Lipschitz minimization and the Goldstein modulus

Siyu Kong A.S. Lewis [*]

May 21, 2024

Abstract

Goldstein's 1977 idealized iteration for minimizing a Lipschitz objective fixes a distance — the step size — and relies on a certain approximate subgradient. That "Goldstein subgradient" is the shortest convex combination of objective gradients at points within that distance of the current iterate. A recent implementable Goldstein-style algorithm allows a remarkable complexity analysis (Zhang et al. 2020), and a more sophisticated variant (Davis and Liang, 2022) leverages typical objective geometry to force near-linear convergence. To explore such methods, we introduce a new modulus, based on Goldstein subgradients, that robustly measures the slope of a Lipschitz function. We relate near-linear convergence of Goldstein-style methods to linear growth of this modulus at minimizers. We illustrate the idea computationally with a simple heuristic for Lipschitz minimization.

Key words: Lipschitz optimization, subgradient, linear convergence,

AMS Subject Classification: 90C56, 49J52, 65Y20

1 Introduction: the Goldstein subgradient

We consider a Euclidean space $\mathbf{X}$ with closed unit ball B, and a Lipschitz function $f\colon \mathbf{X} \to \mathbf{R}$ with Lipschitz constant $L \geq 0$. Following [3], the *Clarke subdifferential* of f at any point $x \in \mathbf{X}$, denoted $\partial f(x)$, is the closed convex hull of all those vectors of the form $\lim_{r\to\infty} \nabla f(x_r)$ that arise from sequences $x_r \to x$. The Clarke subdifferential is always nonempty, compact, and convex. We say that x is *Clarke-critical* when $0 \in \partial f(x)$. For any radius $\epsilon \geq 0$, following [8], the *Goldstein subdifferential* of f at x is the set

$$\partial_\epsilon f(x) \;=\; \mathrm{conv}\big(\partial f(x + \epsilon B)\big),$$

[*] ORIE, Cornell University, Ithaca, NY. people.orie.cornell.edu/aslewis
Research supported in part by National Science Foundation Grant DMS-2006990.

and it too is nonempty, compact, and convex. The associated *Goldstein subgradient*, denoted $g_\epsilon(x)$ is the shortest vector in $\partial_\epsilon f(x)$.

As observed in [8], whenever the Goldstein subgradient is nonzero, it satisfies a striking descent property: for any current point $x \in \mathbf{X}$ and radius $\epsilon \geq 0$, the *Goldstein update*

$$g = g_\epsilon(x) \qquad \text{and} \qquad x_{\text{new}} = x - \epsilon \frac{g}{|g|}, \tag{1.1}$$

ensures

$$f(x_{\text{new}}) \;\leq\; f(x) - \epsilon |g|. \tag{1.2}$$

Iterating this update thus results in a simple conceptual minimization procedure, using a step of constant size ϵ from one iterate to the next.

Practical approximations of this procedure confront two fundamental challenges.

- How should we choose the step size ϵ?
- How should we approximate the Goldstein subgradient $g_\epsilon(x)$?

With respect to the second challenge, the Goldstein subgradient is not computable in practice, except when the objective f is simply structured, such as piecewise linear, or piecewise linear-quadratic. However, approximation schemes have led to several recent advances. Using one such approach, [20] presents a randomized algorithm for general Lipschitz minimization that, given any $\epsilon > 0$, finds a point x satisfying $|g_\epsilon(x)| \leq \epsilon$ using no more than $O(\epsilon^{-4})$ calls to a specialized subgradient oracle. This algorithm is a remarkable accomplishment, notwithstanding the slow rate, inspiring several follow-up studies: [4, 11–14, 19]. All the more striking, then, is a recent breakthrough algorithm [5] whose convergence is "nearly linear" (a term on which we expand later). This subtle algorithm resolves both challenges above in an ingenious fashion.

Extensive earlier literature on linear convergence in nonsmooth optimization includes the seminal work [18] on ϵ-subgradient descent methods, and was surveyed recently in [1], with a particular focus on bundle methods. Those earlier works typically enumerate "serious" steps in procedures rather than all the subgradients needed to approximate the epsilon subdifferential for each serious step. Crucially, the near-linear convergence result of [5] simply counts all subgradients used.

Linear convergence results in nonsmooth optimization typically rely on growth conditions at minimizers. For ϵ-subgradient descent methods, [18] uses an upper Lipschitz condition on the objective subdifferential. For the Goldstein-style method of [5], the proof of near-linear convergence relies on a standard quadratic growth condition, along with several structural assumptions that are sophisticated, albeit generic for semi-algebraic objectives.

Our aim here is to isolate a simple growth condition that might underly near-linear convergence of Goldstein-style subgradient methods. We skirt the sophistication inherent in the algorithm of [5] by decoupling the two challenges above. We lay aside the challenge of approximating the Goldstein subgradient, imagining the following ideal oracle.

Oracle 1.3 (Goldstein subgradient)
Input: a point $x \in \mathbf{X}$ and a radius $\epsilon \geq 0$.
Output: the Goldstein subgradient $g_\epsilon(x)$.

We focus instead on the first challenge — choosing the step size ϵ. We identify a natural growth property of the objective f at its minimizer that, in conjunction with a simple step-size strategy, ensures nearly linear convergence for Goldstein's iteration. We furthermore verify the growth property for a representative class of nonsmooth objectives.

The conceptual optimization method we describe is simple but far from implementable. Nonetheless, we believe that the new growth condition underlying it may prove illuminating, both for variational analysts and for algorithm designers.

2 The Goldstein modulus

We begin by constructing a robust measure of the slope of f, starting with the following simple observation.

Proposition 2.1 *For any L-Lipschitz objective function $f \colon \mathbf{X} \to \mathbf{R}$, any point $x \in \mathbf{X}$, and any radius $\epsilon \geq 0$, the Goldstein subgradient satisfies $|g_\epsilon(x)| \leq L$.*

The Goldstein update (1.1) guarantees an objective decrease of $\epsilon|g_\epsilon(x)|$. If ϵ is small, then this guaranteed decrease is also small, being no larger than $L\epsilon$. On the other hand, the guaranteed decrease usually vanishes for large ϵ, because the ball $x + \epsilon B$ then contains a Clarke critical point. Choosing the step size ϵ thus requires a compromise.

To address this compromise, the following definition introduces a modulus that robustly controls the size of Goldstein subgradients. The growth condition in the definition is a central idea in our development. We argue that this condition often holds in nonsmooth optimization, and illustrate its potential in explaining the convergence rate of Goldstein-style algorithms.

Definition 2.2 For any Lipschitz function $f \colon \mathbf{X} \to \mathbf{R}$ and any point $x \in \mathbf{X}$, the *Goldsten modulus* is the value

$$\Gamma f(x) \;=\; \inf\{\epsilon \geq 0 : |g_\epsilon(x)| \leq \epsilon\}.$$

The Goldstein modulus *grows linearly* at a point $\bar{x} \in \mathbf{X}$ if there exists a constant $\alpha > 0$ such that

$$\Gamma f(x) \geq \alpha|x - \bar{x}| \qquad \text{for all } x \in \mathbf{X} \text{ near } \bar{x}. \tag{2.3}$$

We collect some elementary properties of the Goldstein modulus in the following result.

Proposition 2.4 *The Goldstein modulus for an L-Lipschitz function* $f\colon \mathbf{X} \to \mathbf{R}$ *at a point* $x \in \mathbf{X}$ *satisfies the following properties.*

(i) $0 \leq \Gamma f(x) \leq L$.

(ii)

$$|g_\epsilon(x)| \quad \begin{cases} > \epsilon & \big(\text{if } \epsilon < \Gamma f(x)\big) \\ \leq \epsilon & \big(\text{if } \epsilon = \Gamma f(x)\big) \\ < \epsilon & \big(\text{if } \epsilon > \Gamma f(x)\big) \end{cases}$$

(iii) $\Gamma f(x) = 0$ *if and only if* x *is Clarke critical.*

(iv) *If* $\bar{x}$ *is a Clarke critical point, then* $\Gamma f(x) \leq |x - \bar{x}|$.

(v) *For any positive radius* $\epsilon < \Gamma f(x)$, *the Goldstein update (1.1) ensures an objective decrease larger than* ϵ^2.

Proof Property (i) follows from Proposition 2.1. We next turning to property (ii).

For any value $\epsilon \in \big[0, \Gamma f(x)\big)$, by definition we have $|g_\epsilon(x)| > \epsilon$. Choose a sequence (ϵ_r) decreasing strictly to $\Gamma f(x)$. Let $m = 1 + \dim \mathbf{X}$. By definition we have $|g_{\epsilon_r}(x)| \leq \epsilon_r$ for each $r = 1, 2, 3, \ldots$, so there exist points $x_r^i \in x + \epsilon_r B$, subgradients $y_r^i \in \partial f(x_r^i) \subset LB$, and weights $\lambda_r^i \in [0, 1]$ for $i = 1, 2, \ldots, m$, satisfying

$$\sum_{i=1}^m \lambda_r^i = 1 \qquad \text{and} \qquad \sum_{i=1}^m \lambda_r^i y_r^i = g_{\epsilon_r}(x).$$

After taking a subsequence, we can suppose that, for each $i = 1, 2, \ldots, m$, as $r \to \infty$ the points x_r^i converge to some vector $x^i \in x + \Gamma f(x)B$, the subgradients y_r^i converge to some vector $y^i \in LB$, and the weights λ_r^i converge to some weight $\lambda^i \in [0, 1]$. We deduce $\sum_i \lambda^i = 1$, and the vector $y = \sum_i \lambda^i y^i$ satisfies $|y| \leq \Gamma f(x)$. Since the Clarke subdifferential ∂f has closed graph, we know $y^i \in \partial f(x^i)$ for each i, so $y \in \partial_{\Gamma f(x)} f(x)$. We have thus proved

$$|g_{\Gamma f(x)}(x)| \;\leq\; \Gamma f(x).$$

Finally, for any value $\epsilon > \Gamma f(x)$ we have $\partial_\epsilon f(x) \supset \partial_{\Gamma f(x)}(x)$, and hence

$$|g_\epsilon(x)| \;\leq\; |g_{\Gamma f(x)}(x)| \;\leq\; \Gamma f(x) \;<\; \epsilon.$$

Property (ii) follows.

