Les Cahiers du GERAD ISSN: 0711–2440

Complexity of trust-region methods with unbounded Hessian approximations for smooth and nonsmooth optimization

G. Leconte, D. Orban

G–2023–65

December 2023

La publication de ces rapports de recherche est rendue possible grâce au soutien de HEC Montréal, Polytechnique Montréal, Université McGill, Université du Québec à Montréal, ainsi que du Fonds de recherche du Québec – Nature et technologies.

Dépôt légal – Bibliothèque et Archives nationales du Québec, 2023
– Bibliothèque et Archives Canada, 2023

The publication of these research reports is made possible thanks to the support of HEC Montréal, Polytechnique Montréal, McGill University, Université du Québec à Montréal, as well as the Fonds de recherche du Québec – Nature et technologies.

Legal deposit – Bibliothèque et Archives nationales du Québec, 2023
– Library and Archives Canada, 2023

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

Complexity of trust-region methods with unbounded Hessian approximations for smooth and nonsmooth optimization

Geoffroy Leconte

Dominique Orban

Département de mathématiques et de génie industriel, Polytechnique Montréal, Montréal, (Qc), Canada, H3T 1J4

geoffroy.leconte@polymtl.ca
dominique.orban@gerad.ca

December 2023
Les Cahiers du GERAD
G–2023–65

Abstract : We develop a worst-case evaluation complexity bound for trust-region methods in the presence of unbounded Hessian approximations. We use the algorithm of Aravkin et al. [3] as a model, which is designed for nonsmooth regularized problems, but applies to unconstrained smooth problems as a special case. Our analysis assumes that the growth of the Hessian approximation is controlled by the number of successful iterations. We show that the best known complexity bound of ϵ^{-2} deteriorates to $\epsilon^{-2/(1-p)}$, where $0 \leq p < 1$ is a parameter that controls the growth of the Hessian approximation. The faster the Hessian approximation grows, the more the bound deteriorates. We construct an objective that satisfies all of our assumptions and for which our complexity bound is attained, which establishes that our bound is sharp. Numerical experiments conducted in double precision arithmetic are consistent with the theoretical analysis.

Résumé : Nous présentons une analyse de la borne de complexité dans le pire des cas pour les méthodes de région de confiance en présence d'approximations du Hessien non bornées. Nous utilisons l'algorithme de Aravkin et al. [3] comme modèle, qui, bien qu'étant conçu spécifiquement pour les problèmes non lisses régularisés, s'applique aussi dans le cas particulier des problèmes lisses non contraints. Notre analyse fait l'hypothèse que la croissance des approximations des Hessiens est contrôlée par le nombre d'itérations concluantes. Nous montrons que la borne de complexité bien connue ϵ^{-2} se détériore en $\epsilon^{-2/(1-p)}$, où $0 \leq p < 1$ est un paramètre contrôlant la croissance des approximations des Hessiens. Plus les approximations des Hessiens augmentent, et plus la borne se dégrade. Nous construisons une fonction objectif satisfaisant toutes nos hypothèses pour laquelle la borne de complexité est atteinte, ce qui montre que notre borne est la plus petite possible. Nous présentons des résultats numériques cohérents avec notre analyse théorique.

Acknowledgements: Research supported by an NSERC Discovery grant.

Competing interests: We certify that the research submitted here is original, is our own, and is not being evaluated elsewhere for publication. This work was supported by an NSERC Discovery grant.

Data availability: The code used to produce the numerical results is available from https://github.com/geoffroyleconte/unbounded-hessian-code. The solvers are available from https://github.com/geoffroyleconte/RegularizedOptimization.jl/tree/unbounded.

1 Introduction

We consider the nonsmooth regularized problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \;\; f(x) + h(x) \quad \text{subject to } \ell \le x \le u, \tag{1}$$

where $\ell \in (\mathbb{R} \cup \{-\infty\})^n$, $u \in (\mathbb{R} \cup \{+\infty\})^n$ with $\ell \le u$ componentwise, $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable on an open set containing the feasible set $[\ell, u]$ of (1), and $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is proper and lower semicontinuous (lsc). A component $\ell_i = -\infty$ or $u_i = +\infty$ indicates that x_i is unbounded below or above, respectively. Both f and h may be nonconvex. The nonsmooth regularizer h is often used to identify a local minimizer of f with desirable features, such as sparsity.

