, the point $\tilde{x} = \texttt{linesearch}(x, g, s, T, G)$ satisfies

1. $f(\tilde{x}) - f(\bar{x}) \leq \max\{\rho(f(x) - f(\bar{x})), \epsilon_{1,T}, \epsilon_{1,G}\}$;
2. $\|\tilde{x} - x\| \leq C_9 \max\left\{\epsilon_{1,T}/s, \epsilon_{2,G}/s, \sqrt{2(f(x) - f(\bar{x}))/\min\{s, \gamma\}}\right\}$;

where $C_9 := \max\left\{1, \frac{8(C_{(a)} + C_{(a)}C_8 + 2\beta)}{s}, 2C_8, \frac{\gamma}{s}, \frac{4}{C_8}\right\}$.

Proof. We will first establish the first item of the Proposition. To that end, let us assume that $f(x) - f(\bar{x}) > \max\{\epsilon_{1,T}, \epsilon_{2,G}\}$; otherwise the proof is trivial. In this case, we claim that x must satisfy either Item 1 or Item 2 of Theorem 4.3 for at least one σ_i with $i \leq G-1$. To derive a contradiction, suppose that both items are not satisfied for x with any choice of σ_i with $i = 0, \ldots, G-1$. We will show that neither Item 1b nor its complement can be satisfied, leading to a contradiction.

Throughout the following argument, we will use the following bound:

$$\max\{a_1 \mathrm{dist}(x, \mathcal{M}), a_2\|y - \bar{x}\|\} \leq (a_1 + 2a_2)\,\delta \leq \frac{1}{2} = \sigma_{G-1},$$

where $y := P_{\mathcal{M}}(x)$. Now suppose that Item 1b holds, i.e., $a_2^2\|y-\bar{x}\|^2 \leq \text{dist}(x,\mathcal{M})$. Then Item 1a must fail for any σ_i. We claim that this ensures $\sigma_0 > a_1\text{dist}(x,\mathcal{M})$. Indeed, if $\sigma_0 \leq a_1\text{dist}(x,\mathcal{M})$, we must have

$$\sigma_0 \leq (a_1/10)\text{dist}(x,\mathcal{M}) \leq a_1\text{dist}(x,\mathcal{M}) \leq \sigma_{G-1},$$

since σ_0 cannot satisfy Item 1a. Thus, there exists some $j \leq G-1$ such that $\sigma_j = 2^j\sigma_0$ satisfies Item 1a, a contradiction. Therefore, we have

$$\sigma_0 > a_1\text{dist}(x,\mathcal{M}) \geq a_1a_2^2\|y-\bar{x}\|^2.$$

In this case, by (3.7), we have

$$f(x) - f(\bar{x}) \leq L\text{dist}(x,\mathcal{M}) + \frac{\beta}{2}\|y-\bar{x}\|^2 \leq \left(\frac{L}{a_1} + \frac{\beta}{2a_2^2a_1}\right)\sigma_0 \leq \epsilon_{2,G},$$

which is a contradiction. Therefore, Item 1b cannot hold, so we have $a_2^2\|y-\bar{x}\|^2 > \text{dist}(x,\mathcal{M})$.

Next, for the sake of contradiction, suppose that there exists σ_i satisfying Item 2a. In this case, since $\sigma_i \geq (a_2/10)\|y-\bar{x}\|$, we have

$$\text{dist}(x,\mathcal{M}) < a_2^2\|y-\bar{x}\|^2 \leq 10a_1\sigma_i\|y-\bar{x}\|,$$

i.e., σ_i also satisfies Item 2b, which is a contradiction. Therefore no σ_i satisfies Item 2a. We claim that this ensures $\sigma_0 > a_2\|y-\bar{x}\|$. Indeed, if $\sigma_0 \leq a_2\|y-\bar{x}\|$, we must have

$$\sigma_0 \leq (a_2/10)\|y-\bar{x}\| \leq a_0\|y-\bar{x}\| \leq \sigma_{G-1},$$

since σ_0 cannot satisfy Item 2a. Thus, there exists some $j \leq G-1$ such that $\sigma_j = 2^j\sigma_0$ satisfies Item 2a, a contradiction. Therefore, we have

$$\sigma_0 > a_2\|y-\bar{x}\| \geq \sqrt{\text{dist}(x,\mathcal{M})}.$$

In this case, by (3.7), we have

$$f(x) - f(\bar{x}) \leq L\text{dist}(x,\mathcal{M}) + \frac{\beta}{2}\|y-\bar{x}\|^2 \leq \left(L + \frac{\beta}{2a_2^2}\right)\sigma_0^2 \leq \epsilon_{2,G},$$

which is a contradiction. Therefore, there must exist σ_i satisfying either Item 1 or Item 2 of Theorem 4.3.

Let us now fix a σ_i satisfying either Item 1 or Item 2 of Theorem 4.3. Then, by Theorem 4.3, we have the bound

$$\sigma_i\text{dist}(0,\partial_{\sigma_i}f(x)) \geq 8(1-\rho)(f(x)-f(\bar{x})).$$

In what follows, we will use the above bound to prove that with probability at least $1-p$, we have $f(\tilde{x}) - f(\bar{x}) \leq \rho(f(x)-f(\bar{x}))$ whenever $f(x)-f(\bar{x}) > \max\{\epsilon_{1,T},\epsilon_{2,G}\}$.

Contraction case 1: normal step. We first suppose that there exists σ_i satisfying Item 1. In the interest of analyzing $v_{i+1} \in \partial_{\sigma_i} f(x)$, let us show that x, σ_i, and T satisfy the conditions of Lemma 5.1: indeed, by Item 1a of Theorem 4.3, we have

$$0 < \sigma_i \le a_1 \mathrm{dist}(x, \mathcal{M}) \le D_1 \mathrm{dist}(x, \mathcal{M});$$

in addition, from the definition $D_2 = \mu/2$, it follows that T satisfies the conditions of Lemma 5.1. Therefore, with probability at least $1 - p$, we have

$$f\left(x - \sigma_i \frac{v_{i+1}}{\|v_{i+1}\|}\right) - f(\bar{x}) \le f(x) - f(\bar{x}) - \frac{\sigma_i}{8}\mathrm{dist}(0, \partial_{\sigma_i} f(x)) \le \rho(f(x) - f(\bar{x})).$$

Next, we show that v_{i+1} and σ_i satisfy the trust region condition $\sigma_i \le \frac{\|v_{i+1}\|}{s}$. Indeed, recall that by Lemma 4.1, $\mathrm{dist}(0, \partial_{\sigma_i} f(x)) \ge D_2$. Consequently, we have

$$\sigma_i \le a_1 \mathrm{dist}(x, \mathcal{M}) \le \frac{D_2 \delta}{s} \le \frac{\mathrm{dist}(0, \partial_{\sigma_i} f(x))\delta}{s} \le \frac{\|v_{i+1}\|}{s}.$$

Therefore, with probability at least $1 - p$, we have

$$f(\tilde{x}) - f(\bar{x}) \le f\left(x - \sigma_i \frac{v_{i+1}}{\|v_{i+1}\|}\right) - f(\bar{x}) \le \rho(f(x) - f(\bar{x})),$$

as desired.

Contraction case 2: tangent step. Next, we suppose that there exists σ_i satisfying Item 2 of Theorem 4.3. In the interest of analyzing $u_i \in \partial_{\sigma_i} f(x)$, let us show that x, σ_i, and T satisfy the conditions of Lemma 5.5: indeed, by Item 2a of Theorem 4.3, we have

$$\sigma_i \le a_2 \|y - \bar{x}\| \le \min\{C_2, C_7\}\|y - \bar{x}\|;$$

in addition, by Item 2b of Theorem 4.3, we have

$$\mathrm{dist}(x, \mathcal{M})/\sigma_i \le 10 a_2 \|y - \bar{x}\| \le \min\{C_2, C_7\}\|y - \bar{x}\|.$$

Therefore, since $f(x) - f(\bar{x}) > \epsilon_{1,T}$, Lemma 5.5 implies that

$$f\left(x - \sigma_i \frac{u_i}{\|u_i\|}\right) - f(\bar{x}) \le f(x) - f(\bar{x}) - \frac{\sigma_i}{8}\mathrm{dist}(0, \partial_{\sigma_i} f(x)) \le \rho(f(x) - f(\bar{x})).$$

Next, we show that u_i and σ_i satisfy the trust region condition $\sigma_i \le \frac{\|u_i\|}{s}$. To show this, we first note that σ_i and x satisfy the conditions of Lemma 4.2: indeed, by Item 2a of Theorem 4.3, we have

$$\sigma_i \le a_2 \|y - \bar{x}\| \le C_2 \|y - \bar{x}\|;$$

In addition, by Item 2 of Theorem 4.3, we have

$$\mathrm{dist}(x, \mathcal{M}) \le 10 a_2 \sigma_i \|y - \bar{x}\| \le 10 a_2^2 \|y - \bar{x}\|^2 \le C_2 \|y - \bar{x}\|,$$

where the third inequality follows from the bounds $a_2 \leq \frac{C_2}{10}$ and $\|y - \bar{x}\| \leq 2\delta \leq 1/a_2$. Therefore, by Lemma 4.2 we have $\|u_i\| \geq C_1\|y - \bar{x}\|$. Consequently, we have

$$\sigma_i \leq a_2\|y - \bar{x}\| \leq \frac{C_1\|y - \bar{x}\|}{s} \leq \frac{\|u_i\|}{s}.$$

Observe that $v_{i+1} = u_i$ since `NDescent` will terminate at the first iteration if the descent condition is met. Therefore, we must have

$$f(\tilde{x}) - f(\bar{x}) \leq f\left(x - \sigma_i \frac{v_{i+1}}{\|v_{i+1}\|}\right) - f(\bar{x}) \leq \rho(f(x) - f(\bar{x})),$$

as desired.

Having proved the desired contraction $f(\tilde{x}) - f(\bar{x}) \leq \rho(f(x) - f(\bar{x}))$, we now turn to the bound on $\|\tilde{x} - x\|$.