From property (ii) we deduce $\Gamma f(x) = 0$ if and only if $|g_0(x)| = 0$, or equivalently $g_0(x) = 0$, which amounts the point x being Clarke critical. Property (iv) follows from the observation $0 \in \partial_{|x-\bar{x}|} f(x)$. Property (v) follows from inequality (1.2). □

We return to the question of how to choose the radius ϵ for the Goldstein update (1.1). By property (v) in Proposition 2.4, any choice of radius in the interval

$$\Big[\frac{1}{2}\Gamma f(x)\, ,\, \Gamma f(x)\Big)$$

ensures an objective decrease of at least $\frac{1}{4}\big(\Gamma f(x)\big)^2$. Furthermore, we can use Oracle 1.3 to find such a radius by bisection search, as follows.

Algorithm 2.5 (Goldstein modulus approximation)

```
  input: point $x \in \mathbf{X}$, Lipschitz constant $L > 0$
  if $g_0(x) = 0$ then
    return 0
  end if
  $\epsilon = \frac{1}{2}L$
  while $|g_\epsilon(x)| \le \epsilon$ do
    $\epsilon = \frac{1}{2}\epsilon$
  end while
  return $\epsilon$
```

Proposition 2.6 *For any L-Lipschitz function $f\colon \mathbf{X} \to \mathbf{R}$ and any point $x \in \mathbf{X}$, Algorithm 2.5 returns a radius*

$$\epsilon \;\in\; \Big[\frac{1}{2}\Gamma f(x)\, ,\, \Gamma f(x)\Big). \tag{2.7}$$

The number of oracles calls is 1 *if x is Clarke critical, and otherwise is*

$$2 + \big\lfloor \log L - \log\big(\Gamma f(x)\big)\big\rfloor. \tag{2.8}$$

Proof If x is Clarke critical, then the first oracle call finds $g_0(x) = 0$, and the algorithm returns the value 0. We therefore turn to the case when x is not Clarke critical.

After $1 + m$ oracle calls, for $m = 1, 2, 3, \ldots$, we have $\epsilon = L2^{-m}$. The algorithm terminates as soon as $|g_\epsilon(x)| > \epsilon$, or in other words, by Proposition 2.4, as soon as $\epsilon < \Gamma f(x)$. This latter condition is equivalent to $L2^{-m} < \Gamma f(x)$, or equivalently $m > \log L - \log\big(\Gamma f(x)\big)$. The smallest such integer m is

$$\bar{m} \;=\; 1 + \big\lfloor \log L - \log\big(\Gamma f(x)\big)\big\rfloor$$

and formula (2.8) follows. At termination, we know $|g_\epsilon(x)| > \epsilon$. We furthermore know $|g_{2\epsilon}(x)| \leq 2\epsilon$: if $\bar{m} > 1$, this follows from the condition in the while loop for the previous value of ϵ, and if $\bar{m} = 1$ it follows from the Proposition 2.1. Property (2.7) follows. $\square$

3 A simple Goldstein descent algorithm

By combining the Goldstein descent iteration (1.1) with the modulus approximation procedure, Algorithm 2.5, we arrive at the following simple algorithm.

Algorithm 3.1 (Lipschitz minimization)

```
input: initial point $x \in \mathbf{X}$, Lipschitz constant $L > 0$, maximum iterations $\bar{s}$
$s = 0$                                   {counts Goldstein subgradient evaluations}
loop
  $\epsilon = \frac{1}{2}L$
  $g = g_\epsilon(x)$
  $s = s + 1$
  while $|g| \leq \epsilon$ do
    $\epsilon = \frac{1}{2}\epsilon$                                {bisection search}
    $g = g_\epsilon(x)$
    $s = s + 1$
    if $s > \bar{s}$ then
      return $x$
    end if
  end while
  $x = x - \epsilon\frac{g}{|g|}$
end loop
```

To study the complexity of Algorithm 3.1, we consider how the Goldstein modulus typically grows at a local minimizer. That growth is at most linear, by Proposition 2.4(iv); the following result assumes that it also satisfies a lower linear bound.

Theorem 3.2 (Convergence rate) *Consider an L-Lipschitz objective function $f \colon \mathbf{X} \to \mathbf{R}$ and a minimizer $\bar{x} \in \mathbf{X}$ at which the Goldstein modulus grows linearly. Then, starting from any point $x_0 \in \mathbf{X}$ with sufficiently small initial gap $f(x_0) - f(\bar{x})$, after s subgradient calls to Oracle 1.3, the current point x in Algorithm 3.1 satisfies*

$$f(x) - f(\bar{x}) \;=\; O\Big(\frac{\log s}{s}\Big) \qquad \text{as } s \to \infty,$$

where the implicit constant depends only the Lipschitz constant L, the initial gap, and the Goldstein modulus linear growth constant α in inequality (2.3).

Proof Denote the current point x after $k = 0, 1, 2 \ldots$ bisections searches by x_k. The descent property (1.2) guarantees

$$f(x_k) - f(x_{k+1}) \;\geq\; \frac{1}{4}\big(\Gamma f(x_k)\big)^2.$$

Define the objective gap $\gamma_k = f(x_k) - f(\bar{x})$. The Lipschitz and linear growth conditions imply $\Gamma f(x_k) \geq \frac{\alpha \gamma_k}{L}$, and hence

$$\gamma_k - \gamma_{k+1} \;>\; \Big(\frac{\alpha \gamma_k}{2L}\Big)^2.$$

With the change of variable $\hat{\gamma} = (\frac{\alpha}{2L})^2 \gamma$, we deduce $\hat{\gamma}_k - \hat{\gamma}_{k+1} > \hat{\gamma}_k^2$. Hence

$$\frac{1}{\hat{\gamma}_{k+1}} \;>\; \frac{1}{\hat{\gamma}_k} + \frac{1}{1 - \hat{\gamma}_k} \;>\; \frac{1}{\hat{\gamma}_k} + 1.$$

By induction, we deduce

$$\frac{1}{\hat{\gamma}_k} \;>\; \frac{1}{\hat{\gamma}_0} + k.$$

Summarizing, for some constant β, depending only on the linear growth constant α, the Lipschitz constant L, and the initial gap γ_0, we have bounded the nondecreasing sequence of gaps γ_k in terms of the number of bisection searches k, by

$$\gamma_k \;\leq\; \frac{\beta}{k+1}.$$

Turning to the number of oracle calls, we also have

$$\Gamma f(x_k) \;\geq\; \alpha |x_k - \bar{x}| \;\geq\; \frac{\alpha}{L} \gamma_k.$$

By Proposition 2.6, the bisection search at x_k takes a number of oracle calls not exceeding

$$2 + \log L - \log\big(\Gamma f(x_k)\big) \;\leq\; 2 + \log L - \log\Big(\frac{\alpha}{L}\gamma_k\Big) \;=\; \kappa - \log \gamma_k,$$

for some constant $\kappa > 0$ depending only on α and L. The total number s_k of oracle calls accumulated before updating x_k to x_{k+1} therefore satisfies

$$s_k \;\leq\; \sum_{j=0}^{k} (\kappa - \log \gamma_j) \;\leq\; (k+1)(\kappa - \log \gamma_k) \;\leq\; \frac{\beta(\kappa - \log \gamma_k)}{\gamma_k}.$$

The strictly increasing continuous function $\phi \colon (0, 2^\kappa) \to (0, +\infty)$ defined by

$$\phi(t) \;=\; \frac{t}{\beta(\kappa - \log t)}.$$

has a strictly increasing continuous inverse $\phi^{-1}\colon (0,+\infty) \to (0, 2^{\kappa})$. Notice

$$\lim_{s\to+\infty} \frac{s}{\log s}\phi^{-1}\Big(\frac{1}{s}\Big) \;=\; -\lim_{t\downarrow 0} \frac{t}{\phi(t)\log\phi(t)} \;=\; -\lim_{t\downarrow 0} \frac{\beta(\kappa-\log t)}{\log t - \log(\beta(\kappa - \log t))} \;=\; \beta.$$

We deduce

$$\phi^{-1}\Big(\frac{1}{s}\Big) \;=\; O\Big(\frac{\log s}{s}\Big),$$

where the constant depends only on β. Since we have proved $\phi(\gamma_k) \le \frac{1}{s_k}$ for all $k = 0,1,2,\ldots$, we deduce $\gamma_k = O(\frac{\log s_k}{s_k})$, and the result follows. □

The idealized Goldstein descent complexity result in Theorem 3.2 is illuminating for two reasons. First, it highlights how much stronger our imagined Oracle 1.3 is than a standard subgradient oracle. For comparison, in terms of the number s of calls to standard oracles, the method of [20] for nonsmooth nonconvex objectives has complexity $O(s^{-1/4})$. For convex objectives, the usual subgradient method has complexity $O(s^{-1/2})$, which improves to $O(\frac{1}{s})$ only in the strongly convex or smooth cases [9].

The second and more important reason that we present Theorem 3.2 is to highlight quite simply the impact on complexity of a single assumption: linear growth in the Goldstein modulus. We will argue that this assumption often holds, even in the absence of strong convexity or smoothness. For now, we simply illustrate the linear growth condition with three simple examples.

Example 3.3 Consider three functions, defined, for a constant $\alpha > 0$ and points $x \in \mathbf{X}$, by

$$f_1(x) = \frac{1}{2}\alpha|x|^2, \qquad f_2(x) = \alpha|x|, \qquad f_3(x) = \frac{1}{4}\alpha|x|^4.$$

Each function has a unique minimizer at the point 0, and the functions f_1 and f_2 grow at least quadratically there, but f_3 does not. The corresponding Goldstein modululi for the first two functions, namely

$$\Gamma f_1(x) = \frac{\alpha}{1+\alpha}|x|, \qquad \Gamma f_2(x) = \min\{|x|, \alpha\},$$

both grow linearly at 0, but the third modulus satisfies

$$\Gamma f_3(x) \;\le\; \alpha|x|^3,$$

so its growth is slower than linear.

4 Nearly linear convergence

We turn next from the simple sublinear rate guarantee in Theorem 3.2 to our main focus, which is linear convergence. For many classical first-order algorithms, linear convergence is associated with quadratic growth, as described in the following definition. This property is indispensable, both in this section, and in the remainder of this work.