Algorithms used to solve (1) are often based on the proximal-gradient method [23, 27]. The algorithm that we consider here is the trust-region method (TR) of Aravkin et al. [3], which improves upon the proximal-gradient method by constructing a model of f and a model of h at each iteration in order to compute a step, in the spirit of traditional trust-region methods [15]. To the best of our knowledge, it is the only trust-region method for (1) that allows both f and h to be nonconvex, and that only assumes that h is proper lsc. However, it was developed under the assumption that the Hessian approximations B_k remain bounded, a common, but sometimes restrictive, assumption. A worst-case evaluation complexity bound for a stationarity measure to drop below $\epsilon \in (0, 1)$ of $O(\epsilon^{-2})$ results, which matches the best possible complexity bound in the smooth case, i.e., when $h = 0$ [14].

In the present paper, we examine the situation where $\{B_k\}$ is allowed to grow unbounded. We impose a bound on the growth of $\|B_k\|$ in terms of the number of successful iterations that is slightly more restrictive than bounds used in smooth optimization to establish global convergence—see below. Our tighter growth control, however, allows us to formalize a worst-case evaluation complexity bound, which we then show to be tight. Specifically, we show that the best known complexity bound of $O(\epsilon^{-2})$ deteriorates to $O(\epsilon^{-2/(1-p)})$, where $0 \le p < 1$ is a parameter that controls the growth of $\|B_k\|$. To the best of our knowledge, this is the first formal worst-case analysis in the case of potentially unbounded B_k.

A Julia implementation of TR is available as part of the RegularizedOptimization.jl package [5]. Our findings also apply to Algorithm TRDH of Leconte and Orban [26], which is similar to TR, but uses diagonal Hessian approximations to compute a step without recourse to a subproblem solver.

Unbounded, or potentially unbounded, Hessians are not uncommon in applications. A prime example is interior-point methods for bound-constrained optimization. Consider the minimization of a twice differentiable objective $\phi : \mathbb{R}^n \to \mathbb{R}$ subject to simple bounds $x \ge 0$. Primal interior-point methods [20] consist in applying Newton's method to a sequence of log-barrier subproblems whose objective is $\phi(x) - \mu \sum_i \log(x_i)$ where $\mu > 0$ is a barrier parameter that is eventually driven to zero. Such methods maintain $x > 0$ implicitly but the barrier objective Hessian is $\nabla^2\phi(x) + \mu X^{-2}$, where $X := \text{diag}(x)$. For any $\mu > 0$, the barrier Hessian is unbounded as any component of x approaches a bound, which is often where a solution is located. Primal methods have long been superseded by the better-behaved primal-dual methods—see, e.g., [22] and references therein for an overview of the extensive literature on the subject—in which the barrier Hessian is replaced with $\nabla^2\phi(x) + X^{-1}Z$, where $Z := \text{diag}(z)$ and z is an approximation of the vector of Lagrange multipliers for $x \ge 0$. Even though the primal-dual Hessian does not grow unbounded as fast as the primal Hessian, it nevertheless remains unbounded as any component of x approaches a bound. In order to converge, interior-point methods rely on extra mechanisms that prevent components of x from approaching a bound too fast unless there are indications that a solution is nearby and μ is close to zero. In spite of those mechanisms, x must be allowed to approach bounds, and, therefore, the primal and primal-dual Hessians must be allowed to grow unbounded. Although primal-dual interior-point methods can be shown to have excellent worst-case complexity bounds in convex optimization [31], no such general result is known for nonconvex problems.

Another prime example, often cited in the literature, is when B_k results from a secant approximation [18]. Conn et al. [15, §8.4] suggest that for the BFGS and SR1 approximations, B_k could potentially grow by at most a constant at each update, though it is not clear whether that bound is attained. Aravkin et al. [1], Carter [7] and Lotfi et al. [28] present safeguarding strategies that ensure boundedness of quasi-Newton approximations in order to preserve convergence and $O(\epsilon^{-2})$ worst-case evaluation complexity properties. This point is developed further in the related research below.

The paper is organized as follows. Section 2 provides the nonsmooth analysis background necessary to understand the algorithm of Aravkin et al. [3], a description of how models are constructed at each iteration, and a formal statement of the algorithm. In Section 3, we establish convergence and a worst-case evaluation complexity bound under the assumption that the growth of the model Hessian is controlled by a function of the number of successful iterations, i.e., iterations in which a step is accepted. We show in Section 4 that the worst-case bound is indeed attained, by performing an analysis similar to that of [14, Theorem 2.2.3]. In Section 5, we construct an explicit function that attains the bound and validate our findings numerically. We provide concluding comments and perspectives in Section 6.

Related research

We do not provide an extensive review of trust-region approaches for smooth optimization, but refer the interested reader to [15] for a thorough account, as well as a number of generalizations.