Stepsize bound. We now no longer assume that $f(x) - f(\bar{x}) > \max\{\epsilon_{2,G}, \epsilon_{1,T}\}$. We claim that we have

$$\max_{0 \leq i \leq G-1} \{\sigma_i \colon \sigma_i \leq \|v_{i+1}\|/s\} \leq C_9 \max\left\{\epsilon_{1,T}/s, \epsilon_{2,G}/s, \sqrt{2(f(x) - f(\bar{x}))/\min\{s, \gamma\}}\right\},$$

To prove this, we will apply Lemma 5.7.

To that end, we first verify that there exists an index i such that σ_i satisfies a slightly stronger version of the assumptions of Lemma 5.7. Indeed, recall that by the quadratic growth condition (A1), we have the bound

$$\operatorname{dist}(x, \mathcal{M}) \leq \|x - \bar{x}\| \leq \sqrt{2(f(x) - f(\bar{x}))/\gamma}.$$

Thus, to satisfy the assumptions of Lemma 5.7, we prove that there exists i such that

$$R_x := C_8\sqrt{2(f(x) - f(\bar{x}))/\gamma} \leq \sigma_i \leq \delta_{\text{Grid}}. \tag{6.1}$$

Indeed, first notice that $\sigma_0 \leq \delta_{\text{Grid}}$ since $G \geq \lceil \log_2(1/\delta_{\text{Grid}}) \rceil$. Thus, if $\sigma_0 \geq R_x$, the bound (6.1) holds for σ_0. If instead $\sigma_0 < R_x$, we have

$$\sigma_0 < R_x \leq C_8\sqrt{2L\delta/\gamma} \leq \min\{\delta_{\text{Grid}}/2, 1/4\} \leq \min\{\delta_{\text{Grid}}, 1/2\} \leq 1/2 = \sigma_{G-1},$$

where the second inequality follows by Lipschitz continuity of f and the inclusion $x \in B_\delta(\bar{x})$. Thus, there exists i such that $\sigma_i \in [\min\{\delta_{\text{Grid}}/2, 1/4\}, \min\{\delta_{\text{Grid}}, 1/2\}]$. Since $\min\{\delta_{\text{Grid}}/2, 1/4\} \geq R_x$, inequality (6.1) follows.

Now let i_* be the minimal such index such that (6.1) is satisfied for $i = i_*$. If $i_* \neq 0$, the bound $\sigma_{i_*-1} \leq R_x$ holds. In particular, $\sigma_{i_*} \leq 2R_x$. Therefore, considering the cases $i_* = 0$ and $i_* \neq 0$ separately, we have

$$R_x \leq \sigma_{i_*} \leq \max\{\sigma_0, 2R_x\}.$$

Now we bound the step length $\|x - \tilde{x}\|$ by considering two cases.

First suppose that $\sigma_{i_*} > \|u_{i_*}\|/s$. In this case, (2.1) ensures $\sigma_{i_*} > \|v_{i_*+1}\|/s$. Then, since σ_i is increasing in i, we have

$$\begin{aligned}\|x - \tilde{x}\| &\le \max_{0\le i\le G-1} \{\sigma_i \colon \sigma_i \le \|v_{i+1}\|/s\} \\ &\le \sigma_{i_*} \\ &\le \max\{\sigma_0, 2R_x\} \\ &\le C_9 \max\left\{\epsilon_{2,G}/s, \sqrt{2(f(x) - f(\bar{x}))/\min\{s,\gamma\}}\right\}.\end{aligned}$$

Therefore have

$$\|x - \tilde{x}\| \le C_9 \max\left\{\epsilon_{1,T}/s, \epsilon_{2,G}/s, \sqrt{2(f(x) - f(\bar{x}))/\min\{s,\gamma\}}\right\},$$

as desired.

Next suppose that $\sigma_{i_*} \le \|u_{i_*}\|/s$. We consider two subcases. First suppose that the following bound also holds:

$$\|u_{i_*}\| \le \frac{8(f(x) - f(\bar{x}))}{\sigma_{i_*}}. \tag{6.2}$$

Then, since $\sigma_{i_*} \ge R_x$, we have

$$\|u_{i_*}\| \le \sqrt{32\gamma(f(x) - f(\bar{x}))/C_8^2}.$$

Second, suppose that (6.2) does not hold. Let us apply Lemma 5.7:

$$\begin{aligned}\|u_{i_*}\| &\le \max\left\{\left(1 - \frac{\mu^2}{64L^2}\right)^{T/2} L, 4C_{(a)}\max\{\sigma_0, 2C_8\|x - \bar{x}\|\} + 4(C_{(a)} + 2\beta)\|x - \bar{x}\|\right\} \\ &\le \max\left\{\left(1 - \frac{\mu^2}{64L^2}\right)^{T/2} L, 8C_{(a)}\sigma_0, 8(C_{(a)} + C_8C_{(a)} + 2\beta)\|x - \bar{x}\|\right\} \\ &\le \max\{1\cdot\epsilon_{1,T}, 1\cdot\epsilon_{2,T}, 8(C_{(a)} + C_8C_{(a)} + 2\beta)\|x - \bar{x}\|\} \\ &\le sC_9\max\{\epsilon_{1,T}/s, \epsilon_{2,G}/s, \|x - \bar{x}\|\},\end{aligned}$$

where the second inequality follows by considering cases $C_{(a)}\sigma_0 \le (C_{(a)} + \beta)\|x - \bar{x}\|$ and $C_{(a)}\sigma_0 > (C_{(a)} + \beta)\|x - \bar{x}\|$. Therefore, as long as $\sigma_{i_*} \le \|u_{i_*}\|/s$, we have

$$\|u_{i_*}\| \le \max\left\{sC_9\max\{\epsilon_{1,T}/s, \epsilon_{2,G}/s, \|x - \bar{x}\|\}, \sqrt{32\gamma(f(x) - f(\bar{x}))/C_8^2}\right\}.$$

To complete the proof, recall that by (2.1), for all $j > i_*$, we have $\|v_j\| \le \|u_{i_*}\|$. Thus,

$$\begin{aligned}&\max_{0\le i\le G-1} \{\sigma_i \colon \sigma_i \le \|v_{i+1}\|/s\} \\ &\le \max_{0\le i\le G-1} \{\sigma_{i_*}, \|u_{i_*}\|/s\} \\ &\le \max\left\{\sigma_0, 2C_8\|x - \bar{x}\|, C_9\max\{\epsilon_{1,T}/s, \epsilon_{2,G}/s, \|x - \bar{x}\|\}, \sqrt{32\gamma(f(x) - f(\bar{x}))/(C_8s^2)}\right\}\end{aligned}$$

$$\leq \max\left\{2C_8\|x-\bar{x}\|, C_9\max\{\epsilon_{1,T}/s, \epsilon_{2,G}/s, \|x-\bar{x}\|\}, \sqrt{32\gamma(f(x)-f(\bar{x}))/(C_8 s^2)}\right\}$$
$$\leq C_9\max\left\{\epsilon_{1,T}/s, \epsilon_{2,G}/s, \|x-\bar{x}\|, \sqrt{2(f(x)-f(\bar{x}))/\gamma}\right\}$$
$$\leq C_9\max\left\{\epsilon_{1,T}/s, \epsilon_{2,G}/s, \sqrt{2(f(x)-f(\bar{x}))/\min\{s,\gamma\}}\right\},$$

where the third inequality follows from the bound $\sigma_0 \leq C_9\epsilon_{1,T}/s$; the fourth inequality follows from the bounds $2C_8\|x-\bar{x}\| \leq C_9\|x-\bar{x}\|$ and $C_9 \geq \gamma/s$; and the fifth inequality follows from the bound $\|x-\bar{x}\| \leq \sqrt{2(f(x)-f(\bar{x}))/\gamma}$, a consequence of quadratic growth (A1). Therefore, we have the stepsize bound:

$$\|\tilde{x}-x\| \leq \max_{0\leq i\leq G-1}\{\sigma_i \colon \sigma_i \leq \|v_{i+1}\|/s\} \leq C_9\max\left\{\epsilon_{1,T}/s, \epsilon_{2,G}/s, \sqrt{2(f(x)-f(\bar{x}))/\min\{s,\gamma\}}\right\},$$

as desired. □

6.2 Main convergence theorems

We are now ready to prove the main results of this work. The goal of this section is to prove that the following event occurs with high probability.

Definition 6.2 ($E_{k_0,q,C}$)**.** For any $k_0 > 0$, $q \in (0,1)$ and $C > 0$, let $E_{k_0,q,C}$ denote the event that for all $k \geq k_0$, we have the following two bounds:

$$f(x_k) - f(\bar{x}) \leq \max\{(f(x_{k_0})-f(\bar{x}))q^{k-k_0}, Cq^k\};$$
$$\|x_k-\bar{x}\|^2 \leq \frac{2}{\gamma}\max\{(f(x_{k_0})-f(\bar{x}))q^{k-k_0}, Cq^k\}.$$

We will lower bound the probability of the event $E_{k_0,q,C}$ in both nonconvex and convex settings for a particular choice of q and C. In the nonconvex setting, our result will lower bound the conditional probability of $E_{k_0,q,C}$, given that iterate x_{k_0} enter a sufficiently small neighborhood of $\bar{x}$. To prove the result, we will simply iterate the one-step improvement bound of Proposition 6.1. In the convex setting, we will lower bound the unconditional probability of $E_{k_0,q,C}$. To prove this result, we will combine the conditional result with the sublinear convergence guarantee of Theorem 2.4. We begin with the nonconvex setting.