Definition 4.1 The function $f\colon \mathbf{X} \to \mathbf{R}$ *grows quadratically* at the point $\bar{x}$ when there exists a constant $\delta > 0$ such that

$$f(x) - f(\bar{x}) \;\geq\; \frac{\delta}{2}|x - \bar{x}|^2 \qquad \text{for all } x \in \mathbf{X} \text{ near } \bar{x}.$$

The relationship between linear convergence and quadratic growth is explored in detail in [7].

Two distinct avenues to proving linear convergence suggest themselves. We might focus on the iterates themselves, proving the existence of a constant $\theta \in (0,1)$ such that for all $r = 0, 1, 2, \ldots$,

$$(4.2) \qquad |x_{r+1} - \bar{x}| \;\leq\; \theta|x_r - \bar{x}|.$$

Alternatively, we might consider instead the objective values, instead proving

$$(4.3) \qquad f(x_{r+1}) - f(\bar{x}) \;\leq\; \theta\big(f(x_r) - f(\bar{x})\big).$$

Instead of linear convergence, suppose that we are prepared to accept a somewhat relaxed rate. Following [5], we say that $x_r \to \bar{x}$ *nearly linearly* if, for some exponent $m > 0$, ensuring an error $|x_r - \bar{x}|$ less than any small tolerance $\delta > 0$ requires only $r = O\big((\log \frac{1}{\delta})^m\big)$ iterations. (Linear convergence corresponds to the case $m = 1$.) We prove next that if each iteration satisfies *either* inequality (4.2) *or* inequality (4.3), then $x_r \to \bar{x}$ nearly linearly.

Theorem 4.4 (Near-linear convergence) *Consider a locally Lipschitz function $f\colon \mathbf{X} \to \mathbf{R}$ that grows quadratically at a point $\bar{x} \in \mathbf{X}$, and a sequence of points (x_r) in X with initial point x_0 near $\bar{x}$. For some constant $\theta \in (0,1)$, suppose that successive points in the sequence always satisfy $f(x_{r+1}) \leq f(x_r)$ and one of the inequalities (4.2) and (4.3). Then there exist constants $\alpha, \beta > 0$ such that*

$$|x_r - \bar{x}| \;\leq\; \alpha e^{-\beta\sqrt{r}}$$

for all r.

Proof To simplify notation, we suppose $\bar{x} = 0$ and $f(0) = 0$. For sufficiently small initial points x_0, we know that there exist constants $\delta, L > 0$ such that each point x_r satisfies

$$\frac{\delta}{2}|x_r|^2 \leq f(x_r) \leq L|x_r|.$$

For any nonnegative integers r, s, if $f(x_{r+s}) > \theta f(x_r)$, then

$$\frac{\theta}{L} f(x_r) \ < \ \frac{1}{L} f(x_{r+s}) \ \leq \ |x_{r+s}| \ \leq \ \theta^s |x_r| \ \leq \ \theta^s \sqrt{\frac{2f(x_r)}{\delta}}$$

so

$$\theta^s \ > \ \frac{\theta\sqrt{\delta}}{L\sqrt{2}} \sqrt{f(x_r)},$$

and hence

$$s < \mu - \nu \log\big(f(x_r)\big)$$

for suitable constants $\mu, \nu > 0$. We deduce

$$f(x_{r+s}) \leq \theta f(x_r) \qquad \text{for } s = \big\lceil \mu - \nu \log\big(f(x_r)\big) \big\rceil.$$

Arguing inductively, we see, for $t = 0, 1, 2, \ldots$,

$$f(x_r) \leq \theta^t f(x_0)$$

for

$$r \ \geq \ \sum_{u=0}^{t-1} \big\lceil \mu - \nu \log\big(\theta^u f(x_0)\big) \big\rceil$$

and hence in particular for

$$\begin{aligned} r \ &\geq \ \sum_{u=0}^{t-1} \Big(1 + \mu - \nu(u \log\theta + \log\big(f(x_0)\big)\Big) \\ &= \ t(1 + \mu - \nu \log\big(f(x_0)\big) \ - \ \frac{\nu}{2} t(t-1) \log\theta. \end{aligned}$$

Thus there exists an integer $\psi > 0$ such that, for all $t = 0, 1, 2 \ldots$,

$$f(x_r) \ \leq \ f(x_{\psi t^2}) \ \leq \ f(x_0)\theta^t$$

for any integer $r \geq \psi t^2$. We deduce

$$\frac{\delta}{2}|x_r|^2 \ \leq \ f(x_r) \ \leq \ f(x_0)\theta^{\left\lfloor \sqrt{\frac{1}{\psi} r} \right\rfloor}$$

and hence

$$2 \log|x_r| \ \leq \ \log\Big(\frac{2f(x_0)}{\delta}\Big) \ + \ \Big(\sqrt{\frac{1}{\psi} r} - 1\Big) \log\theta,$$

from which the result follows. □

To capture the two possible ways to control linear convergence, (4.2) and (4.3), in the context of the Goldstein iteration (1.1), we make the following definition.

Definition 4.5 A locally Lipschitz function $f\colon \mathbf{X} \to \mathbf{R}$ has *the Goldstein property* at a point $\bar{x} \in \mathbf{X}$ if there exist constants $\nu > \mu > 0$ and $\theta \in (0,1)$ such that, for any point $x \neq \bar{x}$ near $\bar{x}$, and any step size ϵ satisfying

$$\mu \le \frac{\epsilon}{|x-\bar{x}|} \le \nu,$$

the Goldstein subgradient $g = g_\epsilon(x)$ is nonzero, and the point $x^+ = x - \epsilon\frac{g}{|g|}$ satisfies

$$\begin{array}{lrcl} \text{either} & f(x^+) - f(\bar{x}) & \le & \theta\big(f(x) - f(\bar{x})\big) \\ \text{or} & |x^+ - \bar{x}| & \le & \theta|x-\bar{x}|. \end{array}$$

We call the interval $[\mu,\nu]$ a *proportionality bracket.*

Corollary 4.6 *Consider a locally Lipschitz function $f\colon \mathbf{X} \to \mathbf{R}$ that grows quadratically and has the Goldstein property at point $\bar{x} \in \mathbf{X}$, with proportionality bracket $[\mu,\nu]$. Then any sequence of points generated iteratively from an initial point near $\bar{x}$ and updating iterates $x \neq \bar{x}$ according to the rule*

$$x \;\leftarrow\; x - \epsilon\frac{g}{|g|} \qquad \text{for any } \epsilon \text{ satisfying } \mu \le \frac{\epsilon}{|x-\bar{x}|} \le \nu \text{ and } g = g_\epsilon(x)$$

converges nearly linearly to $\bar{x}$ in the sense of Theorem 4.4.

Even assuming access to Oracle 1.3 for Goldstein subgradients, the updating rule in Corollary 4.6 is not realistic in general because we do not know the distance between the current iterate x and the minimizer $\bar{x}$. Our strategy will be to estimate that distance from the Goldstein modulus, using the linear growth property.

5 Robust growth relative to a manifold

To verify linear growth of the Goldstein modulus for a locally Lipschitz function $f\colon \mathbf{X} \to \mathbf{R}$ at a point $\bar{x} \in \mathbf{X}$, we will rely on two further conditions, each of which concern some distinguished set $\mathcal{M} \subset \mathbf{X}$. The first is a type of Lipschitz condition on the subdifferential ∂f.

Definition 5.1 The subdifferential ∂f is *upper Lipschitz* relative to a set $\mathcal{M} \subset \mathbf{X}$ at a point $\bar{x} \in \mathcal{M}$ when there exists a constant $K > 0$ such that all points x, y near $\bar{x}$ with $x \in \mathcal{M}$ satisfy

$$\partial f(y) \;\subset\; \partial f(x) + K|x-y|B.$$

The second property, which plays an important role in [5], describes how the value $f(x)$ grows as the point $x \in \mathbf{X}$ moves away from the set $\mathcal{M}$. It assumes in particular that the set $\mathcal{M}$ is $\mathcal{C}^2$-smooth manifold around $\bar{x}$. In that case, every point x near $\bar{x}$ has a unique nearest point in $\mathcal{M}$, which we denote $P_{\mathcal{M}}(x)$.

Definition 5.2 If a set $\mathcal{M} \subset \mathbf{X}$ is a $\mathcal{C}^2$-smooth manifold around a point $\bar{x} \in \mathcal{M}$, then a locally Lipschitz function $f\colon \mathbf{X} \to \mathbf{R}$ satisfies the *aiming condition* relative to $\mathcal{M}$ at $\bar{x}$ when there exists a constant $\mu > 0$ such that all points $x \in \mathbf{X}$ near $\bar{x}$ and subgradients $v \in \partial f(x)$ satisfy

$$\langle v, x - P_{\mathcal{M}}(x) \rangle \;\geq\; \mu d_{\mathcal{M}}(x).$$

Some standard variational-analytic terminology [3] helps illuminate the aiming condition. We therefore pause to review this language.

Consider a locally Lipschitz function $f\colon \mathbf{X} \to \mathbf{R}$, and a point $x \in \mathbf{X}$. The *Clarke directional derivative* of f at x in a direction $y \in \mathbf{X}$ is the quantity

$$f^{\circ}(x; y) \;=\; \max_{v \in \partial f(x)} \langle v, y \rangle.$$

The function f is *subdifferentially regular* at x when the Clarke and classical directional derivatives agree in every direction:

$$\lim_{t \downarrow 0} \frac{1}{t}\big(f(x+ty) - f(x)\big) \;=\; f^{\circ}(x; y) \qquad \text{for all } y \in \mathbf{X}.$$

For example, sums of smooth and continuous convex functions are everywhere subdifferentially regular. The *slope* of f at x is the quantity

$$|\nabla f|(x) \;=\; \limsup_{y \to x} \frac{f(x) - f(y)}{|x - y|},$$

unless x is a local minimizer, in which case the slope is zero. The following result is well known [10, Proposition 8.5].

Proposition 5.3 (Slope and subgradients) *At every point $x \in \mathbf{X}$, the slope of a locally Lipschitz function $f\colon \mathbf{X} \to \mathbf{R}$ satisfies the inequality*

$$|\nabla f|(x) \;\geq\; \min |\partial f(x)|,$$

with equality if f is subdifferentially regular at x.