We begin by reviewing milestones in the convergence analysis of trust-region methods with potentially unbounded model Hessians. Powell [34] first showed convergence of a trust-region algorithm for smooth optimization that allows unbounded Hessian approximations B_k. Specifically, he assumes that there exist nonnegative α and β such that $\|B_k\| \le \alpha + \beta \sum_{i=0}^{k-1} \|s_j\|$, where s_j is the trust-region step at iteration j. Under that and other standard assumptions, he established that $\liminf \|\nabla f(x_k)\| = 0$. Powell hints that his motivation lies in Hessian approximations arising from secant updates [18]. To the best of our knowledge, it is not known whether secant approximations and their limited-memory counterparts remain bounded. However, Fletcher [21] establishes that the quasi-Newton update that bears Powell's name, the Powell symmetric Broyden update, derived in [33], satisfies the bound above.

Powell [35] refines his earlier analysis by showing global convergence under the weaker assumption $\|B_k\| \le \alpha + \beta k$. Under the weaker yet assumption

$$\sum_{k=0}^{\infty} \frac{1}{1 + \max_{0 \le j \le k} \|B_j\|} = \infty, \tag{2}$$

which is hinted at in the proofs of Powell [35], Toint [39] shows that global convergence is preserved.

When f is convex with uniformly bounded Hessian, Conn et al. [15, §8.4] indicate that the BFGS update satisfies $\|B_{k+1}\| \le \|B_k\| + \beta$ for some $\beta \ge 0$. Therefore, $\|B_{k+1}\| \le \|B_0\| + (k+1)\beta$, and the assumption of Powell [35], and hence that of Toint [39], are satisfied. The SR1 update with safeguards satisfies a similar inequality without the convexity assumption.

Carter [7] presents procedures to safeguard Hessian approximations in trust-region algorithms for smooth problems. The goal of these procedures is to satisfy the *uniform predicted decrease condition*

$$\varphi_k(x_k) - \varphi_k(x_{k+1}) \ge \tfrac{1}{2}\beta_1 \|\nabla f(x_k)\| \min\left(\Delta_k, \frac{\|\nabla f(x_k)\|}{\beta_0}\right),$$

where β_0 and $\beta_1 > 0$. When $\|B_k\| \le \beta_0$ for all k, this condition is satisfied, but the author shows that it can also be satisfied under milder assumptions. Carter's procedures are used to correct B_k so that such assumptions hold. Aravkin et al. [3] and Lotfi et al. [28] instead maintain estimates of the largest and smallest eigenvalues of limited-memory BFGS and SR1 approximations and use them to ensure updates generate bounded Hessian approximations.

We now review determinant complexity analyses of trust-region and related methods for smooth optimization. Cartis et al. [8] show that the steepest descent method and Newton's method for smooth problems may converge in as many as $O(\epsilon^{-2})$ iterations, and that the bound is sharp for the steepest descent method. The analysis assumes that the Hessian remains uniformly bounded. In addition, they prove that it is possible to construct an example where Newton's method is arbitrarily slow when allowing unbounded Hessians.

Our main contribution is to establish that TR, the trust-region algorithm of [3], may converge in as many as $O(\epsilon^{-2/(1-p)})$ iterations, where $p \in [0, 1)$ is a parameter that controls the growth of the model Hessian—the larger p, the larger the allowed growth. Because $\epsilon^{-2/(1-p)} \to +\infty$ as $p \nearrow 1$, our results reinforces that of Cartis et al. [8] and makes it more precise. Our analysis applies to smooth optimization—indeed, the example that we construct to establish sharpness of the complexity bound is smooth—but it is general enough to apply to (1).

Cartis et al. [14, Section 2.2] show that the steepest-descent algorithm with backtracking Armijo linesearch results in an $O(\epsilon^{-2})$ complexity bound, and a function is constructed by polynomial interpolation to prove that the bound is sharp, with a technique that is different from that of [8].

The complexity of other methods for smooth optimization was subsequently analyzed using techniques similar to those of [8]. The Adaptive Regularization with Cubics algorithm (ARC, or AR2 because it uses second-order derivatives) [9, 19] minimizes at each iteration the model

$$\varphi_k(x_k + s) = f(x_k) + \nabla f(x_k)^T s + \tfrac{1}{2} s^T B_k s + \tfrac{1}{3} \sigma_k \|s\|^3, \tag{3}$$

where B_k must remain bounded. It is known to require at most $O(\epsilon^{-3/2})$ iterations to reach $\|\nabla f(x_k)\| \le \epsilon$, and this bound is sharp [9, 32]. Curtis et al. [16] and Martínez and Raydan [30] present modified trust-region algorithms with bounded model Hessians to solve nonconvex smooth problems that also have a complexity bound of $O(\epsilon^{-3/2})$.