Theorem 6.3 (Main Theorem: Nonconvex Setting)**.** *Let $C' = \frac{2048L^2}{\mu^2}$. Then there exists $K > 0$ and a neighborhood U of $\bar{x}$ such that the following holds: For every $k_0 \geq K$, we have*

$$P(E_{k_0,q,C} \mid x_{k_0} \in U) \geq 1 - \frac{\exp\left(-\frac{k_0+1}{C'}\right)}{1-\exp\left(-\frac{k_0+1}{C'}\right)},$$

where $q := \max\left\{\rho, \sqrt{1-\frac{3\mu^2}{256L^2}}, \frac{1}{2}\right\}$ *and*

$$C := \max\left\{\frac{(\min\{C_2^2, C_7^2\}L+\beta)L}{C_4}, L, \frac{L}{\min\{1,a_1\}} + \frac{\beta}{2\min\{1,a_1\}a_2^2}, 8C_{(a)}, s\right\}.$$

Proof. Let δ be small enough and let K be large enough that the conclusions of Propositions 3.5 and 6.1 holds for all x in $B_\delta(\bar{x})$ and $G, T \geq K$. Assume that K is large enough that for all $k \geq K$, we have

$$\epsilon_{1,T_k}/s \leq \sqrt{2\epsilon_{1,T_k}/\gamma'} \text{ and } \epsilon_{2,G_k}/s \leq \sqrt{2\epsilon_{2,G_k}/\gamma'} \qquad \text{where } \gamma' := \min\{s, \gamma\}. \tag{6.3}$$

We now choose a neighborhood $U := B_{\delta'}(\bar{x})$ of $\bar{x}$ with $\delta' > 0$ small enough that x_k does not escape $B_\delta(\bar{x})$ for all $k \geq k_0$. To that end, for all $k_0 \geq K$ and $\delta' > 0$, define

$$D_{k_0,\delta'} := \sum_{k=k_0}^{\infty} C_9 \max\left\{\sqrt{2\max\{\delta' L q^{(k-k_0)}, Cq^k\}/\gamma'}, \sqrt{2\epsilon_{1,T_k}/\gamma'}, \sqrt{2\epsilon_{2,G_k}/\gamma'}\right\}.$$

Notice that $D_{k_0,\delta'}$ is finite due to the lower bound $\min\{T_k, G_k\} \geq k+1$. Now fix K large enough and δ' small enough that

$$D_{k_0,\delta'} + \delta' \leq \delta/2 \qquad \text{for all } k_0 \geq K.$$

Using this notation, our neighborhood is $U := B_{\delta'}(\bar{x})$.

We now turn to the proof. Consider the following sequence defined for all $k \geq k_0$:

$$b_k := \delta' + \sum_{j=k_0}^{k-1} C_9 \max\left\{\sqrt{2\max\{\delta' L q^{(k-k_0)}, Cq^k\}/\gamma'}, \sqrt{2\epsilon_{1,T_k}/\gamma'}, \sqrt{2\epsilon_{2,G_k}/\gamma'}\right\}.$$

Note that $b_k \leq \delta/2$ for all $k \geq k_0$. Define the event $F_{k_0} := \{x_{k_0} \in U\}$. For every $k \geq k_0$, define the quantity $R_k := \max\{(f(x_{k_0}) - f(\bar{x}))q^{k-k_0}, Cq^k\}$ and the decreasing sequence of events

$$A_k := \bigcap_{j=k_0}^{k} \{f(x_j) - f(\bar{x}) \leq R_j \text{ and } \|x_j - \bar{x}\| \leq b_j\}.$$

We claim that for all $k \geq k_0$, we have

$$P(A_{k+1} \mid A_k \cap F_{k_0}) \geq 1 - \exp(-T_k/C'). \tag{6.4}$$

Indeed, Proposition 6.1 implies that, conditioned on $A_k \cap F_{k_0}$, the following three inequalities are satisfied with probability at least $1 - \exp(-T_k/C')$:

1. $\|x_k - \bar{x}\| \leq b_k$;
2. $\|x_{k+1} - x_k\| \leq C_9 \max\left\{\epsilon_{1,T_k}/s, \epsilon_{2,G_k}/s, \sqrt{2(f(x_k) - f(\bar{x}))/\gamma'}\right\}$;
3. $f(x_{k+1}) - f(\bar{x}) \leq \max\{\rho(f(x_k) - f(\bar{x})), \epsilon_{1,T_k}, \epsilon_{2,G_k}\}$.

Thus, the bound (6.4) will follow if we can prove that whenever the above three conditions hold, we have $\|x_{k+1} - \bar{x}\| \leq b_{k+1}$ and $f(x_{k+1}) - f(\bar{x}) \leq R_{k+1}$.

To that end, we first prove $\|x_{k+1} - \bar{x}\| \leq b_{k+1}$. Indeed,

$$\|x_{k+1} - \bar{x}\| \leq \|x_{k+1} - x_k\| + \|x_k - \bar{x}\|$$

$$
\begin{aligned}
&\leq C_9 \max\left\{\epsilon_{1,T_k}/s, \epsilon_{2,G_k}/s, \sqrt{2(f(x_k) - f(\bar{x}))/\gamma'}\right\} + b_k \\
&\leq C_9 \max\left\{\sqrt{2\max\{(f(x_{k_0}) - f(\bar{x}))q^{(k-k_0)}, Cq^k\}/\gamma'}, \sqrt{2\epsilon_{1,T_k}/\gamma'}, \sqrt{2\epsilon_{2,G_k}/\gamma'}\right\} + b_k \\
&\leq C_9 \max\left\{\sqrt{2\max\{\delta' L q^{(k-k_0)}, Cq^k\}/\gamma'}, \sqrt{2\epsilon_{1,T_k}/\gamma'}, \sqrt{2\epsilon_{2,G_k}/\gamma'}\right\} + b_k = b_{k+1},
\end{aligned}
$$

where the third inequality follows from (6.3) and the definition of A_k; and the last inequality follows by Lipschitz continuity: $f(x_{k_0}) - f(\bar{x}) \leq L\|x_{k_0} - \bar{x}\| \leq L\delta'$. Next, we prove the bound on $f(x_{k+1}) - f(\bar{x}) \leq R_{k+1}$. Indeed,

$$
\begin{aligned}
f(x_{k+1}) - f(\bar{x}) &\leq \max\{\rho(f(x_k) - f(\bar{x})), \epsilon_{1,T_k}, \epsilon_{2,G_k}\} \\
&\leq \max\{\rho \max\{(f(x_{k_0}) - f(\bar{x}))q^{k-k_0}, Cq^k\}, \epsilon_{1,T_k}, \epsilon_{2,G_k}\} \\
&\leq \max\{R_{k+1}, \epsilon_{1,T_k}, \epsilon_{2,G_k}\}.
\end{aligned}
$$

Next, we prove that $\max\{\epsilon_{1,T_k}, \epsilon_{2,G_k}\} \leq R_{k+1}$. Beginning with ϵ_{1,T_k}, we have

$$
\begin{aligned}
\epsilon_{1,T_k} &= \max\left\{\frac{(\min\{C_2^2, C_7^2\}L + \beta)L}{C_4}\left(1 - \frac{3\mu^2}{256L^2}\right)^{T_k/2}, \left(1 - \frac{\mu^2}{64L^2}\right)^{T_k/2} L\right\} \\
&\leq C \max\left\{\left(1 - \frac{3\mu^2}{256L^2}\right)^{\frac{T_k}{2}}, \left(1 - \frac{\mu^2}{64L^2}\right)^{\frac{T_k}{2}}\right\} \\
&\leq Cq^{k+1} \leq R_{k+1},
\end{aligned}
$$

where the first and second inequalities follow from the definitions of C and q together with the lower bound $T_k \geq k + 1$. Turning to ϵ_{2,G_k}, we have

$$
\epsilon_{2,G_k} = \max\left\{\frac{L}{\min\{1, a_1\}} + \frac{\beta}{2\min\{1, a_1\}a_2^2}, 8C_{(a)}, s\right\} 2^{-G_k} \leq C2^{-G_k} \leq Cq^{k+1} \leq R_{k+1},
$$

where the first and second inequalities follow from the definition of C and q together with the lower bound $G_k \geq k+1$. Putting these bounds together, arrive at the desired inequality: $f(x_{k+1}) - f(\bar{x}) \leq R_{k+1}$. Consequently, the bound (6.4) holds. Moreover, due to the bound $T_k \geq k + 1$, we have

$$
P(A_{k+1} \mid A_k \cap F_{k_0}) \geq 1 - \exp(-T_k/C') \geq 1 - \exp(-(k+1)/C'). \tag{6.5}
$$

Now we relate A_k to E_{k_0}. To that end, by the conditional law of total probability, for all $k \geq k_0$, we have

$$
P(A_{k+1} \mid F_{k_0}) \geq P(A_{k+1} \mid A_k \cap F_{k_0})P(A_k \mid F_{k_0}) \geq P(A_k \mid F_{k_0}) - \exp(-(k+1)/C'),
$$

Therefore, for all $k \geq k_0$, we have

$$
P(A_k \mid F_{k_0}) \geq P(A_{k_0} \mid F_{k_0}) - \sum_{j=k_0+1}^{\infty} \exp(j/C') \geq 1 - \frac{\exp(-\frac{k_0+1}{C'})}{1 - \exp(-\frac{k_0+1}{C'})},
$$

where the final inequality follows since $P(A_{k_0} \mid F_{k_0}) = 1$. Now recall that $\sup_{k \geq k_0} b_k \leq \delta/2$. Therefore, defining the event

$$E'_{k_0,q,C} := \{f(x_k) - f(\bar{x}) \leq R_k \text{ for all } k \geq k_0 \text{ and and } x_k \in B_\delta(\bar{x})\},$$

we have

$$P(E'_{k_0,q,C} \mid F_{k_0}) \geq \lim_{k \to \infty} P(A_k \mid F_{k_0}) \geq 1 - \frac{\exp(-\frac{k_0+1}{C'})}{1 - \exp(-\frac{k_0+1}{C'})}.$$

Next, recall that since $\delta \leq \delta_{\mathrm{A}}$, the quadratic growth bound (A1)

$$\|x_k - \bar{x}\|^2 \leq \frac{2}{\gamma}(f(x_k) - f(\bar{x})) \leq \frac{2}{\gamma} R_k$$

holds for every $k \geq k_0$ within the event $E'_{k_0,q,C}$. Thus, $E_{k_0,q,C} \supseteq E'_{k_0,q,C}$. Therefore, we have

$$P(E_{k_0,q,C} \mid F_{k_0}) \geq P(E'_{k_0,q,C} \mid F_{k_0}) \geq 1 - \frac{\exp(-\frac{k_0+1}{C'})}{1 - \exp(-\frac{k_0+1}{C'})},$$

as desired. □

Now we turn to the convex setting. Our goal is to prove a lower bound on $P(E_{k_0,q,C})$ for all sufficiently large k_0. Before stating the result, we recall a simple fact about convex functions satisfying Assumption A; we place the proof in Appendix C.