With this terminology in hand, we return to the aiming condition. Definition 5.2 states that, at points x outside the manifold $\mathcal{M}$ but near the point $\bar{x}$, the Clarke directional derivative of the function f in the unit direction from x towards its nearest point in $\mathcal{M}$ is uniformly negative:

$$f^\circ\Big(x\,;\ \frac{1}{d_{\mathcal{M}}(x)}(P_{\mathcal{M}}(x)-x)\Big) \ \le\ -\mu.$$

We can understand the aiming condition further by considering both the direction and the norm of subgradients $v \in \partial f(x)$. First, the angle between v and the direction from x to its nearest point $P_{\mathcal{M}}(x)$ is uniformly larger than $\frac{\pi}{2}$: this is precisely the "aiming" behavior from which the property derives its name. Secondly, the subgradients v are uniformly bounded away from zero:

$$\liminf_{x\to\bar{x},\ x\notin\mathcal{M}} |\nabla f|(x) \ >\ 0, \tag{5.4}$$

In [15], property (5.4) is called *identifiability* of the set $\mathcal{M}$ for the function f at the point $\bar{x}$. To summarize, if the aiming condition holds for f relative to $\mathcal{M}$ at $\bar{x} \in \mathcal{M}$, then $\mathcal{M}$ is identifiable at $\bar{x}$ for f and subgradients at nearby points outside $\mathcal{M}$ aim uniformly away from $\mathcal{M}$.

At any point $x \in \mathbf{X}$ and for any radius $\epsilon \ge 0$, the slope and the Goldstein subgradient clearly satisfy

$$|\nabla f|(x) \ \ge\ |g_\epsilon(x)|,$$

by Proposition 5.3. With this in mind, we note that the aiming condition in fact implies a stronger property than identifiability, captured in the following crucial tool [5, Lemma 4.1].

Lemma 5.5 *The aiming condition, Definition 5.2, implies the existence of a constant $\gamma > 0$ such that*

$$\liminf_{x\to\bar{x},\ x\notin\mathcal{M}} |g_{\gamma d_{\mathcal{M}}(x)}(x)| \ >\ 0.$$

We next combine the conditions we need for linear growth of the Goldstein modulus in the following property.

Definition 5.6 If a set $\mathcal{M} \subset \mathbf{X}$ is a $\mathcal{C}^2$-smooth manifold around a point $\bar{x} \in \mathcal{M}$, then a locally Lipschitz function $f\colon \mathbf{X} \to \mathbf{R}$ *grows robustly* relative to $\mathcal{M}$ at $\bar{x}$ when the following properties hold.

- f grows quadratically at $\bar{x}$ (Definition 4.1).
- f is subdifferentially regular throughout $\mathcal{M}$.
- The restriction $f_{\mathcal{M}}\colon \mathcal{M} \to \mathbf{R}$ is $\mathcal{C}^2$-smooth.
- The subdifferential ∂f is upper Lipschitz relative to $\mathcal{M}$ at $\bar{x}$ (Definition 5.1).

- f satisfies the aiming condition relative to $\mathcal{M}$ at $\bar{x}$ (Definition 5.2).

Two simple examples are worth keeping in mind.

Example 5.7 (Strongly convex functions) If the function f is $\mathcal{C}^2$-smooth and strongly convex then it grows robustly at a minimizer relative to the whole space $\mathbf{X}$.

Example 5.8 (The norm) The norm grows robustly at 0 relative to the manifold $\{0\}$.

As a first step towards proving linear growth of the Goldstein modulus, we first observe an easier version: linear growth of the slope.

Proposition 5.9 (Local linear growth of slope) *Consider a set $\mathcal{M} \subset \mathbf{X}$ that is a $\mathcal{C}^2$-smooth manifold around a point $\bar{x} \in \mathcal{M}$, and a locally Lipschitz function $f \colon \mathbf{X} \to \mathbf{R}$ that grows robustly relative to $\mathcal{M}$ at $\bar{x}$. Then, at $\bar{x}$, the slope of f grows linearly: there exists a constant $\beta > 0$ such that*

$$|\nabla f|(x) \;\geq\; \beta|x - \bar{x}| \qquad \text{for all } x \in \mathbf{X} \text{ near } \bar{x}. \tag{5.10}$$

Proof The $\mathcal{C}^2$-smooth restriction $f_{\mathcal{M}} \colon \mathcal{M} \to \mathbf{R}$ grows quadratically at its local minimizer $\bar{x}$, so it is strongly convex on a neighborhood of $\bar{x}$ [2, Theorem 11.21], and hence, by [2, Lemma 11.28], its Riemannian gradient grows linearly at $\bar{x}$. In other words, using the notation of [2], there exists a constant $\beta > 0$ such that all points $x \in \mathcal{M}$ near $\bar{x}$ satisfy

$$\|\mathrm{grad}(f_{\mathcal{M}})\|_x \;\geq\; \beta|x - \bar{x}|$$

and consequently

$$|\nabla f|(x) \;\geq\; |\nabla f_{\mathcal{M}}|(x) \;=\; \|\mathrm{grad}(f_{\mathcal{M}})\|_x \;\geq\; \beta|x - \bar{x}|.$$

Combined with the identifiability condition (5.4), we deduce local linear growth of the slope. □

We can strengthen this result, as follows.

Theorem 5.11 *If a locally Lipschitz function $f \colon \mathbf{X} \to \mathbf{R}$ grows robustly at a point, then its Goldstein modulus grows linearly there.*

Proof To simplify notation, suppose that the point of interest is $\bar{x} = 0$. We argue by contradiction. If the result fails, then there exists a sequence of values $0 < \alpha_r \downarrow 0$ and a sequence of nonzero points $x_r \to 0$ in $\mathbf{X}$ such that $\Gamma f(x_r) < \alpha_r |x_r|$ for each $r = 1, 2, \ldots$. Each corresponding Goldstein subgradient must then satisfy $|g_{\alpha_r|x_r|}(x_r)| < \alpha_r|x_r|$. After taking a subsequence, we can suppose that the normalized vectors $\frac{x_r}{|x_r|}$ converge to some unit direction $u \in \mathbf{X}$.

Following the notation of Lemma 5.5 and Definitions 5.1 and 5.6, suppose that the direction u lies outside the tangent space T to the manifold $\mathcal{M}$ at 0. As $r \to \infty$, we have $x_r = |x_r|u + o(|x_r|)$, and hence $d_{\mathcal{M}}(x_r) = d_T(u)|x_r| + o(|x_r|)$. For all large r we deduce $\gamma d_{\mathcal{M}}(x_r) > \alpha_r|x_r|$ and hence

$$|g_{\gamma d_{\mathcal{M}}(x_r)}(x_r)| \;\le\; |g_{\alpha_r|x_r|}(x_r)| \;<\; \alpha_r|x_r| \;\to\; 0 \qquad \text{as } r \to \infty,$$

contradicting Lemma 5.5.

The direction u must therefore lies in the tangent space T, so there exists a sequence of points $x'_r \in \mathcal{M}$ satisfying $x'_r - |x_r|u = o(|x_r|)$ and hence $x_r - x'_r = o(|x_r|)$ as $r \to \infty$. The upper-Lipschitz property ensures

$$\begin{aligned} g_{\alpha_r|x_r|}(x_r) \;&\in\; \partial_{\alpha_r|x_r|}f(x_r) \;=\; \mathrm{conv}\big(\partial f(x_r + \alpha_r|x_r|B)\big) \\ &\subset\; \partial f(x'_r) \;+\; \big(K\alpha_r|x_r| + o(|x_r|)\big)B, \end{aligned}$$

so there exist subgradients $y_r \in \partial f(x'_r)$ satisfying

$$|g_{\alpha_r|x_r|}(x_r) - y_r| \;\le\; K\alpha_r|x_r| + o(|x_r|).$$

The linear growth property (5.10) along with the subdifferential regularity of f at x'_r shows $|y_r| \ge |\nabla f|(x'_r) \ge \beta|x'_r|$, so we deduce

$$\alpha_r|x_r| \;>\; |g_{\alpha_r|x_r|}(x_r)| \;\ge\; \beta|x'_r| - K\alpha_r|x_r| + o(|x_r|) \;=\; \beta|x_r| - K\alpha_r|x_r| + o(|x_r|,$$

which is a contradiction for large r. □

6 Robust growth for max functions

Classical nonlinear programming furnishes a central example of robust growth. We consider *smooth max functions* $f\colon \mathbf{X} \to \mathbf{R}$, by which we mean functions having a representation of the form

$$f(x) \;=\; \max_{i \in I} f_i(x) \qquad (x \in \mathbf{X}) \tag{6.1}$$

for some family of continuously differentiable functions $f_i\colon \mathbf{X} \to \mathbf{R}$ indexed by i in some finite index set I. Under reasonable conditions, we will show that smooth max functions grow robustly. We will furthermore develop a Goldstein-style descent method that converges nearly linearly for such objectives.

At any point $x \in \mathbf{X}$, the subdifferential of the function (6.1) is given by

$$\partial f(x) \;=\; \mathrm{conv}\{\nabla f_i(x) : f_i(x) = f(x)\} \tag{6.2}$$

(see [3]). The standard second-order sufficient conditions in nonlinear programming motivate the following definition.

Definition 6.3 A function $f\colon \mathbf{X} \to \mathbf{R}$ is a *strong* $\mathcal{C}^2$ *max function* at a point $\bar{x} \in \mathbf{X}$ if the following conditions hold.

- The point $\bar{x}$ is Clarke critical and *nondegenerate*: $0 \in \mathrm{ri}\big(\partial f(\bar{x})\big)$.
- The function f grows quadratically at $\bar{x}$ (Definition 4.1).
- For some finite set I and $\mathcal{C}^2$-smooth functions $f_i\colon \mathbf{X} \to \mathbf{R}$ (for $i \in I$),
$$f(x) \;=\; \max_{i \in I} f_i(x) \qquad \text{for all } x \text{ near } \bar{x},$$
and furthermore the values $f_i(\bar{x})$ (for $i \in I$) are all equal, and the gradients $\nabla f_i(\bar{x})$ (for $i \in I$) are affinely independent.

In that case, we call the set of those points $x \in \mathbf{X}$ where the values $f_i(x)$ (for $i \in I$) are all equal the *active manifold*.