The analysis of [14, Section 2.2] shows that the steepest-descent algorithm with backtracking Armijo linesearch technique results in an $O(\epsilon^{-2})$ complexity bound, and a function is constructed by polynomial interpolation to prove that the bound is sharp, with a technique that is different to that of [8].

Cartis et al. [12] show that Algorithm ARp for smooth problems, a generalization of ARC using a model of order $p \ge 1$, requires at most $O(\epsilon^{-(p+1)/p})$ iterations to satisfy $\|\nabla f(x_k)\| \le \epsilon$, and that the bound is sharp. They introduce a generalization of the first-order stationarity measure $\|\nabla f(x_k)\| \le \epsilon$ to q-th order stationarity, where $q \in \mathbb{N}_0$, and show that at most $O(\epsilon^{-(p+1)/(p-q+1)})$ evaluations of the objective and the derivatives are required with this measure. They require that the p-th derivative of f be globally Hölder continuous. For $p = 2$ and $q = 1$, we recover the bound of [9].

For smooth nonconvex problems with bounded Hessians, the number of iterations required to satisfy the conditions on the gradient $\|\nabla f(x_k)\| \le \epsilon_g$ and on the smallest eigenvalue of the Hessian $\lambda_{\min}(\nabla^2 f(x_k)) \ge -\epsilon_H$, where ϵ_g, $\epsilon_H \in (0, 1)$, have also been studied. Cartis et al. [11] show that their trust-region algorithm needs at most $O(\max\{\epsilon_g^{-2}\epsilon_H^{-1}, \epsilon_H^{-3}\})$ iterations to satisfy these conditions, and $O(\max\{\epsilon_g^{-3/2}, \epsilon_H^{-3}\})$ iterations for ARC. The latter bound is also obtained for the trust-region algorithms in [16, 30]. Royer and Wright [38] use a second-order linesearch method to obtain the bound $O(\max\{\epsilon_g^{-3}\epsilon_H^{3}, \epsilon_g^{-3/2}, \epsilon_H^{-3}\})$.

Aravkin et al. [3] provide an overview of the literature on convergence of methods for nonsmooth optimization, and we now summarize the review with an eye to trust-region methods. Methods prior to their work were restricted to special cases. Most were developed for $f = 0$, i.e., in a purely nonsmooth context. Yuan [40] considers a nonsmooth term of the form $h(c(x))$, where $c \in \mathcal{C}^1$ and convex. Dennis et al. [17] take $f = 0$ and assume that h is Lipschitz-continuous. Qi and Sun [36] relax the assumptions of [17] to h locally Lipschitz-continuous with bounded level sets. Martínez and Moretti [29] add treatment of equality constraints to the method of Qi and Sun [36]. The only prior trust-region method for $f \neq 0$ and more general h that we are aware of is that of Kim et al. [25], who assume that f and h are convex. None of those works provides a complexity analysis.

Finally, we review complexity analyses of trust-region methods for nonsmooth problems. Cartis et al. [10] describe a first-order trust-region method and a quadratic regularization algorithm to solve nonsmooth problem of the form

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \;\; f(x) + h(c(x)), \tag{4}$$

where f and c are continuously differentiable and may be nonconvex, and h is convex but may be nonsmooth, and is Lipschitz-continuous. Note that (4) is a special case of (1), but the convexity assumption on h is strong. They show that both algorithms have a complexity bound of $O(\epsilon^{-2})$. Grapiglia et al. [24] provide a unified convergence theory for smooth optimization that has trust-region methods as a special case. They also generalize the results of [10] under the same assumptions.

Aravkin et al. [3] describe a proximal trust-region algorithm to solve (1) using bounded model Hessians. They also present a quadratic regularization variant. They establish that their criticality measure is smaller than ϵ in at most $O(\epsilon^{-2})$ iterations for both algorithms. Aravkin et al. [1] adapt these algorithms to solve nonsmooth regularized least-squares problems and obtain the same complexity bound under the assumption that the residual Jacobian is uniformly bounded. As far as we know, the complexity analyses of [1, 3] make the weakest assumptions on h so far, that h be lsc.