Lemma 6.4. *Suppose that function f is convex. Then f has bounded sublevel sets. In addition, for every neighborhood U of $\bar{x}$, there exists a constant $a > 0$ such that*

$$\{x \in \mathbb{R}^d \colon f(x) - f(\bar{x}) \leq a\} \subseteq U.$$

We now turn to our main theorem.

Theorem 6.5 (Main Theorem: Convex setting)**.** *Suppose that f is convex. Fix a failure probability $p \in (0, 1)$. Then there exists a constant $K_p > 0$ such that for any $k_0 \geq K_p$,*

$$P(E_{k_0,q,C}) \geq 1 - p,$$

where q and C are defined as in Theorem 6.3.

Proof. Choose $K > 0$ and U as in Theorem 6.3. Let a satisfy

$$\{x \in \mathbb{R}^d \colon f(x) - f(\bar{x}) \leq a\} \subseteq U.$$

Using Theorem 6.3, choose $K_1 \geq K$ large enough that for $k_0 \geq K_1$, we have

$$P(E_{k_0,q,C} \mid x_{k_0} \in U) \geq 1 - p/2. \tag{6.6}$$

In the remainder of the proof, we prove there exists $K_p \geq K_1$ such that $P(x_{k_0} \in U) \geq 1 - p/2$ for all $k_0 \geq K_p$. Note that this yield the proof, since in that case

$$P(E_{k_0,q,C}) \geq P(E_{k_0,q,C} \mid x_{k_0} \in U) P(x_{k_0} \in U) \geq 1 + p^2/4 - p \geq 1 - p.$$

To lower bound $P(x_{k_0} \in U)$, we apply the Theorem 2.4. To that end, first note that f is Lipschitz continuous on

$$S := \{x + u \colon f(x) \leq f(x_0) \text{ and } u \in \overline{B}(x)\}.$$

since S is bounded by Lemma 6.4. Let L' denote this Lipschitz constant. Now choose K_2 large enough that there exists $i \leq \min_{K_2 \leq k \leq 2K_2 - 1}\{G_k\}$ satisfying

$$(1/2)K_2^{-1/2} \leq \sigma_i \leq K_2^{-1/2}.$$

In addition, assume K_2 is large enough that

$$\operatorname{diam}(S) \max\left\{\frac{16(f(x_{K_2}) - \inf f)}{K_2^{1/2}}, \frac{32L'\sqrt{2\log(2K_2^2/p)}}{K_2^{1/2}}, \sqrt{128}L'K_2^{-1/2}\right\} + 4\frac{L'}{K_2^{1/2}} \leq a. \tag{6.7}$$

Then, by Theorem 2.4, we have with probability at least $1 - p/2$, the bound

$$f(x_{2K_2-1}) - f(\bar{x}) \leq a.$$

When this bound holds, we have $x_k \in U$ for all $k \geq 2K_2 - 1$ since $f(x_k)$ is nonincreasing. Therefore, defining $K_p = \max\{2K_2 - 1, K_1\}$, we have

$$P(x_{k_0} \in U) \geq 1 - p/2 \qquad \text{for all } k_0 \geq K_p,$$

as desired. □

Thus, we have established a local nearly linear convergence rate for `NTDescent`. Before moving to a brief numerical illustration, we explain how Theorem 1.1 from the introduction follows from the above results.

Remark 1 (Establishing Theorem 1.1)**.** Theorem 1.1 from the introduction immediately follows from Theorems 6.3 and 6.5. Indeed, first the event $E_{k_0,q,C}$ from Theorems 6.3 and 6.5 is slightly stronger than the corresponding event $E_{k_0,q,C}$ from Theorem 1.1 for particular q and C. Second, from looking at the proofs, we note that U depends solely on f, while K_p depends only on p and f. The main loose end is the dependence of K_p on p, which we seek to show is on the order of $K_p = O(\max\{\log(1/p), 1\})$. Looking through the proofs, we find this dependence of K_p on p arises from two required bounds: first from (6.6)

$$1 - \frac{\exp\left(-\frac{K_p+1}{C'}\right)}{1 - \exp\left(-\frac{K_p+1}{C'}\right)} \geq 1 - p/2;$$

second from (6.7)

$$\frac{32L\sqrt{2\log(2K_p^2/p)}}{K_p^{1/2}} \leq C''.$$

for some constant $C'' > 0$ depending only on f. From these bounds we see that we may choose $K_p = O(\max\{\log(1/p), 1\})$, as desired.

7 Numerical illustration

In this section, we briefly illustrate the numerical performance of `NTDescent` on two nonsmooth objective functions, borrowed from [1, 9, 29, 31]. In both experiments, we compare `NTDescent` to the subgradient method with the popular Polyak stepsize (`PolyakSGM`) [38], which iterates

$$x_{k+1} = x_k - \frac{f(x_k) - \inf f}{\|w_k\|^2} w_k \qquad \text{for some } w_k \in \partial f(x_k).$$

In the first example, $\inf f$ is known, in the second, we estimate $\inf f$ from multiple runs of `NTDescent`. We choose to compare against the subgradient method because it is a simple first-order method with strong convergence guarantees in convex [38] and nonconvex settings [16]. Importantly, `PolyakSGM` accesses the objective solely through function and subgradient evaluations. Thus, we compare the accuracy achieved by `PolyakSGM` and `NTDescent` after a fixed number of oracle calls, i.e., evaluations of ∂f.

Let us comment on the implementation of `NTDescent`. First, in both experiments, we do not tune parameters of `NTDescent`. Instead, we simply choose

$$T_k = k + 1 \qquad \text{and} \qquad G_k = \min\{k+1, \lceil \log_2(10^{-16}) \rceil\} \quad \text{for all } k \geq 0.$$

Second, we attempt to save first-order oracle calls by breaking the loop on Lines 2 through 6 of Algorithm 3 whenever we find that $\sigma_i > \|v_{i+1}\|/s$. Since σ_i is increasing in i and $\|v_{i+1}\|$ is nonincreasing in i, this does not affect the iterates x_k of `NTDescent`; see Lemma 2.1.

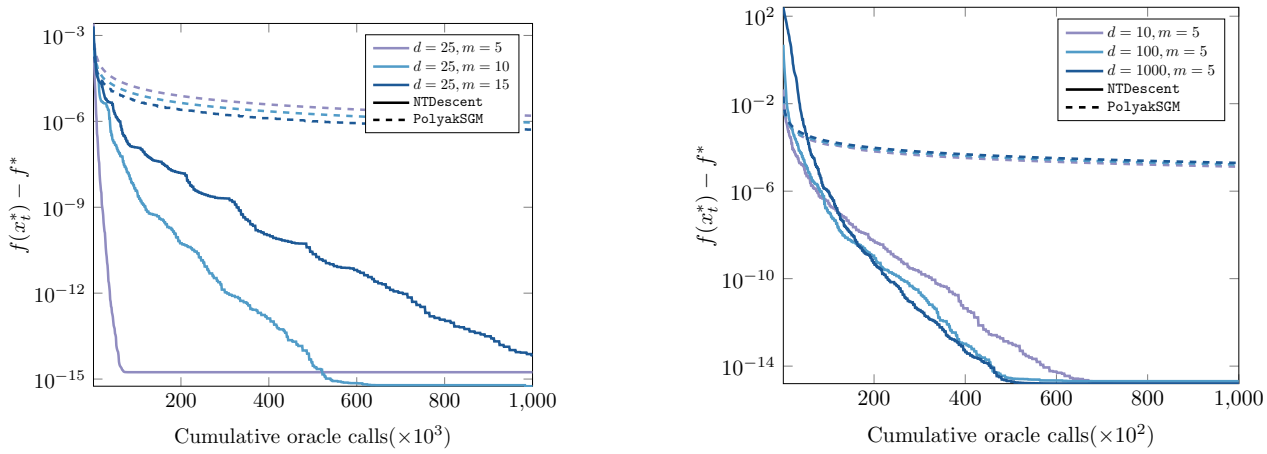
We now turn to the examples.

7.1 A max-of-smooth function

In this example, f takes the following form

$$f(x) = \max_{i=1,\ldots,m} \left\{ g_i^\top x + \frac{1}{2} x^T H_i x \right\}, \tag{7.1}$$

where we generate a random vector $\lambda \in \mathbb{R}^m$ in $\{\lambda > 0 \colon \sum_{i=1}^m \lambda_i = 0\}$, a random positive semi-definite matrix H_i, and a random vector g_i satisfying that $\sum_{i=1}^m \lambda_i g_i = 0$. In this case, one can show that with probability 1, f satisfies Assumption A at its unique minimizer 0.



(a) (b)

Figure 4: Comparison of `NTDescent` with `PolyakSGM` on (7.1). Left: we fix the dimension of the problem and vary the number of smooth functions; Right: we fix the number of smooth functions and vary the dimension of problem. For both algorithms the value $f(x_t^*)$ denotes the best function seen after t oracle evaluations.

In Figure 4 we plot the performance of `NTDescent` and `PolyakSGM` for multiple pairs of (d, m). Figure (4a) shows that the performance of `NTDescent` depends on m. On the other hand, Figure (4b) shows `NTDescent` performance is independent of d, as expected. Both plots show that `NTDescent` outperforms `PolyakSGM`.