The various ingredients of Definition 6.3 correspond exactly to standard second-order sufficient conditions for the nonlinear program

$$\inf_{x \in \mathbf{X},\; t \in \mathbf{R}} \{t : f_i(x) \le t \text{ for all } i \in I\},$$

as presented in [17], for example. Denote the active manifold by $\mathcal{M}$. The classical first-order necessary optimality condition requires the existence of a Lagrange multiplier vector $\lambda \in \mathbf{R}^I_+$ satisfying $\sum_i \lambda_i = 1$ and $\sum_i \lambda_i \nabla f_i(\bar{x}) = 0$: exactly the condition that $\bar{x}$ is Clarke critical. The affine independence condition in the definition is just the usual linear independence constraint qualification, which ensures that λ is unique, and furthermore that the set $\mathcal{M}$ is a $\mathcal{C}^2$-smooth manifold around $\bar{x}$, with tangent space

$$T \;=\; \{\nabla f_i(\bar{x}) - \nabla f_j(\bar{x}) : i \neq j\}^{\perp}.$$

Nondegeneracy reduces to the condition $\lambda_i > 0$ for all $i \in I$, which is the classical strict complementarity condition. The classical theory of second-order sufficient conditions shows that this condition is equivalent to positive-definiteness of the operator $\sum_i \lambda_i \nabla^2 f_i(\bar{x})$ on the subspace T.

In Definition 6.3, the active manifold is well defined in the following sense. Although it depends on the functions f_i involved in the representation of the function f, the active manifold is identifiable for f at $\bar{x}$, in the sense of inequality (5.4), and identifiable manifolds must be locally unique around $\bar{x}$: see [6].

Theorem 6.4 *If $f\colon \mathbf{X} \to \mathbf{R}$ is a strong $\mathcal{C}^2$ max function at a point $\bar{x} \in \mathbf{X}$, with active manifold $\mathcal{M}$, then f grows robustly relative to $\mathcal{M}$ at $\bar{x}$, and hence its Goldstein modulus grows linearly at $\bar{x}$.*

Proof Robust growth is proved in [5], and the result then follows. □

The analogous result holds for any function f satisfying [5, Assumption A].

Definition 6.3 implies in particular that the restriction $f_{\mathcal{M}} \colon \mathcal{M} \to \mathbf{R}$ is $\mathcal{C}^2$ smooth around the point $\bar{x}$. At any nearby point $x \in \mathcal{M}$, we can identify the Riemannian gradient of $f_{\mathcal{M}}$ with a vector $\nabla f_{\mathcal{M}}(x)$ in the tangent space $T_{\mathcal{M}}(x)$. This vector has the following property.

Proposition 6.5 *If $f \colon \mathbf{X} \to \mathbf{R}$ is a strong $\mathcal{C}^2$ max function at a point $\bar{x} \in \mathbf{X}$, with active manifold $\mathcal{M}$, then for all points $x \in \mathcal{M}$ near $\bar{x}$, the Riemannian gradient $\nabla f_{\mathcal{M}}(x)$ is the shortest convex combination of the set of gradients $\{\nabla f_i(x) : i \in I\}$.*

Proof Consider any point $x \in \mathcal{M}$ near $\bar{x}$. For all $i \in I$ we know $f = f_i$ throughout the manifold $\mathcal{M}$, the vector $\nabla f_i(x) - \nabla f_{\mathcal{M}}(x)$ must lie in the normal space $N_{\mathcal{M}}(x)$. Equation (6.2) implies

$$\partial f(x) \;=\; \mathrm{conv}\{\nabla f_i(x) : i \in I\} \;\subset\; \nabla f_{\mathcal{M}}(x) + N_{\mathcal{M}}(x).$$

Furthermore, a partial smoothness argument [16, Proposition 4.3] shows $\nabla f_{\mathcal{M}}(x) \in \partial f(x)$. Since $\nabla f_{\mathcal{M}}(x)$ is the shortest vector in the affine subspace $\nabla f_{\mathcal{M}}(x) + N_{\mathcal{M}}(x)$, it must also be the shortest vector in $\partial f(x)$, so the result follows. □

The following tool is useful in what follows.

Proposition 6.6 *Consider the map $\Lambda \colon \mathbf{X}^k \to \mathbf{X}$ that maps any list of k vectors to its shortest convex combination. Then, around any affinely independent list whose image lies in the relative interior of its convex hull, the map Λ is smooth.*

Proof We map any list $v = (v^1, v^2, \ldots, v^k) \in \mathbf{X}^k$ to the shortest convex combination of the vectors in the list. Denote the given affinely independent list by $\bar{v}$. Then the relative interior assumption ensures $\Lambda(\bar{v}) = \sum_i \bar{\lambda}_i \bar{v}^i$ for some vector $\bar{\lambda} > 0$ solving the optimization problem

$$\min_{\lambda \in \mathbf{R}^k_+} \Big\{ \frac{1}{2} \Big| \sum_i \lambda_i \bar{v}^i \Big|^2 : \sum_i \lambda_i = 1 \Big\}$$

Since $\bar{\lambda} > 0$, convexity implies that $\bar{\lambda}$ also solves the problem

$$\min_{\lambda \in \mathbf{R}^k} \Big\{ \frac{1}{2} \Big| \sum_i \lambda_i v^i \Big|^2 : \sum_i \lambda_i = 1 \Big\}$$

when $v = \bar{v}$. The solutions of this latter problem are characterized by the linear system in $(\alpha, \lambda) \in \mathbf{R} \times \mathbf{R}^k$

$$(6.7) \qquad \left\{ \begin{array}{rcl} \alpha \;+\; \sum_i \langle v^i, v^j \rangle \lambda_i & = & 0 \qquad (j = 1, 2, \ldots, k) \\ \sum_i \lambda_i & = & 1. \end{array} \right.$$

When $v = \bar{v}$, this square system is invertible, because

$$\begin{array}{rcll} \alpha \ + \ \sum_i \langle \bar{v}^i, \bar{v}^j \rangle \lambda_i & = & 0 & \qquad (j = 1, 2, \ldots, k) \\ \sum_j \lambda_j & = & 0 & \end{array}$$

implies

$$0 \ = \ \sum_i \sum_j \langle \bar{v}^i, \bar{v}^j \rangle \lambda_i \lambda_j \ = \ \Big| \sum_i \lambda_i \bar{v}^i \Big|^2$$

so $\sum_i \lambda_i \bar{v}^i = 0$ and hence $\alpha = 0$, and furthermore, by affine independence, $\lambda = 0$. We deduce that the solution (α, λ) depends smoothly on the list v around $v = \bar{v}$, and furthermore satisfies $\lambda > 0$, since $\bar{\lambda} > 0$. Consequently we know $\Lambda(v) = \sum_i \lambda_i v^i$ also depends smoothly on v. □

We end this section by proving that small Goldstein subgradients of strong $\mathcal{C}^2$ max functions must approximate Riemannian gradients on the active manifold.

Theorem 6.8 *Consider a strong $\mathcal{C}^2$ max function f at a local minimizer $\bar{x}$ with active manifold $\mathcal{M}$. Then there exists a constant $\mu > 0$ such that all Goldstein subgradients at points $x \in \mathbf{X}$ near $\bar{x}$ for small radii $\epsilon \geq 0$ satisfy*

$$|g_\epsilon(x)| \leq \mu \qquad \Rightarrow \qquad \left\{ \begin{array}{rcl} x & = & P_{\mathcal{M}}(x) + O(\epsilon) \qquad \textit{and} \\ g_\epsilon(x) & = & \nabla_{\mathcal{M}} f\big(P_{\mathcal{M}}(x)\big) + O(\epsilon). \end{array} \right.$$

Proof Using the terminology of Definition 6.3, for each $i \in I$, consider the set

$$X_i \ = \ \{x \in \mathbf{X} : f_i(x) = f(x)\}.$$

The active manifold $\mathcal{M}$ is just $\cap_i X_i$. The affine independence assumption and a standard metric regularity argument shows the existence of a constant $C > 0$ such that

$$d_{\mathcal{M}}(x) \ \leq \ C \max_i d_{X_i}(x)$$

for all points $x \in \mathbf{X}$ near $\bar{x}$.

For each $j \in I$, denote by μ_j the distance from zero to the set

$$\text{conv}\{\nabla f_i(\bar{x}) : i \neq j\}, \tag{6.9}$$

which is strictly positive by Definition 6.3. Fix any constant μ in the interval $(0, \min_i \mu_i)$, and consider any point x satisfying $|g_\epsilon(x)| \leq \mu$.

We first claim that, if the point x is near $\bar{x}$ and the radius $\epsilon \geq 0$ is small, then $d_{X_j}(x) \leq \epsilon$ for all j, and hence $d_{\mathcal{M}}(x) \leq C\epsilon$. To see this, we argue by contradiction. If the claim fails, then there exists a sequence of points $x_r \to \bar{x}$ in $\mathbf{X}$ and radii $\epsilon_r \downarrow 0$, and elements $j_r \in I$, such that $|g_{\epsilon_r}(x_r)| \leq \mu$ and $d_{X_{j_r}}(x_r) > \epsilon_r$ for all $r = 1, 2, 3, \ldots$. After taking a subsequence, we can suppose that some element $j \in I$ satisfies $j_r = j$

for all r. Using Caratheodory's theorem, for $m = \dim \mathbf{X} + 1$, there exist scalars $\lambda_{ik}^r \geq 0$ and points $y_{ik}^r \in x_r + \epsilon_r B$ for all $i \neq j$ in I, $k = 1, 2, \ldots, m$, such that

$$g_{\epsilon_r}(x_r) \;=\; \sum_{i \neq j} \sum_{k=1}^{m} \lambda_{ik}^r \nabla f_i(y_{ik}^r) \quad \text{and} \quad \sum_{i \neq j} \sum_{k=1}^{m} \lambda_{ik}^r \;=\; 1 \quad \text{for all } r = 1, 2, 3, \ldots.$$

Taking another subsequence, we can suppose the existence of the limits $\lambda_{ik} = \lim_r \lambda_{ik}^r \in [0, 1]$ for each $i \neq j$ and $k = 1, 2, \ldots, m$, so some limit point $\hat{g}$ of the Goldstein subgradients $g_{\epsilon_r}(x_r)$ has the form

$$\hat{g} \;=\; \sum_{i \neq j} \sum_{k=1}^{m} \lambda_{ik} \nabla f_i(\bar{x}), \qquad \text{where} \qquad \sum_{i \neq j} \sum_{k=1}^{m} \lambda_{ik} \;=\; 1.$$

We deduce that $\hat{g}$ lies in the set (6.9), and yet $|\hat{g}| \leq \mu$, contradicting the definition of μ. We have thus proved our claim.