Baraldi and Kouri [4] also describe a proximal trust-region algorithm for convex h. In addition, they allow the use of inexact objective and gradient evaluations. As Toint [39] in the smooth case, they assume that

$$\sum_{k=0}^{\infty} \frac{1}{1 + \max_{0 \le j \le k} \omega_j} = \infty, \tag{5}$$

where

$$\omega_k = \sup \left\{ \frac{2}{\|s\|^2} |\varphi_k(x_k + s) - \varphi_k(x_k) - \nabla \varphi_k(x_k)^T s| \mid 0 < \|s\| \le \Delta_k \right\},$$

and φ_k is a smooth model of f about x_k. In particular, if φ_k is a second-order Taylor approximation at x_k with Hessian approximation B_k, $\omega_k = \sup \left\{ s^T B_k s / \|s\|^2 \mid 0 < \|s\| \le \Delta_k \right\}$, so that (5) is reminiscent of (2). If ω_k is bounded independently of k, which is the case for bounded Hessian approximations, they show that their algorithm enjoys a complexity bound of $O(\epsilon^{-2})$.

Cartis et al. [13] present a similar concept of high-order approximate minimizers to that of [12] for nonsmooth problems such as (4) where f, c are smooth, and h is nonsmooth but Lipschitz-continuous. They present an algorithm of adaptive regularization of order p, and derive several bounds depending on the properties of (4) and of the order of the desired approximate minimizer. In particular, for $q = 1$ and convex h, their complexity bound is $O(\epsilon^{-(p+1)/p})$, and they show that it is sharp.

To the best of our knowledge, previous literature does not provide a complexity analysis in the case of potentially unbounded model Hessians.

Notation. $\mathbb{B}$ denotes the unit ball at the origin in a certain norm dictated by the context, $\Delta\mathbb{B}$ is the ball of radius $\Delta > 0$ centered at the origin, and $x + \Delta\mathbb{B}$ is the ball of radius $\Delta > 0$ centered at $x \in \mathbb{R}^n$. For $A \subseteq \mathbb{R}^n$, the indicator of A is $\chi(\cdot \mid A) : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ defined as $\chi(x \mid A) = 0$ if $x \in A$ and $+\infty$ otherwise. If $A \neq \varnothing$, $\chi(\cdot \mid A)$ is proper. If A is closed, $\chi(\cdot \mid A)$ is lsc. For a finite set $A \subset \mathbb{N}$, we denote $|A|$ its cardinality. If f_1 and f_2 are two positive functions of $\epsilon > 0$, we say that $f_1(\epsilon) = O(f_2(\epsilon))$ if there exists a constant $C > 0$ such that $f_1(\epsilon) \le C f_2(\epsilon)$ for all $\epsilon > 0$ sufficiently small.

2 Context

2.1 Background

We recall relevant concepts of variational analysis—see, e.g., [37].

Consider $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$ and $\bar{x} \in \mathbb{R}^n$ with $\phi(\bar{x}) < \infty$. The *Fréchet subdifferential* of ϕ at $\bar{x}$ is the closed convex set $\widehat{\partial}\phi(\bar{x})$ of $v \in \mathbb{R}^n$ such that

$$\liminf_{\substack{x \to \bar{x} \\ x \neq \bar{x}}} \frac{\phi(x) - \phi(\bar{x}) - v^T(x - \bar{x})}{\|x - \bar{x}\|} \geq 0.$$

The *limiting subdifferential* of ϕ at $\bar{x}$ is the closed, but not necessarily convex, set $\partial\phi(\bar{x})$ of $v \in \mathbb{R}^n$ for which there exist $\{x_k\} \to \bar{x}$ and $\{v_k\} \to v$ such that $\{\phi(x_k)\} \to \phi(\bar{x})$ and $v_k \in \widehat{\partial}\phi(x_k)$ for all k. $\widehat{\partial}\phi(\bar{x}) \subset \partial\phi(\bar{x})$ always holds.

We say that $\bar{x}$ is *stationary* for the problem of minimizing ϕ if $0 \in \partial\phi(\bar{x})$.

The *horizon subdifferential* of ϕ at $\bar{x}$ is the closed, but not necessarily convex, cone $\partial^\infty\phi(\bar{x})$ of $v \in \mathbb{R}^n$ for which there exist $\{x_k\} \to \bar{x}$, $\{v_k\}$ and $\{\lambda_k\} \downarrow 0$ such that $\{\phi(x_k)\} \to \phi(\bar{x})$, $v_k \in \widehat{\partial}\phi(x_k)$ for all k, and $\{\lambda_k v_k\} \to v$.

If $C \subseteq \mathbb{R}^n$ and $\bar{x} \in C$, the closed convex cone $\widehat{N}_C(\bar{x}) := \widehat{\partial}\chi(\bar{x} \mid C)$ is the regular normal cone to C at $\bar{x}$. The closed cone $N_C(\bar{x}) := \partial\chi(\bar{x} \mid C) = \partial^\infty\chi(\bar{x} \mid C)$ is the normal cone to C at $\bar{x}$. $\widehat{N}_C(\bar{x}) \subseteq N_C(\bar{x})$ always holds, and is an equality if C is convex.