Before turning to our second experiment, let us briefly mention two alternative methods – Prox-linear [10, 21, 22, 43, 46] and Survey Descent [24] – which could be applied to this problem. In order to explain these algorithms, let us write $f = \max_{i=1,\dots,m}\{f_i\}$, where the f_i are the quadratic function from (7.1).

Prox-linear method. Given a point $x \in \mathbb{R}^d$, the Prox-linear update x_+ solves

$$x_+ = \operatorname*{argmin}_{y \in \mathbb{R}^d} \max_{i=1,\dots,m} \{f_i(x) + \langle \nabla f_i(x), y - x\rangle\} + \frac{\rho}{2}\|y - x\|^2.$$

One may show that x_+ geometrically improves on x; see [18]. However, in contrast to `NTDescent`, the prox-linear method requires that the components f_i are known. This is stronger than the first-order oracle model considered in this work. Thus, we do not compare `NTDescent` with prox-linear.

Survey Descent The Survey Descent method is a multi-point generalization of gradient descent, designed for max-of-smooth functions. Rather than maintaining a single iterate sequence, the Survey Descent maintains a *survey* S of points, meaning a collection of points $\{s_i\}_{i=1}^m$ at which f is differentiable. A single iteration of the Survey Descent method then aims to produce a new survey $S^+ = \{s_i^+\}_{i=1}^m$ satisfying

$$s_i^+ := \operatorname*{argmin}_{x \in \mathbb{R}^d} \left\| x - \left(s_i - \frac{1}{L}\nabla f(s_i)\right)\right\|^2$$

$$\text{subject to: } f(s_j) + \langle \nabla f(s_j), x - s_j \rangle + \frac{L}{2}\|x - s_j\|^2 \leq f(s_i) + \langle \nabla f(s_i), x - s_i \rangle \quad \forall j \neq i.$$

Here, L is an upper bound on the Lipschitz constant of ∇f_i for all $i = 1, \ldots, m$. In [24], Han and Lewis study linear convergence of Survey Descent on max-of-smooth functions under the conditions of Corollary 3.4. Given a survey S, they show that the updated survey S^+ geometrically improves on S (in an appropriate sense) whenever the following conditions are satisfied: (i) all elements of the survey S are near $\bar{x}$; (i) the survey S is *valid*, meaning there exists a permutation a such that

$$f_{a(i)}(s_i) = f(s_i) \qquad \text{and} \qquad \partial f(s_i) = \{\nabla f_{a(i)}(s_i)\} \qquad \text{for all } i = 1, \ldots, m.$$

To estimate the number of components m and find a valid initial survey S sufficiently close to $\bar{x}$, Han and Lewis suggest an empirical procedure based on running a nonsmooth variant of BFGS [31] for several iterations. After running BFGS, they suggest to (i) compute an estimate $\hat{m}$ of m from a singular value decomposition of the computed gradients, and (ii) build the survey from $\hat{m}$ past iterates in such a way that the computed gradients form an affine independent set. From the numerical illustration in [24], Survey Descent performs well on several small problems. However, since the initialization procedure and implementation of Survey Descent are somewhat sophisticated, we leave a detailed comparison between `NTDescent` and Survey Descent and to future work.

7.2 An eigenvalue product function

In this example, we aim to optimize a function $\tilde{f}$ that takes the following form

$$\tilde{f}(X) = \log E_K(A \odot X),$$

where A is a fixed positive semi-definite data matrix, $E_K(Y)$ denotes the product of K largest eigenvalues of a symmetric matrix $Y \in \mathbb{S}^N$, and $\odot$ denotes the Hadamard (entrywise) matrix product, subject to the constraint that X is positive semi-definite and its diagonal entries are 1. This example is a nonconvex relaxation of an entropy minimization problem arising in an environmental application [1, 9]. In our experiments, we choose A as in [1]: A is the leading $N \times N$ submatrix of a 63×63 covariance matrix, scaled so that the largest entry is 1. As suggested by [9], we reformulate this problem as an unconstrained optimization problem using a Burer-Monteiro type factorization

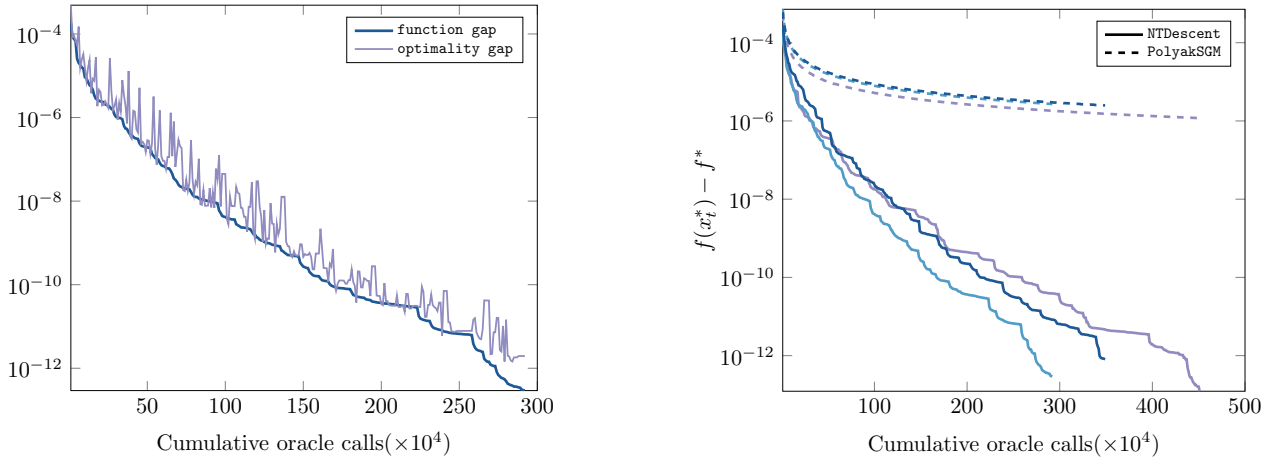
$$\min_{V \in \mathbb{R}^{N \times N}} f(V) = \tilde{f}(c(V)), \tag{7.2}$$

where $c \colon \mathbb{R}^{N \times N} \to \mathbb{S}^N$ satisfies $c(V) = \mathrm{Diag}([\mathrm{diag}(VV^\top)]^{-1/2})V$ for all $V \in \mathbb{R}^{N \times N}$. Here, the mapping $\mathrm{diag}(\cdot)$ takes a matrix an $N \times N$ matrix A to the N dimensional vector with ith entry A_{ii}. On the other hand, the mapping $\mathrm{Diag}(\cdot)$ takes an N dimensional vector v to the $N \times N$ diagonal matrix with ith diagonal entry v_i. A formula for the subgradient of f may be found [9]. We do not attempt to verify that f satisfies the full Assumption A. Instead, we point out that under a "transversality condition," function f admits an active manifold at local minimizers [31].

Turning to the experiment, we consider the case where $N = 14$ and $K = 7$. In this example, the optimal function value $\inf f$ is not known. Thus, we run `NTDescent` from four random initial starting points. We terminate each run of `NTDescent` when a certain "optimality gap" R_k satisfies $R_k \leq 10^{-12}$. We denote the minimal function value achieved across all four runs by f^*. Let us now define and motivate the optimality gap. For iteration k in Algorithm 4, define

$$R_k = \min\left\{\max\{\sigma_i^{(k)}, \|v_{i+1}^{(k)}\|^2\} \colon \sigma_i^{(k)} \leq \|v_{i+1}^{(k)}\|\right\},$$

where $\sigma_i^{(k)}$ and $v_{i+1}^{(k)}$ are computed in Lines 2 through 6 of Algorithm 3 at iteration k. Provided that x_k is sufficiently close to a point $\bar{x}$ at which function f satisfies Assumption A, it is possible to show that R_k satisfies $f(x_k) - f(\bar{x}) \lesssim R_k$. This is illustrated in Figure 5a: there the optimality gap closely tracks the estimated function gap, when approximating by $\inf f$ by f^*. In Figure 5b, we compare the performance of `NTDescent` on the three runs which did not achieve function value f^* before termination. In all three cases, we see similar performance. Next, for each run of `NTDescent`, we also run `PolyakSGM` from the same initial starting point, estimating $\inf f$ by f^*. We see that `NTDescent` outperforms `PolyakSGM`.



(a) (b)

Figure 5: Numerical performance on (7.2). Left: close relationship between "optimality gap" and function gap; Right: comparison of `PolyakSGM` and `NTDescent` from three initial starting points. For both algorithms the value $f(x_t^*)$ denotes the best function seen after t oracle evaluations. See text for detail.

References

[1] Kurt M Anstreicher and Jon Lee. A masked spectral bound for maximum-entropy sampling. In *mODa 7—Advances in Model-Oriented Design and Analysis*, pages 1–12. Springer, 2004.

[2] Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An ap-

proach based on the kurdyka-łojasiewicz inequality. *Mathematics of operations research*, 35(2):438–457, 2010.

[3] Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137(1):91–129, 2013.

[4] Pascal Bianchi, Walid Hachem, and Sholom Schechtman. Stochastic subgradient descent escapes active strict saddles. *arXiv preprint arXiv:2108.02072*, 2021.

[5] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1):459–494, 2014.

[6] Jérôme Bolte, Aris Daniilidis, Adrian Lewis, and Masahiro Shiota. Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization*, 18(2):556–572, 2007.

[7] Nicolas Boumal. An introduction to optimization on smooth manifolds. *Available online, Aug*, 2020.

[8] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[9] Samuel Burer and Jon Lee. Solving maximum-entropy sampling problems using factored masks. *Mathematical Programming*, 109(2):263–281, 2007.

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

[11] Francis H Clarke, Yuri S Ledyaev, Ronald J Stern, and Peter R Wolenski. *Nonsmooth analysis and control theory*, volume 178. Springer Science & Business Media, 2008.