Assuming that the point x is near $\bar{x}$ and the radius $\epsilon \geq 0$ is small, we now know $d_{X_i}(x) \leq \epsilon$ for all $i \in I$, so there exists a point $y_i \in (x + \epsilon B) \cap X_i$. There exist scalars $\lambda_{ik} \geq 0$ and points $y_{ik} \in (x + \epsilon B) \cap X_i$ for $i \in I$ and $k = 1, 2, \ldots, m$, such that

$$g_\epsilon(x) \;=\; \sum_{i \in I} \sum_{k=1}^{m} \lambda_{ik} \nabla f_i(y_{ik}) \quad \text{and} \quad \sum_{i \in I} \sum_{k=1}^{m} \lambda_{ik} \;=\; 1.$$

For convenience, we can suppose $y_{ik} = y_i$ whenever $\lambda_{ik} = 0$. The nonnegative scalars $\lambda_i = \sum_k \lambda_{ik}$, for $i \in I$, sum to one. Define

$$g_i \;=\; \begin{cases} \frac{1}{\lambda_i} \sum_{k=1}^{m} \lambda_{ik} \nabla f_i(y_{ik}) & (\lambda_i > 0) \\ \\ \nabla f_i(y_i) & (\lambda_i = 0). \end{cases}$$

For each $i \in I$, we also know

$$g_i \;\in\; \mathrm{conv}\big(\nabla f_i(x + \epsilon B)\big) \;\subset\; \nabla f_i(x) + \epsilon L B,$$

for a suitable Lipschitz constant $L > 0$ for the gradients ∇f_i around $\bar{x}$. We have

$$g_\epsilon(x) \;=\; \sum_{i \in I} \lambda_i g_i \;\in\; \mathrm{conv}\{g_i : i \in I\} \;\subset\; \mathrm{conv}\{\nabla f_i(y_{ik}) : i \in I,\; k = 1, 2, \ldots, m\}$$

From its definition, $g_\epsilon(x)$ must therefore be the shortest convex combination of the vectors $\nabla f_i(y_{ik})$, so it must also be the shortest convex combination of the vectors g_i, for $i \in I$.

On the other hand, by Proposition 6.5, the Riemannian gradient $\nabla f_{\mathcal{M}}\big(P_{\mathcal{M}}(x)\big)$ is the shortest convex combination of the gradients $\nabla f_i(P_{\mathcal{M}}(x)\big)$, for $i \in I$. Furthermore, for each $i \in I$ we have

$$\begin{aligned} |\nabla f_i(P_{\mathcal{M}}(x)\big) - g_i| \;&\leq\; |\nabla f_i(P_{\mathcal{M}}(x)\big) - \nabla f_i(x)| + |\nabla f_i(x) - g_i| \\ &\leq\; L d_{\mathcal{M}}(x) + \epsilon L \;\leq\; L(1 + C)\epsilon. \end{aligned}$$

By assumption, when $x = \bar{x}$, the Riemannian gradient $\nabla f_{\mathcal{M}}\big(P_{\mathcal{M}}(x)\big) = \nabla f_{\mathcal{M}}(\bar{x})$ is zero, which lies in the relative interior of the convex hull of the corresponding gradients $\nabla f_i(P_{\mathcal{M}}(x)\big) = \nabla f_i(\bar{x})$. The result therefore now follows by Proposition 6.6. □

7 Tempered growth relative to a manifold

Returning to our general theme, given a function $f \colon \mathbf{X} \to \mathbf{R}$ and a minimizer $\bar{x}$, we are interested in the behavior of Goldstein subgradients $g_\epsilon(x)$ for nearby points x and radii $\epsilon > 0$. Suppose that f grows robustly relative to a manifold $\mathcal{M}$ at $\bar{x}$, so the aiming condition holds. Lemma 5.5 then guarantees that, in the "small radius" regime when the radius ϵ is small compared with the distance to the manifold $d_{\mathcal{M}}(x)$, the Goldstein subgradient cannot be too small. Consequently, the Goldstein update (1.1) ensures a reasonable decrease in the objective value. We now consider, by contrast, the "large radius" regime, where the radius is of the same order of magnitude as the distance to the minimizer $\bar{x}$.

Theorem 6.8 demonstrated, for strong $\mathcal{C}^2$ max functions, how small Goldstein subgradients of the function f at points x must approximate Riemannian gradients of the restriction $f_{\mathcal{M}}$ at the corresponding nearest points on the manifold $\mathcal{M}$. We crystallize this general behavior in the following definition.

Definition 7.1 Consider a set $\mathcal{M} \subset \mathbf{X}$ that is a $\mathcal{C}^2$-smooth manifold around a point $\bar{x} \in \mathcal{M}$, and a locally Lipschitz function $f \colon \mathbf{X} \to \mathbf{R}$ whose restriction $f|_{\mathcal{M}}$ is $\mathcal{C}^2$-smooth. We say that f has *tempered growth* at $\bar{x}$ relative to $\mathcal{M}$ if, given any angle $\theta > 0$, for any sufficiently small $\beta > 0$ and sequences of points $\bar{x} \neq x_r \to \bar{x}$ and radii $\epsilon_r \leq \beta|x_r - \bar{x}|$ with corresponding Goldstein subgradients $g_{\epsilon_r}(x_r)$ converging to zero, the subgradients and corresponding Riemannian gradients $\nabla_{\mathcal{M}} f\big(P_{\mathcal{M}}(x_r)\big)$ are eventually nonzero and subtend an angle less than θ.

Theorem 7.2 *Strong $\mathcal{C}^2$ max functions have tempered growth.*

Proof Consider a strong $\mathcal{C}^2$ max function f as in Definition 6.3. By Theorem 6.8, there exist constants C and D such that the projections $x'_r = P_{\mathcal{M}}(x_r)$ satisfy

$$\begin{aligned} |x_r - x'_r| &\leq C\epsilon_r \leq \beta C|x_r - \bar{x}| \\ |g_{\epsilon_r}(x_r) - \nabla_{\mathcal{M}} f(x'_r)| &\leq D\epsilon_r \leq \beta D|x_r - \bar{x}| \end{aligned}$$

for all large r. Quadratic growth (Definition 4.1) ensures

$$\begin{aligned} |\nabla_{\mathcal{M}} f(x'_r)| &\geq \delta|x'_r - \bar{x}| \\ &\geq \delta(|x_r - \bar{x}| - |x_r - x'_r|) \geq \delta(1 - \beta C)(|x_r - \bar{x}|). \end{aligned}$$

We deduce

$$\frac{|g_{\epsilon_r}(x_r) - \nabla_{\mathcal{M}} f(x'_r)|}{|\nabla_{\mathcal{M}} f(x'_r)|} \;\leq\; \frac{\beta D}{\delta(1-\beta C)} \;<\; \sin\theta,$$

providing that β is sufficiently small. □

Theorem 7.3 *Consider a set $\mathcal{M} \subset \mathbf{X}$ that is a $\mathcal{C}^2$-smooth manifold around a point $\bar{x} \in \mathcal{M}$, and a locally Lipschitz function $f\colon \mathbf{X} \to \mathbf{R}$ that grows robustly and has tempered growth at $\bar{x}$ relative to $\mathcal{M}$. Then f has the Goldstein property at $\bar{x}$, and indeed, given any constant $\gamma \in (0,1)$, the interval $[\beta\gamma, \beta]$ is a proportionality bracket for all sufficiently small $\beta > 0$. Consequently, any sequence of points generated iteratively from an initial point near $\bar{x}$ and updating iterates $x \neq \bar{x}$ according to the rule*

$$x \;\leftarrow\; x - \epsilon\frac{g}{|g|} \qquad \textit{for any } \epsilon \textit{ satisfying } \gamma \leq \frac{\epsilon}{\beta|x-\bar{x}|} \leq 1 \textit{ and } g = g_\epsilon(x)$$

converges nearly linearly to $\bar{x}$ in the sense of Theorem 4.4.

Proof By way of contradiction, consider any small constant $\beta > 0$, and suppose the property in Definition 7.1 fails. Then there exists sequences of values $0 < \mu_r \to 0$, points $\bar{x} \neq x_r \to \bar{x}$, radii ϵ_r satisfying

$$\gamma \;\leq\; \frac{\epsilon_r}{\beta|x_r - \bar{x}|} \;\leq\; 1$$

and Goldstein subgradients $g_r = g_{\epsilon_r}(x_r)$, such that either $g_r = 0$ or the updates

$$x_r^+ \;=\; x_r - \epsilon_r \frac{g_r}{|g_r|}$$

satisfy

$$\begin{aligned} f(x_r^+) \;&>\; f(x_r) - L\mu_r|x_r - \bar{x}| \\ |x_r^+ - \bar{x}| \;&>\; (1-\mu_r)|x_r - \bar{x}|. \end{aligned}$$

Taking a subsequence, we can suppose

$$\frac{\epsilon_r}{|x_r - \bar{x}|} \;\to\; \text{some } \alpha \in (0, \beta]. \tag{7.4}$$

If β is small, so is α, in which case g_r cannot be zero infinitely often, by the definition of tempered growth. Taking a subsequence, we can therefore suppose each g_r is nonzero.