ϕ is *proper* if $\phi(x) > -\infty$ for all x, and $\phi(x) < \infty$ for at least one x. ϕ is *lower semicontinuous* (*lsc*) at $\bar{x}$ if $\liminf_{x \to \bar{x}} \phi(x) = \phi(\bar{x})$.

Let $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper lsc, and $C \subseteq \mathbb{R}^n$ be closed. We say that the *constraint qualification* is satisfied at $\bar{x} \in C$ for the constrained problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \ \phi(x) \quad \text{subject to } x \in C \tag{6}$$

if

$$\partial^\infty\phi(\bar{x}) \cap N_C(\bar{x}) = \{0\}. \tag{7}$$

If $\bar{x}$ solves (6) and (7) is satisfied at $\bar{x}$, [37, Theorem 8.15] yields

$$0 \in \partial(\phi + \chi(\cdot \mid C))(\bar{x}) = \partial\phi(\bar{x}) + N_C(\bar{x}).$$

In the case of (1), this first-order necessary condition for optimality reads

$$0 \in \nabla f(\bar{x}) + \partial h(\bar{x}) + N_{[\ell,u]}(\bar{x})$$

thanks to [37, Exercise 8.8*c*].

The *proximal operator* associated with a proper lsc function ϕ is

$$\underset{\nu\phi}{\text{prox}}(q) := \underset{x}{\text{argmin}} \ \tfrac{1}{2}\nu^{-1}\|x - q\|_2^2 + \phi(x), \tag{8}$$

where $\nu > 0$ is a preset steplength.

If ϕ is prox-bounded and $\nu > 0$ is sufficiently small, $\text{prox}_{\nu\phi}(q)$ is a nonempty and closed set. It may contain multiple elements.

The proximal gradient method [23, 27] for (1) is a generalization of the gradient method that takes the nonsmooth term into account. It generates iterates $\{s_j\}$ according to

$$s_{j+1} \in \underset{\nu h}{\text{prox}}(s_j - \nu\nabla f(s_j)). \tag{9}$$

2.2 Models and trust-region algorithm

At $x \in \mathbb{R}^n$ where h is finite, we define models

$$\begin{aligned} \varphi(s;x) &\approx f(x+s) && (10a)\\ \psi(s;x) &\approx h(x+s) && (10b)\\ m(s;x) &:= \varphi(s;x) + \psi(s;x), && (10c) \end{aligned}$$

Our assumptions on (10) are the same as those of Aravkin et al. [3]:

Model Assumption 2.1. For any $x \in \mathbb{R}^n$, $\varphi(\cdot;x) \in \mathcal{C}^1$, and satisfies $\varphi(0;x) = f(x)$ and $\nabla\varphi(0;x) = \nabla f(x)$. For any $x \in \mathbb{R}^n$ where h is finite, ψ is proper lsc, and satisfies $\psi(0;x) = h(x)$ and $\partial\psi(0) = \partial h(x)$.

The following result states that if $s = 0$ minimizes (10c) and (7) is satisfied, x must be stationary.

Proposition 1 (26, Proposition 1)**.** *Let $C \subset \mathbb{R}^n$ be nonempty and compact, and let Model Assumption 2.1 be satisfied. Let (1) satisfy the constraint qualification (7) at $x \in C$. Assume $0 \in \operatorname{argmin}_s m(s;x) + \chi(x+s \mid C)$, and let the latter subproblem satisfy the constraint qualification (7) at $s = 0$. Then x is first-order stationary for (1).*

A useful model is based on the second-order Taylor expansion

$$\begin{aligned} \varphi(s;x,B) &:= f(x) + \nabla f(x)^T s + \tfrac{1}{2} s^T B s, && (11a)\\ m(s;x,B) &:= \varphi(s;x,B) + \psi(s;x), && (11b) \end{aligned}$$

where $B = B^T \in \mathbb{R}^{n\times n}$.

Each iteration is divided into two parts. In the first part, Aravkin et al. [2] define the following model based on a first-order Taylor expansion to compute a *Cauchy point*

$$\begin{aligned} \varphi_{\text{cp}}(s;x) &:= f(x) + \nabla f(x)^T s, && (12a)\\ m(s;x,\nu) &:= \varphi_{\text{cp}}(s;x) + \tfrac{1}{2}\nu^{-1}\|s\|^2 + \psi(s;x), && (12b) \end{aligned}$$

where $\nu_k > 0$ and "cp" stands for "Cauchy point." We compute a first step

$$s_{k,1} \in \operatorname*{argmin}_s \; m(s;x_k,\nu_k) + \chi(x_k + s \mid [\ell, u] \cap (x_k + \Delta_k \mathbb{B})), \tag{13}$$

for an appropriate value of $\nu_k > 0$.