[12] Damek Davis and Dmitriy Drusvyatskiy. Subgradient methods under weak convexity and tame geometry. *SIAG/OPT Views and News*, 28(1):1–10, 2020.

[13] Damek Davis and Dmitriy Drusvyatskiy. A gradient sampling method with complexity guarantees for general lipschitz functions. *arXiv preprint arXiv:2112.06969*, 2021.

[14] Damek Davis, Dmitriy Drusvyatskiy, and Vasileios Charisopoulos. Stochastic algorithms with geometric step decay converge linearly on sharp functions. *arXiv preprint arXiv:1907.09547*, 2019.

[15] Damek Davis, Dmitriy Drusvyatskiy, and Liwei Jiang. Subgradient methods near active manifolds: saddle point avoidance, local convergence, and asymptotic normality. *arXiv preprint arXiv:2108.11832*, 2021.

[16] Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D Lee. Stochastic subgradient method converges on tame functions. *Foundations of computational mathematics*, 20(1):119–154, 2020.

[17] Damek Davis, Dmitriy Drusvyatskiy, Yin Tat Lee, Swati Padmanabhan, and Guanghao Ye. A gradient sampling method with complexity guarantees for general lipschitz functions. *arXiv preprint arXiv:2112.06969*, 2022.

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

[19] Dmitriy Drusvyatskiy, Alexander D Ioffe, and Adrian S Lewis. Generic minimizing behavior in semialgebraic optimization. *SIAM Journal on Optimization*, 26(1):513–534, 2016.

[20] Dmitriy Drusvyatskiy and Adrian S Lewis. Optimality, identifiability, and sensitivity. *Mathematical Programming*, 147(1):467–498, 2014. Citations refer to long version arXiv:1207.6628.

[21] Paul J Enright and Bruce A Conway. Discrete approximations to optimal trajectories using direct transcription and nonlinear programming. *Journal of Guidance, Control, and Dynamics*, 15(4):994–1002, 1992.

[22] R Fletcher. A model algorithm for composite nondifferentiable optimization problems. In *Nondifferential and Variational Techniques in Optimization*, pages 67–76. Springer, 1982.

[23] AA Goldstein. Optimization of lipschitz continuous functions. *Mathematical Programming*, 13(1):14–22, 1977.

[24] XY Han and Adrian S Lewis. Survey descent: A multipoint generalization of gradient descent for nonsmooth optimization. *arXiv preprint arXiv:2111.15645*, 2021.

[25] A.D. Ioffe. An invitation to tame optimization. *SIAM J. Optim.*, 19(4):1894–1917, 2009.

[26] John M Lee. Smooth manifolds. In *Introduction to Smooth Manifolds*, pages 1–31. Springer, 2013.

[27] Claude Lemarechal. An extension of davidon methods to non differentiable problems. In *Nondifferentiable optimization*, pages 95–109. Springer, 1975.

[28] Claude Lemaréchal, François Oustry, and Claudia Sagastizábal. The $\mathcal{U}$-lagrangian of a convex function. *Transactions of the American mathematical Society*, 352(2):711–729, 2000.

[29] Adrian Lewis and Calvin Wylie. A simple newton method for local nonsmooth optimization. *arXiv preprint arXiv:1907.11742*, 2019.

[30] Adrian S Lewis. Active sets, nonsmoothness, and sensitivity. *SIAM Journal on Optimization*, 13(3):702–725, 2002.

[31] Adrian S Lewis and Michael L Overton. Nonsmooth optimization via quasi-newton methods. *Mathematical Programming*, 141(1):135–163, 2013.

[32] Robert Mifflin. Semismooth and semiconvex functions in constrained optimization. *SIAM J. Control Optim.*, 15(6):959–972, 1977.

[33] Robert Mifflin and Claudia Sagastizábal. A $\mathcal{VU}$-algorithm for convex minimization. *Mathematical Programming*, 104(2):583–608, 2005.

[34] Robert Mifflin and Claudia Sagastizábal. A VU-algorithm for convex minimization. *Mathematical programming*, 104(2):583–608, 2005.

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

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

[37] Welington de Oliveira and Claudia Sagastizábal. Bundle methods in the xxist century: A bird's-eye view. *Pesquisa Operacional*, 34:647–670, 2014.

[38] B. T. Polyak. Minimization of unsmooth functionals. *USSR Computational Mathematics and Mathematical Physics*, 9(3):14–29, 1969.

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

[40] Alexander Shapiro. On a class of nonsmooth composite functions. *Mathematics of Operations Research*, 28(4):677–692, 2003.

[41] Alexander Shapiro. On a class of nonsmooth composite functions. *Mathematics of Operations Research*, 28(4):677–692, 2003.

[42] Philip Wolfe. A method of conjugate subgradients for minimizing nondifferentiable functions. In *Nondifferentiable optimization*, pages 145–173. Springer, 1975.

[43] Stephen J Wright. Convergence of an inexact algorithm for composite nonsmooth optimization. *IMA journal of numerical analysis*, 10(3):299–321, 1990.

[44] Stephen J Wright. Identifiable surfaces in constrained optimization. *SIAM Journal on Control and Optimization*, 31(4):1063–1079, 1993.

[45] Yangyang Xu and Wotao Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on imaging sciences*, 6(3):1758–1789, 2013.

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

[47] Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Ali Jadbabaie, and Suvrit Sra. Complexity of finding stationary points of nonsmooth nonconvex functions. *arXiv preprint arXiv:2002.04130*, 2020.

A Proof of Lemma 2.2

Let g denote the minimal norm element of $\partial_\sigma f(x)$. Write g as a convex combination of subgradients: $g = \sum_{i=1}^n \lambda_i g_i$ where $\sum_{i=1}^n \lambda_i = 1$ and $g_i \in \partial f(x_i)$ for some $x_i \in B_\sigma(x)$ and $n > 0$. Then

$$\begin{aligned} f(x) &\leq \sum_{i=1}^n \lambda_i f(x_i) + L\sigma \\ &\leq f(y) + \sum_{i=1}^n \langle \lambda_i g_i, x_i - y \rangle + L\sigma \\ &\leq f(y) + \langle g, x - y \rangle + \sum_{i=1}^n \lambda_i \langle g_i, x_i - x \rangle + L\sigma \\ &\leq f(y) + \operatorname{dist}(0, \partial_\sigma f(x)) \|x - y\| + 2L\sigma, \end{aligned}$$

as desired.

B Proof of Proposition 3.5

First, since $\mathcal{M}$ is C^4 smooth, the projection $P_{\mathcal{M}}$ is C^3 smooth near $\bar{x}$. In addition, since f is C^3 smooth along $\mathcal{M}$ near $\bar{x}$, the composition $f_{\mathcal{M}} := f \circ P_{\mathcal{M}}$ is also C^3 smooth near $\bar{x}$. Note that $\mu > 0$ due to the active manifold assumption. Now, choose $\delta > 0$ small enough that the following hold:

1. $\nabla P_{\mathcal{M}}$ is $C_{\mathcal{M}}$-Lipschitz on $B_\delta(\bar{x})$;
2. $\nabla f_{\mathcal{M}}$ is β-Lipschitz on $B_\delta(\bar{x})$;
3. $\nabla^2 f_{\mathcal{M}}$ is ρ-Lipschitz on $B_\delta(\bar{x})$ in the operator norm, where $\rho := 2\mathrm{lip}^{\mathrm{op}}_{\nabla^2 f_{\mathcal{M}}}(\bar{x})$;
4. f is L-Lipschitz on $B_\delta(\bar{x})$;
5. the quadratic growth bound (A1) holds:
$$f(x) - f(\bar{x}) \geq \frac{\gamma}{2}\|x - \bar{x}\|^2 \qquad \text{for all } x \in \overline{B}_\delta(\bar{x});$$
6. the strong (a) bound (A3) holds:
$$\|P_{T_{\mathcal{M}}(y)}(v - \nabla_{\mathcal{M}} f(y))\| \leq C_{(a)}\|x - y\| \tag{B.1}$$
for all $x \in \overline{B}_\delta(\bar{x})$, $v \in \partial f(x)$, and $y \in \mathcal{M} \cap \overline{B}_\delta(\bar{x})$.
7. the $(b_\leq)$ regularity bound (A4) holds:
$$f(x') \geq f(x) + \langle v, x' - x \rangle - \frac{\mu}{2}\|x - \hat{x}\| \tag{B.2}$$
for all $x \in B_\delta(\bar{x})$, $v \in \partial f(x)$, and $x' \in B_\delta(\bar{x}) \cap \mathcal{M}$.

8. the sharpness condition holds:

$$\operatorname{dist}(0, \partial f(x)) > 2\mu \qquad \text{for all } x \in B_\delta(\bar{x}) \backslash \mathcal{M}.$$

Given these bounds, let us define

$$\delta_{\mathrm{A}} := \frac{1}{2} \min \left\{ \delta, \frac{9\gamma}{32\rho}, \frac{\mu}{2(C_{(a)} + 2\beta + 2C_{\mathcal{M}} L)} \right\}.$$

For this choice of δ_{A}, Item 1 holds automatically. We now prove the remaining items.

B.1 Item 2: Smoothness of $P_{\mathcal{M}}$.

Fix $x' \in B_{2\delta_{\mathrm{A}}}(\bar{x})$ and $x \in B_{\delta_{\mathrm{A}}}(x)$. Observe that $P_{\mathcal{M}}(x) \in B_{2\delta_{\mathrm{A}}}(\bar{x})$ and we have the inclusion $x - P_{\mathcal{M}}(x) \in N_{\mathcal{M}}(P_{\mathcal{M}}(x))$. Consequently, we have

1. $P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(x) = P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(P_{\mathcal{M}}(x))$;
2. $P_{\mathcal{M}}(x) = P_{\mathcal{M}}(P_{\mathcal{M}}(x))$;
3. $\nabla P_{\mathcal{M}}(P_{\mathcal{M}}(x)) = P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}$.