By inequality (1.2) we have

$$f(x_r^+) \;\leq\; f(x_r) - \epsilon_r|g_r|,$$

so $L\mu_r|x_r - \bar{x}| > \epsilon_r|g_r|$ and hence by property (7.4) we deduce $g_r \to 0$. By Lemma 5.5, there exists a constant $\gamma > 0$ such that $\epsilon_r > \gamma d_{\mathcal{M}}(x_r)$ for all large r, so each projection $x'_r = P_{\mathcal{M}}(x_r)$ satisfies $|x_r - x'_r| < \frac{1}{\gamma}\epsilon_r$, and the definition of tempered growth now ensures that each corresponding Riemannian gradient $g'_r = \nabla_{\mathcal{M}} f(x'_r)$ is nonzero and, with g_r, subtends an arbitrarily small angle: specifically, assuming the quadratic growth condition (4.1), then for all sufficiently small β, we can guarantee

$$\Big|\frac{g_r}{|g_r|} - \frac{g'_r}{|g'_r|}\Big| \;\le\; \frac{\delta}{3}.$$

Quadratic growth guarantees

$$\Big\langle \frac{x'_r - \bar{x}}{|x'_r - \bar{x}|}\,,\, \frac{g'_r}{|g'_r|}\Big\rangle \;\ge\; \frac{2\delta}{3},$$

so

$$\Big\langle \frac{x'_r - \bar{x}}{|x'_r - \bar{x}|}\,,\, \frac{g_r}{|g_r|}\Big\rangle \;\ge\; \frac{\delta}{3}.$$

Now notice

$$\begin{aligned}
(1-\mu_r)^2|x_r - \bar{x}|^2 \;&<\; |x_r^+ - \bar{x}|^2 \\
&=\; \Big|x_r - \epsilon_r\frac{g_r}{|g_r|} - \bar{x}\Big|^2 \\
&=\; |x_r - \bar{x}|^2 + \epsilon_r^2 - 2\epsilon_r\Big\langle x_r - \bar{x}\,,\, \frac{g_r}{|g_r|}\Big\rangle \\
&\le\; |x_r - \bar{x}|^2 + \epsilon_r^2 + 2\epsilon_r|x_r - x'_r| - 2\epsilon_r\Big\langle x'_r - \bar{x}\,,\, \frac{g_r}{|g_r|}\Big\rangle \\
&\le\; |x_r - \bar{x}|^2 + \epsilon_r^2 + \frac{2}{\gamma}\epsilon_r^2 - 2\epsilon_r|x'_r - \bar{x}|\Big\langle \frac{x'_r - \bar{x}}{|x'_r - \bar{x}|}\,,\, \frac{g_r}{|g_r|}\Big\rangle \\
&\le\; |x_r - \bar{x}|^2 + \Big(1 + \frac{2}{\gamma}\Big)\epsilon_r^2 - \frac{2\delta}{3}\epsilon_r|x'_r - \bar{x}| \\
&\le\; |x_r - \bar{x}|^2 + \Big(1 + \frac{2}{\gamma}\Big)\epsilon_r^2 - \frac{2\delta}{3}\epsilon_r(|x_r - \bar{x}| - |x_r - x'_r|) \\
&\le\; |x_r - \bar{x}|^2 + \Big(1 + \frac{2}{\gamma} + \frac{2\delta}{3\gamma}\Big)\epsilon_r^2 - \frac{2\delta}{3}\epsilon_r|x_r - \bar{x}|.
\end{aligned}$$

Dividing both sides by $|x_r - \bar{x}|^2$ and letting $r \to \infty$ shows

$$1 \;\le\; 1 + \Big(1 + \frac{2}{\gamma} + \frac{2\delta}{3\gamma}\Big)\alpha^2 - \frac{2\delta}{3}\alpha,$$

which is a contradiction for all sufficiently small positive β (and hence α). □

The following algorithm realizes the near linear convergence property in Theorem 7.3 by approximating the distance to the minimizer using the Goldstein modulus.

Algorithm 7.5 (Minimization for Lipschitz f)

```
  input: Lipschitz constant $L$, initial point $x \in \mathbf{X}$, multiplier $\beta > 0$
  for iteration $= 1, 2, 3, \ldots$ do
    $\epsilon = \frac{1}{2}L$
    while $|g_\epsilon(x)| \leq \epsilon$ do
      $\epsilon = \frac{1}{2}\epsilon$
    end while
    $\epsilon = \beta\epsilon$
    $g = g_\epsilon(x)$
    $x = x - \epsilon\frac{g}{|g|}$
  end for
```

Theorem 7.6 *With the assumptions of Theorem 7.3, for any sufficiently small multiplier $\beta > 0$, if the initial point x is sufficiently close to the minimizer $\bar{x}$, then Algorithm 7.5 converges nearly linearly to $\bar{x}$ in the sense of Theorem 4.4.*

Proof After each `while` loop, setting $\epsilon = \beta\epsilon$ ensures that the radius satisfies

$$\frac{1}{2}\Gamma f(x) \;\leq\; \frac{\epsilon}{\beta} \;<\; \Gamma f(x),$$

by Proposition 2.6. By Theorem 5.11, the Goldstein modulus grows linearly: for some constant $\alpha > 0$, we know

$$\alpha|x - \bar{x}| \;\leq\; \Gamma f(x) \leq |x - \bar{x}|$$

where the second inequality follows from the fact that $\bar{x}$ is Clarke critical. We deduce

$$\frac{\alpha}{2} \;\leq\; \frac{\epsilon}{\beta|x - \bar{x}|} \;<\; 1.$$

Providing that β is sufficiently small, the result now follows from Theorem 7.3. $\square$

8 A Goldstein-style heuristic

In practice, we cannot implement Algorithm 7.5 to minimize a Lipschitz function $f\colon \mathbf{X} \to \mathbf{R}$, because, given a point $x \in \mathbf{X}$ and a radius $\epsilon > 0$, we cannot usually compute the Goldstein subgradient $g_\epsilon(x)$. To explore the effectiveness of the underlying idea — adjusting the radius ϵ adaptively by estimating the Goldstein modulus — we therefore resort to approximating the Goldstein subgradient $g = g_\epsilon(x)$, using a simple, easily implementable heuristic.

Our approach is guided by the fundamental descent property that we noted at the outset:

$$g = g_\epsilon(x) \neq 0 \quad \Rightarrow \quad f\Big(x - \frac{\epsilon}{|g|}g\Big) \;\leq\; f(x) - \epsilon|g|.$$

The following definition relaxes that property.

Definition 8.1 Consider a locally Lipschitz function $f \colon \mathbf{X} \to \mathbf{R}$, a point $x \in \mathbf{X}$, and a radius $\epsilon > 0$ An *approximate Goldstein subgradient* of f at x is a subgradient $g \in \partial_\epsilon f(x)$ satisfying the following property:

$$|g| \geq \epsilon \quad \Rightarrow \quad f\Big(x - \frac{\epsilon}{|g|} g\Big) \ < \ f(x) - \frac{\epsilon |g|}{2}.$$

We can compute approximate Goldstein subgradients almost surely using a simple but ingenious randomized procedure from [20]. Consider any vector g in the Goldstein subdifferential $\partial_\epsilon f(x)$ that is not an approximate Goldstein subgradient. Then, the shortest convex convex combination g' of g and any subgradient of f at a point uniformly distributed between x and $x - \frac{\epsilon}{|g|} g$ is likely to be substantially shorter than g. Updating $g = g'$ and repeating eventually produces an approximate Goldstein subgradient, as shown in [20]. We describe the procedure more formally as follows.

Algorithm 8.2 (Approximate Goldstein subgradient for Lipschitz f)

```
input: center $x \in \mathbf{X}$, radius $\epsilon > 0$
output: approximate Goldstein subgradient $g$
choose $g \in \partial f(x)$
$\gamma = |g|$
while $\gamma \geq \epsilon$ do
  $y = x - \frac{\epsilon}{\gamma} g$                    {$g$ is not small so check descent property}
  if $f(x) - f(y) > \frac{\epsilon\gamma}{2}$ then
    break                                 {$g$ is an approximate Goldstein subgradient}
  end if
  sample $z \in [x, y]$ uniformly at random
  choose $h \in \partial f(z)$
  $g$ = shortest vector in $[g, h]$
  $\gamma = |g|$
end while
return $g$
```

Consider any point $x \in \mathbf{X}$ that is not Clarke critical. By Proposition 2.4, for all small $\epsilon > 0$, every element of the Goldstein subdifferential $\partial_\epsilon f(x)$ has norm at least ϵ. In particular, the approximate Goldstein subgradient produced by Algorithm 8.2 has norm at least ϵ. Consequently, starting from any initial radius $\epsilon > 0$, if we mimic our conceptual Algorithm 7.5 by repeatedly shrinking the radius and running Algorithm 8.2, then we eventually balance the sizes of the radius and a corresponding approximate Goldstein subgradient. The final radius is our estimate of the Goldstein modulus $\Gamma(x)$. We describe the procedure below, including a tolerance $\bar{\epsilon} > 0$ that triggers termination if we encounter a small approximate Goldstein subgradient corresponding to a small radius.

Algorithm 8.3 (Goldstein modulus estimation for Lipschitz f)

```
input: center $x \in \mathbf{X}$, initial radius $\epsilon > 0$, tolerance $\bar{\epsilon} > 0$
output: Goldstein modulus estimate $\epsilon > 0$,
        approximate Goldstein subgradient $g$ satisfying $|g| \geq \epsilon$
repeat
  $\epsilon = \frac{1}{2}\epsilon$
  find approximate Goldstein subgradient $g \in \partial_\epsilon f(x)$ by Algorithm 8.2
  if $|g| < \epsilon < \bar{\epsilon}$ then
    print "$x$ is approximately stationary"
    terminate
  end if
until $|g| \geq \epsilon$
return radius $\epsilon$, subgradient $g$
```

We now mimic the philosophy of Algorithm 7.5, replacing its Goldstein subgradients by their approximate versions.

Algorithm 8.4 (Minimization for Lipschitz f)

```
input: Lipschitz constant $L$, initial point $x \in \mathbf{X}$,
       tolerance $\bar{\epsilon} > 0$, multiplier $\beta > 0$, maximum iterations $n$
output: approximate stationary point $x$
for iteration $= 1, 2, 3, \ldots, n$ do
  $\epsilon = L$
  run Algorithm 8.3 to set $\epsilon =$ Goldstein modulus estimate
  $\epsilon = 2\beta\epsilon$
  run Algorithm 8.3 again:
    • shrink radius $\epsilon$ further, if necessary
    • find approximate Goldstein subgradient $g \in \partial_\epsilon f(x)$.
  $x = x - \epsilon\frac{g}{|g|}$
end for
return $x$
```

We illustrate Algorithm 8.4 on a simple random example. We define a nonsmooth nonconvex function $f \colon \mathbf{R}^{10} \to \mathbf{R}$ by

$$f(x) \;=\; \max_{1 \leq i \leq 5} \{g_i^T x + x^T H_i x\} \qquad (x \in \mathbf{R}^{10}) \tag{8.5}$$

for vectors $g_i \in \mathbf{R}^{10}$ and 10-by-10 symmetric matrices H_i (for $i = 1, 2, 3, 4$) with entries uniformly distributed on the interval $[-1, 1]$, and with $\sum_{i \leq 5} g_i = 0$ and $\sum_{i \leq 5} H_i$ equal to the identity matrix. This construction ensures that f is almost surely a strong $\mathcal{C}^2$ max function at its global minimizer, $x_{\min} = 0$. At any point $x \in \mathbf{R}^{10}$, to calculate a subgradient $h \in \partial f(x)$, we choose the index i attaining the