In the notation of [2], let

$$\xi_{\text{cp}}(\Delta_k;x_k,\nu_k) := f(x_k) + h(x_k) - \varphi_{\text{cp}}(s_{k,1};x_k) - \psi(s_{k,1};x_k), \tag{14}$$

denote the optimal model decrease for (12). The following proposition indicates that $\xi_{\text{cp}}(\Delta;x,\nu)$ can be used to determine whether x is first-order stationary for (1).

Proposition 2 (3, Proposition 3.3 and 2)**.** *Let Model Assumption 2.1 be satisfied, $\Delta > 0$, and $\nu > 0$. In addition, let (1) satisfy the constraint qualification at x and (12) satisfy the constraint qualification at 0. Then, $\xi_{\text{cp}}(\Delta;x,\nu) = 0 \iff 0$ is a solution of (13) $\implies x$ is first-order stationary for (1).*

In the second part of iteration k, we construct $m_k(s;x_k,B_k) := \varphi(s;x_k,B_k) + \psi(s;x_k) \approx f(x_k + s) + h(x_k+s)$, and compute an approximate solution of

$$\operatorname*{minimize}_s \; m_k(s;x_k) \quad \text{subject to } \|s\| \le \Delta_k, \tag{15}$$

using $s_{k,1}$ as starting point.

Algorithm 2.1 Nonsmooth trust-region algorithm with potentially unbounded Hessian.

Choose constants

$$0 < \eta_1 \le \eta_2 < 1, \quad 0 < 1/\gamma_3 \le \gamma_1 \le \gamma_2 < 1 < \gamma_3 \le \gamma_4, \quad \Delta_{\max} > \Delta_0, \quad \alpha > 0, \quad \text{and} \quad \beta \ge 1.$$

Choose a stopping tolerance $\epsilon > 0$.
Choose $x_0 \in \mathbb{R}^n$ where h is finite, $\Delta_0 > 0$, compute $f(x_0) + h(x_0)$.
for $k = 0, 1, \ldots$ **do**
  Choose

$$0 < \nu_k \le \frac{\alpha \Delta_k}{1 + \|B_k\|(1 + \alpha \Delta_k)} = \frac{1}{\alpha^{-1}\Delta_k^{-1} + \|B_k\|(1 + \alpha^{-1}\Delta_k^{-1})}.$$

  Define $m_k(s; x_k, \nu_k)$ as in (12) and compute $s_{k,1}$ as in (13).
  If $\nu_k^{-1/2} \xi_{\mathrm{cp}}(\Delta_k; x_k, \nu_k)^{1/2} \le \epsilon$, terminate and claim that x_k is approximately stationary.
  Define $m_k(s; x_k, B_k)$ as in (11) according to Model Assumption 2.1 and compute a solution s_k of (15) with Δ_k replaced by $\min(\Delta_k, \beta\|s_{k,1}\|)$.
  Compute the ratio

$$\rho_k := \frac{f(x_k) + h(x_k) - (f(x_k + s_k) + h(x_k + s_k))}{m_k(0; x_k, B_k) - m_k(s_k; x_k, B_k)}. \tag{16}$$

  If $\rho_k \ge \eta_1$, set $x_{k+1} = x_k + s_k$. Otherwise, set $x_{k+1} = x_k$.
  Update the trust-region radius according to

$$\bar{\Delta}_{k+1} \in \begin{cases} [\gamma_3 \Delta_k, \, \gamma_4 \Delta_k] & \text{if } \rho_k \ge \eta_2, & \text{(very successful iteration)} \\ [\gamma_2 \Delta_k, \, \Delta_k] & \text{if } \eta_1 \le \rho_k < \eta_2, & \text{(successful iteration)} \\ [\gamma_1 \Delta_k, \, \gamma_2 \Delta_k] & \text{if } \rho_k < \eta_1, & \text{(unsuccessful iteration)} \end{cases}$$

  and $\Delta_{k+1} = \min(\bar{\Delta}_{k+1}, \Delta_{\max})$

We focus on the trust-region (TR) algorithm formally stated as Algorithm 2.1. It consists of the algorithm of Aravkin et al. [3] with a modified maximum allowable stepsize ν_k.

In Algorithm 2.1, $s_{k,1}$ is used to check for stationarity, and to set the trust-region radius for the computation of the search direction s_k.

Let us now briefly turn our attention to unconstrained smooth problems. In this case, the following lemma gives a global minimizer of (12) and (11).