Therefore, we have

$$\begin{aligned}
&\|P_{\mathcal{M}}(x') - P_{\mathcal{M}}(x) - P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(x' - x)\| \\
&= \|P_{\mathcal{M}}(x') - P_{\mathcal{M}}(P_{\mathcal{M}}(x)) - \nabla P_{\mathcal{M}}(P_{\mathcal{M}}(x))(x' - P_{\mathcal{M}}(x))\| \\
&\leq \frac{C_{\mathcal{M}}}{2} \|x' - P_{\mathcal{M}}(x)\|^2 \\
&\leq C_{\mathcal{M}}(\|x' - x\|^2 + \operatorname{dist}^2(x, \mathcal{M})),
\end{aligned}$$

where the first inequality follows from Lipschitz continuity of $\nabla P_{\mathcal{M}}$ on $B_{2\delta_{\mathrm{A}}}(\bar{x}) \subseteq B_\delta(\bar{x})$.

B.2 Item 3: Bounds on $\nabla_{\mathcal{M}} f$

Recall that $P_{\mathcal{M}}(x) \in B_{2\delta_{\mathrm{A}}}(\bar{x})$ whenever $x \in B_{\delta_{\mathrm{A}}}(\bar{x})$. Thus, below we prove that

$$\frac{\gamma}{2} \|y - \bar{x}\| \leq \|\nabla f_{\mathcal{M}}(y)\| \leq \beta \|y - \bar{x}\| \qquad \text{for all } y \in B_{2\delta_{\mathrm{A}}}(\bar{x}) \cap \mathcal{M}.$$

This is equivalent to the claimed bound since $\nabla f_{\mathcal{M}}(y) = \nabla_{\mathcal{M}} f(y)$ for all $y \in B_{2\delta_{\mathrm{A}}}(\bar{x}) \cap \mathcal{M}$.

Let us first prove the claimed upper bound. Due to the inequality,

$$f_{\mathcal{M}}(x) - f_{\mathcal{M}}(\bar{x}) \geq \frac{\gamma}{2} \|P_{\mathcal{M}}(x) - \bar{x}\|^2 \qquad \text{for all } x \in B_\delta(\bar{x}),$$

it follows that $\bar{x}$ is a local minimizer of $f_{\mathcal{M}}$. Consequently, $\nabla f_{\mathcal{M}}(\bar{x}) = 0$. Thus, since β is a local Lipschitz constant of $\nabla f_{\mathcal{M}}$ on $B_\delta(\bar{x})$, we have

$$\|\nabla f_{\mathcal{M}}(y)\| \leq \beta \|y - \bar{x}\| \qquad \text{for all } y \in B_\delta(\bar{x}) \cap \mathcal{M}.$$

Since $2\delta_{\mathrm{A}} \le \delta$, this proves the claimed upper bound.

Next we prove the claimed lower bound. It suffices to establish the following convexity inequality:

$$f_{\mathcal{M}}(y) + \langle \nabla f_{\mathcal{M}}(y), \bar{x} - y \rangle \le f_{\mathcal{M}}(\bar{x}) \qquad \text{for all } y \in B_{2\delta_{\mathrm{A}}}(\bar{x}) \cap \mathcal{M}. \tag{B.3}$$

Indeed, if this inequality holds, we have

$$\langle \nabla f_{\mathcal{M}}(y), y - \bar{x} \rangle \ge f_{\mathcal{M}}(y) - f_{\mathcal{M}}(\bar{x}) \ge \frac{\gamma}{2}\|y - \bar{x}\|^2 \qquad \text{for all } y \in B_{2\delta_{\mathrm{A}}}(\bar{x}) \cap \mathcal{M},$$

and the desired result follows from Cauchy-Schwarz.

To that end, observe that since $\nabla f_{\mathcal{M}}(\bar{x}) = 0$ and $\nabla^2 f_{\mathcal{M}}$ is ρ-Lipschitz in $B_{2\delta_{\mathrm{A}}}(\bar{x})$, we have

$$f_{\mathcal{M}}(y) \le f_{\mathcal{M}}(\bar{x}) + \frac{1}{2}\langle \nabla^2 f_{\mathcal{M}}(\bar{x})(y - \bar{x}), (y - \bar{x}) \rangle + \frac{\rho}{6}\|y - \bar{x}\|^3 \qquad \text{for all } y \in B_{2\delta_{\mathrm{A}}}(\bar{x}).$$

Consequently, we have the lower bound on the quadratic form: for all $y \in B_{2\delta_{\mathrm{A}}}(\bar{x}) \cap \mathcal{M}$, we have

$$\begin{aligned}
\frac{1}{2}\langle \nabla^2 f_{\mathcal{M}}(\bar{x})(y - \bar{x}), (y - \bar{x}) \rangle &\ge f_{\mathcal{M}}(y) - f_{\mathcal{M}}(\bar{x}) - \frac{\rho}{6}\|y - \bar{x}\|^3 \\
&\ge \frac{\gamma}{2}\|y - \bar{x}\|^2 - \frac{\rho}{6}\|y - \bar{x}\|^3 \\
&\ge \frac{3\gamma}{8}\|y - \bar{x}\|^2,
\end{aligned} \tag{B.4}$$

where the second inequality follows from the quadratic growth bound and the third follows from the bound $\|y - \bar{x}\| \le 2\delta_{\mathrm{A}} \le \frac{3\gamma}{4\rho}$. Therefore, for all $y \in \mathcal{M} \cap B_{2\delta_{\mathrm{A}}}(\bar{x})$, we have

$$\begin{aligned}
f_{\mathcal{M}}(\bar{x}) &\ge f_{\mathcal{M}}(y) + \langle \nabla f_{\mathcal{M}}(y), \bar{x} - y \rangle + \frac{1}{2}\langle \nabla^2 f_{\mathcal{M}}(y)(\bar{x} - y), (\bar{x} - y) \rangle - \frac{\rho}{6}\|y - \bar{x}\|^3 \\
&\ge f_{\mathcal{M}}(y) + \langle \nabla f_{\mathcal{M}}(y), \bar{x} - y \rangle + \frac{1}{2}\langle \nabla^2 f_{\mathcal{M}}(\bar{x})(\bar{x} - y), (\bar{x} - y) \rangle - \frac{2\rho}{3}\|y - \bar{x}\|^3 \\
&\ge f_{\mathcal{M}}(y) + \langle \nabla f_{\mathcal{M}}(y), \bar{x} - y \rangle + \frac{3\gamma}{8}\|y - \bar{x}\|^2 - \frac{2\rho}{3}\|y - \bar{x}\|^3 \\
&\ge f_{\mathcal{M}}(y) + \langle \nabla f_{\mathcal{M}}(y), \bar{x} - y \rangle ,
\end{aligned}$$

where the first and second inequalities follow by Lipschitz continuity of $\nabla^2 f_{\mathcal{M}}$; the third inequality follows from (B.4); and the fourth inequality follows from the bound $\|y - \bar{x}\| \le 2\delta_{\mathrm{A}} \le \frac{9\gamma}{16\rho}$. This completes the proof.

B.3 Item 4: Consequences of strong (a)-regularity

Fix $x \in B_{\delta_{\mathrm{A}}}(\bar{x})$ and $\sigma \le \delta_{\mathrm{A}}$. Recall that $y := P_{\mathcal{M}}(x) \in B_{2\delta_{\mathrm{A}}}(\bar{x})$ since $x \in B_{\delta_{\mathrm{A}}}(\bar{x})$. Fix $g \in \partial_\sigma f(x)$. By definition of $\partial_\sigma f(x)$, there exists a family of coefficients $\lambda_i \in [0, 1]$, points $x_i \in \overline{B}_\sigma(x) \subseteq \overline{B}_\delta(\bar{x})$, and subgradients $g_i \in \partial f(x_i)$ indexed by a finite set $i \in I$ such that

$\sum_{i\in I}\lambda_i = 1$ and $g = \sum_{i\in I}\lambda_i g_i$. Therefore, by averaging the strong (a) bound (B.1) over g_i, we find that

$$\begin{aligned}\|P_{T_{\mathcal{M}}(y)}(g - \nabla_{\mathcal{M}} f(y))\| &\le \sum_{i\in I}\lambda_i \|P_{T_{\mathcal{M}}(y)}(g_i - \nabla_{\mathcal{M}} f(y))\| \\ &\le \sum_{i\in I}\lambda_i C_{(a)}\|x_i - y\|. \\ &\le C_{(a)}(\operatorname{dist}(x,\mathcal{M}) + \sigma).\end{aligned}$$

Since g was arbitrary, it follows that for all $x \in B_{\delta_{\mathrm{A}}}(\bar{x})$ and $\sigma \le \delta_{\mathrm{A}}$, we have

$$\sup_{g\in\partial_\sigma f(x)} \|P_{T_{\mathcal{M}}(y)}(g - \nabla_{\mathcal{M}} f(y))\| \le C_{(a)}(\operatorname{dist}(x,\mathcal{M}) + \sigma). \tag{B.5}$$

Now we apply this bound to establish the two remaining inequalities.

Indeed, first observe that for all $x \in B_{\delta_{\mathrm{A}}}(\bar{x})$ and $\sigma \le \delta_{\mathrm{A}}$, we have

$$\sup_{g\in\partial_\sigma f(x)} \|P_{T_{\mathcal{M}}(y)} g\| \le \|\nabla_{\mathcal{M}} f(y)\| + C_{(a)}(\operatorname{dist}(x,\mathcal{M}) + \sigma) \le \beta\|y - \bar{x}\| + C_{(a)}(\operatorname{dist}(x,\mathcal{M}) + \sigma),$$

where the first inequality follows from (B.5) and the second inequality follows from Item 3. This proves the first claimed bound. Second, observe that for all $x \in B_{\delta_{\mathrm{A}}}(\bar{x})$ and $\sigma \le \delta_{\mathrm{A}}$, we have

$$\begin{aligned}\sup_{g,g'\in\partial_\sigma f(x)} \|P_{T_{\mathcal{M}}(y)}(g - g')\| &\le \sup_{g\in\partial_\sigma f(x)} \|P_{T_{\mathcal{M}}(y)}(g - \nabla_{\mathcal{M}} f(y))\| + \sup_{g'\in\partial_\sigma f(x)} \|P_{T_{\mathcal{M}}(y)}(g' - \nabla_{\mathcal{M}} f(y))\| \\ &\le 2C_{(a)}(\operatorname{dist}(x,\mathcal{M}) + \sigma).\end{aligned}$$

where the second inequality follows from (B.5). This completes the proof.