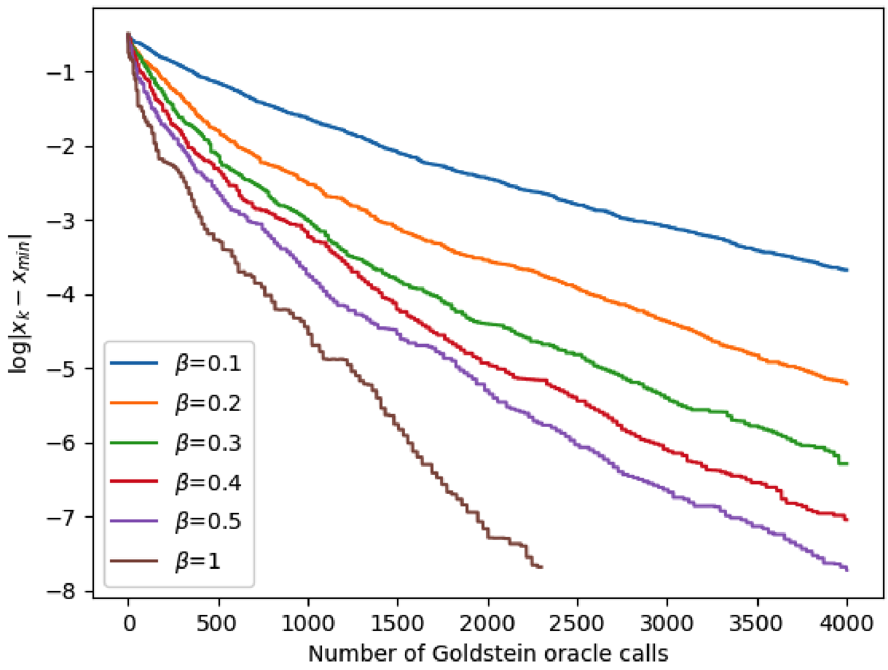

Figure 1: Algorithm 8.4 minimizing a maximum of five nonconvex quadratics on $\mathbf{R}^{10}$, using an approximate Goldstein subgradient oracle, illustrating near-linear convergence of the iterates to the minimizer.

max in equation (8.5), and set $h = g_i + 2H_i x$. We can then run Algorithm 8.4 from a random initial point, for various values of the multiplier β. We plot the progress of both the distance to the minimizer, $|x - x_{\min}|$ (in Figure 1) and the objective gap, $f(x) - f(x_{\min})$ (in Figure 2), against the number of calls to Algorithm 8.2 — our surrogate for the Goldstein subgradient oracle.

We emphasize that Algorithm 8.4 merely a heuristic. We have not explored if and why the approximate Goldstein subgradients produced by Algorithm 8.2 can serve as a useful substitute for the true Goldstein subgradient. Nonetheless, the behavior of Algorithm 8.4, as illustrated in Figure 1, is strikingly suggestive of the near-linear convergence that our theory predicts for the idealized method, Algorithm 7.5.

Our surrogate for the Goldstein subgradient oracle, Algorithm 8.2, is written for simplicity rather than with any aim at efficiency with respect to the number of subgradients computed. Nonetheless, for interest, Figure 3 plots the distance to the minimizer as a function of subgradient calls.

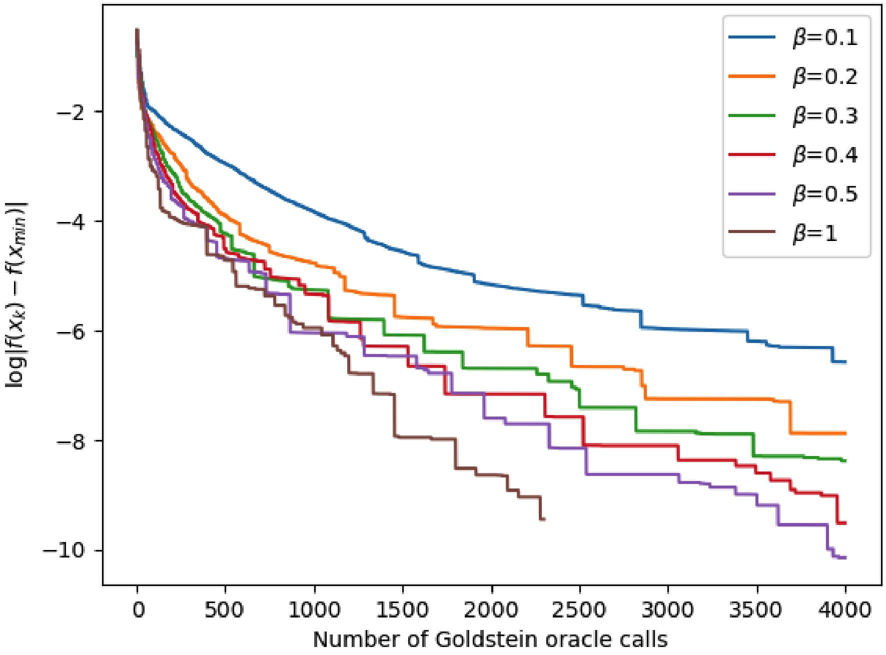

Figure 2: The example of Figure 1, illustrating convergence of the objective value.

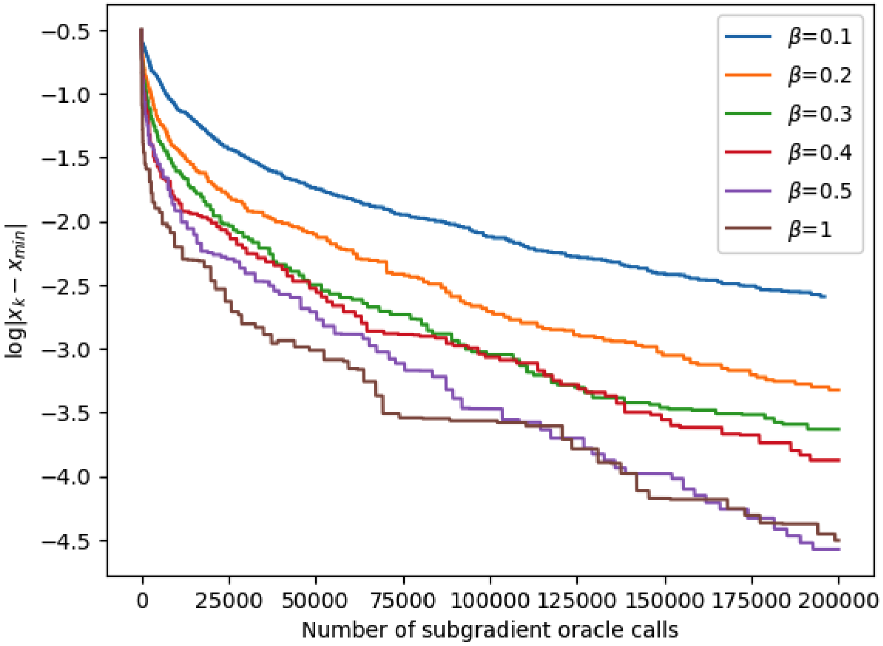

Figure 3: The example of Figure 1, illustrating convergence with respect to subgradient evaluations.

References

[1] F. Atenas, C. Sagastizábal, P. J. S. Silva, and M. Solodov. A unified analysis of descent sequences in weakly convex optimization, including convergence rates for bundle methods. *SIAM Journal on Optimization*, 33:89–115, 2023.

[2] N. Boumal. *An Introduction to Optimization on Smooth Manifolds.* Cambridge University Press, Cambridge, 2023.

[3] F.H. Clarke. *Optimization and Nonsmooth Analysis.* Wiley Interscience, New York, 1983.

[4] D. Davis, D. Drusvyatskiy, Yin Tat Lee, S. Padmanabhan, and Guanghao Ye. A gradient sampling method with complexity guarantees for lipschitz functions in high and low dimensions. In *NeurIPS Proceedings*, 2022.

[5] D. Davis and Liwei Jiang. A nearly linearly convergent first-order method for nonsmooth functions with quadratic growth. *Found. Comput. Math.*, to appear, 2024. `arXiv:2205.00064v3`.

[6] D. Drusvyatskiy and A.S. Lewis. Optimality, identifiability, and sensitivity. *Math. Program.*, 147:467–498, 2014.

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

[8] A.A. Goldstein. Optimization of Lipschitz continuous functions. *Math. Programming*, 13:14–22, 1977.

[9] E. Hazan and S. Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *J. Mach. Learn. Res.*, 15:2489–2512, 2014.

[10] A.D Ioffe. *Variational Analysis of Regular Mappings.* Springer US, 2017.

[11] M.I. Jordan, G. Kornowski, Tianyi Lin, O. Shamir, and M. Zampetakis. Deterministic nonsmooth nonconvex optimization. In *Proceedings of Machine Learning Research*, volume 195, pages 1–28, 2023.

[12] M.I. Jordan, Tianyi Lin, and M. Zampetakis. On the complexity of deterministic nonsmooth and nonconvex optimization. arXiv:2209.12463, 2022.

[13] Siyu Kong and A.S. Lewis. The cost of nonconvexity in deterministic nonsmooth optimization. *Mathematics of Operations Research*, `doi.org/10.1287/moor.2022.0289`, 2023.

[14] G. Kornowski and O. Shamir. On the complexity of finding small subgradients in nonsmooth optimization. arXiv:2209.10346, 2022.

[15] A.S. Lewis and Tonghua Tian. Identifiability, the KL property in metric spaces, and subgradient curves. *Fourndations of Computational Mathematics*, 2024. To appear.

[16] A.S. Lewis and S. Zhang. Partial smoothness, tilt stability, and generalized Hessians. *SIAM J. Optim.*, 23(1):74–94, 2013.

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

[18] S.M. Robinson. Linear convergence of epsilon-subgradient descent methods for a class of convex functions. *Math. Program.*, 86:41–50, 1999.

[19] Lai Tian and Anthony Man-Cho So. Computing Goldstein (ϵ, δ)-stationary points of Lipschitz functions in $\widetilde{O}(\epsilon^{-3}\delta^{-1})$ iterations via random conic perturbation. `arxiv.org/abs/2112.09002`, 2021.

[20] Jingzhao Zhang, Hongzhou Lin, S. Jegelka, S. Sra, and A. Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In *ICML Proceedings*, 2020.