B.4 Item 5: Aiming inequality

Consider a point $x \in B_{\delta_{\mathrm{A}}}(\bar{x})$, let $\kappa = 2\mu$, and define

$$\hat{x} \in \operatorname*{argmin}_{x'\in\overline{B}_{2\delta_{\mathrm{A}}}(\bar{x})} \{f(x') + \kappa\|x' - x\|\}.$$

We claim that $\hat{x} \in \mathcal{M} \cap B_{2\delta_{\mathrm{A}}}(\bar{x})$. Indeed, first note that by definition of $\hat{x}$ and the inclusion $\hat{x} \in \overline{B}_{2\delta_{\mathrm{A}}}(\bar{x})$, we have

$$\|\hat{x} - x\| \le \frac{f(\bar{x}) - f(\hat{x})}{\kappa} + \|\bar{x} - x\| \le \|\bar{x} - x\| < \delta_{\mathrm{A}},$$

where the second inequality follows since $\bar{x}$ is a minimizer of f on $B_{2\delta_{\mathrm{A}}}(\bar{x})$, a consequence of quadratic growth. Thus, by the triangle inequality, we have $\hat{x} \in B_{2\delta_{\mathrm{A}}}(\bar{x})$. By Fermat's rule, we therefore have the inclusion:

$$0 \in \partial(f + \kappa\|\cdot - x\|)(\hat{x}) \subseteq \partial f(\hat{x}) + \kappa\overline{B}.$$

If $\hat{x} \notin \mathcal{M}$, then $\operatorname{dist}(0, \partial f(\hat{x})) > \kappa$, contradicting the above inclusion. Therefore, we have $\hat{x} \in \mathcal{M} \cap B_{2\delta_{\mathrm{A}}}(\bar{x})$.

Turning to the aiming inequality, apply the $(b_{\leq})$-regularity bound B.2 to $\hat{x}$:

$$f(\hat{x}) \geq f(x) + \langle v, \hat{x} - x \rangle - \varepsilon \|x - \hat{x}\| \geq f(\hat{x}) + \langle v, \hat{x} - x \rangle + (\kappa - \varepsilon)\|x - \hat{x}\|,$$

where we define $\varepsilon := \mu/2$. Consequently, we have

$$\langle v, x - P_{\mathcal{M}}(x) \rangle \geq (\kappa - \varepsilon)\|x - \hat{x}\| + \langle v, \hat{x} - P_{\mathcal{M}}(x) \rangle \qquad \text{for all } v \in \partial f(x). \tag{B.6}$$

We now bound the term $\langle v, \hat{x} - P_{\mathcal{M}}(x) \rangle$: By the conclusion of Item 2, we have

$$\|P_{\mathcal{M}}(\hat{x}) - P_{\mathcal{M}}(x) - P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(\hat{x} - x)\| \leq C_{\mathcal{M}}(\|x - \hat{x}\|^2 + \operatorname{dist}^2(x, \mathcal{M})) \leq 2C_{\mathcal{M}}\|x - \hat{x}\|^2,$$

where the second inequality follows since $\hat{x} \in \mathcal{M}$. Thus, we have

$$\begin{aligned}
|\langle v, \hat{x} - P_{\mathcal{M}}(x) \rangle| &\leq |\langle v, P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}(\hat{x} - x) \rangle| + 2C_{\mathcal{M}}\|v\|\|x - \hat{x}\|^2 \\
&\leq \|P_{T_{\mathcal{M}}(P_{\mathcal{M}}(x))}v\|\|\hat{x} - x\| + 2C_{\mathcal{M}}L\|x - \hat{x}\|^2 \\
&\leq (C_{(a)}\operatorname{dist}(x, \mathcal{M}) + \beta\|P_{\mathcal{M}}(x) - \bar{x}\|)\|\hat{x} - x\| + 2C_{\mathcal{M}}L\|x - \hat{x}\|^2 \\
&\leq (C_{(a)}\delta_{\mathrm{A}} + 2\beta\delta_{\mathrm{A}} + 2C_{\mathcal{M}}L\delta_{\mathrm{A}})\|\hat{x} - x\| \\
&\leq \varepsilon\|\hat{x} - x\|.
\end{aligned}$$

where the second inequality follows from Item 4 and the third inequality follows from the inclusion $P_{\mathcal{M}}(x) \in B_{2\delta_{\mathrm{A}}}(\bar{x})$. Therefore, plugging this bound into (B.6), we arrive at

$$\langle v, x - P_{\mathcal{M}}(x) \rangle \geq (\kappa - 2\varepsilon)\|x - \hat{x}\| \geq \mu \operatorname{dist}(x, \mathcal{M}),$$

as desired.

B.5 Item 6: Bounding subgradients

Fix $x \in B_{\delta_{\mathrm{A}}}(\bar{x})$, $\sigma \leq \delta_{\mathrm{A}}$, and $g \in \partial_\sigma f(x)$. By definition of $\partial_\sigma f(x)$, there exists a family of coefficients $\lambda_i \in [0, 1]$, points $x_i \in \overline{B}_\sigma(x) \subseteq \overline{B}_\delta(\bar{x})$, and subgradients $g_i \in \partial f(x_i)$ indexed by a finite set $i \in I$ such that $\sum_{i \in I} \lambda_i = 1$ and $g = \sum_{i \in I} \lambda_i g_i$. Recall that by Lipschitz continuity of f on $B_\delta(\bar{x})$, we have $\|g_i\| \leq L$ for $i \in I$. Therefore,

$$\|g\| \leq \sum_{i \in I} \lambda_i \|g_i\| \leq L,$$

as desired.

B.6 Item 7: Bounding the function gap

Fix a point $x \in B_{\delta_{\mathrm{A}}}(\bar{x})$ and recall that $P_{\mathcal{M}}(x) \in B_{2\delta_{\mathrm{A}}}(\bar{x})$. Then by Lipschitz continuity of f on $B_\delta(\bar{x})$, we have

$$f(x) - f(P_{\mathcal{M}}(\bar{x})) \leq L \operatorname{dist}(x, \mathcal{M}).$$

Next, arguing as in the proof of Item 3, we find that $\nabla f_{\mathcal{M}}(\bar{x}) = 0$. Thus, since $\nabla f_{\mathcal{M}}$ is β-Lipschitz on $B_\delta(\bar{x})$, we have

$$f(P_{\mathcal{M}}(\bar{x}))-f(\bar{x}) = f_{\mathcal{M}}(P_{\mathcal{M}}(x))-f(\bar{x}) \leq \langle \nabla f_{\mathcal{M}}(\bar{x}), P_{\mathcal{M}}(x) - \bar{x}\rangle + \frac{\beta}{2}\|P_{\mathcal{M}}(x)-\bar{x}\|^2 = \frac{\beta}{2}\|P_{\mathcal{M}}(x)-\bar{x}\|^2.$$

Putting both bounds together, we have

$$f(x) - f(\bar{x}) = f(x) - f(P_{\mathcal{M}}(x)) + f(P_{\mathcal{M}}(x)) - f(\bar{x}) \leq L\mathrm{dist}(x, \mathcal{M}) + \frac{\beta}{2}\|P_{\mathcal{M}}(x) - \bar{x}\|^2,$$

as desired.

C Proof of Lemma 6.4

We remind the reader that Assumption A is in force. Consequently, by Item 1 of Proposition 3.5, we have:

$$f(x) - f(\bar{x}) \geq \frac{\gamma}{2}\|x - \bar{x}\|^2 \qquad \text{for all } x \in \overline{B}_{\delta_A}(\bar{x}).$$

For any $a > 0$, we claim that

$$\{x \in \mathbb{R}^d \colon f(x) - f(\bar{x}) \leq a\} \subseteq B_{r_a}(\bar{x}) \qquad \text{where } r_a := \max\left\{\frac{2a}{\gamma\delta_A}, \sqrt{\frac{2a}{\gamma}}\right\}.$$

Note that the inclusion follows from the following bound:

$$f(x) - f(\bar{x}) \geq \frac{\gamma}{2}\min\{\delta_A, \|x - \bar{x}\|\}\|x - \bar{x}\| \qquad \text{for all } x \in \mathbb{R}^d.$$

Thus, in the remainder of the proof, we prove this bound.

To that end, note that if $x \in B_{\delta_A}(\bar{x})$, the bound is immediate. Now fix $x \in \mathbb{R}^d \backslash B_{\delta_A}(\bar{x})$ and define the curve $x_t \colon t \mapsto (1-t)x + t\bar{x}$. Let $t_0 \in [0,1]$ to satisfy $x_{t_0} \in \mathrm{bdry}\, B_{\delta_A}(\bar{x})$. Then by Jensen's inequality, we have

$$(1-t_0)f(x) \geq f(x_{t_0})-tf(\bar{x}) \geq (1-t_0)f(\bar{x})+\frac{\gamma}{2}\|x_{t_0}-\bar{x}\|^2 = (1-t_0)f(\bar{x})+\frac{\gamma(1-t_0)}{2}\|x-\bar{x}\|\|x_{t_0}-\bar{x}\|.$$

Consequently, since $\|x_{t_0} - \bar{x}\| = \delta_A$, we have

$$f(x) - f(\bar{x}) \geq \frac{\gamma\delta_A}{2}\|x - \bar{x}\| \geq \frac{\gamma}{2}\min\{\delta_A, \|x - \bar{x}\|\}\|x - \bar{x}\|,$$

as desired. This completes the proof.