Lemma 1. *We consider the special case of (1) where $h = 0$, $\ell_i = -\infty$ and $u_i = +\infty$ for $i = 1, \ldots, n$. Let $B = B^T \in \mathbb{R}^{n \times n}$ be positive definite and $\psi = 0$. Then for any $x \in \mathbb{R}^n$,*

$$\operatorname*{argmin}_s m(s; x, B) = \operatorname*{argmin}_s \varphi(s; x, B) = \{-B^{-1}\nabla f(x)\}. \tag{17}$$

In particular, if $B = \nu^{-1} I$ with $\nu > 0$,

$$\operatorname*{argmin}_s m(s; x, \nu) = \operatorname*{argmin}_s \varphi_{\mathrm{cp}}(s; x) + \tfrac{1}{2}\nu^{-1}\|s\|^2 = \{s_{k,1}\} = \{-\nu \nabla f(x)\}. \tag{18}$$

Proof. The objective of (17) is convex because B is positive definite. Its global minimizer satisfies the first-order necessary condition $\nabla f(x) + Bs = 0$, i.e., $s = -B^{-1}\nabla f(x)$. With $B = \nu^{-1} I$, the first-order necessary condition is $s = -\nu \nabla f(x)$. □

The following proposition draws a parallel between $\xi_{\mathrm{cp}}(\Delta_k; x_k, \nu_k)$ and $\|\nabla f(x_k)\|$ for smooth problems when the trust-region constraint is inactive, as is expected to occur when close to a stationary point.

Proposition 3. *We consider the special case of (1) where $h = 0$, $\ell_i = -\infty$ and $u_i = +\infty$ for $i = 1, \ldots, n$. If $\|s_{k,1}\| < \Delta_k$, then $\xi_{\mathrm{cp}}(\Delta_k; x_k, \nu_k) = \nu_k \|\nabla f(x_k)\|^2$.*

Proof. If the trust-region constraint is inactive, Lemma 1 indicates that $s_{k,1} = -\nu_k \nabla f(x_k)$. Thus, (14) yields $\xi_{\mathrm{cp}}(\Delta_k; x_k, \nu_k) = -\nabla f(x_k)^T s_{k,1} = \nu_k \|\nabla f(x_k)\|^2$. □

3 Convergence and complexity with potentially unbounded Hessian

From this section onwards, we consider the model defined in (11), and we aim to establish convergence and worst-case complexity results for Algorithm 2.1 in the presence of potentially unbounded Hessian approximations B_k.

The following two assumptions are essential. Assumption 1 is [3, Step Assumption 3.8b], whereas Assumption 2 is a relaxed version of [3, Step Assumption 3.8a] that takes into account potentially unbounded Hessian approximations. Indeed, assuming, for simplicity, that $\nabla^2 f(x_k)$ exists, a second-order Taylor expansion of f about x_k yields

$$f(x_k + s_k) - \varphi(s_k; x_k, B_k) = \tfrac{1}{2} s_k^T (\nabla^2 f(x_k) - B_k) s_k + o(\|s_k\|^2),$$

which is not necessarily $O(\|s_k\|^2)$ if $\{B_k\}$ is unbounded.

Assumption 1. There exists $\kappa_{\text{mdc}} \in (0, 1)$ such that

$$m(0; x_k, B_k) - m(s_k; x_k, B_k) \geq \kappa_{\text{mdc}} \xi_{\text{cp}}(\Delta_k; x_k, \nu_k). \tag{19}$$

Assumption 2. There exists $\kappa_{\text{ubd}} > 0$ such that

$$|(f + h)(x_k + s_k) - m(s_k; x_k, B_k)| \leq \kappa_{\text{ubd}} (1 + \|B_k\|) \|s_k\|_2^2. \tag{20}$$

Leconte and Orban [26, Proposition 2] and Aravkin et al. [2] already indicate that Assumption 1 holds for TRDH and TR. We now justify that it also holds for Algorithm 2.1 with potentially unbounded Hessian approximations.

Proposition 4. *If* Model Assumption 2.1 *is satisfied, there exists* $\kappa_{\text{mdc}} \in (0, 1)$ *such that* Assumption 1 *holds.*

Proof. We proceed similarly as in [26, Proposition 2]. The definition of s_k implies that

$$m(s_k; x_k, B_k) \leq m(s_{k,1}; x_k, B_k) = \varphi_{\text{cp}}(s_{k,1}; x_k) + \tfrac{1}{2} s_{k,1}^T B_k s_{k,1} + \psi(s_{k,1}; x_k).$$

As

$$s_{k,1}^T B_k s_{k,1} \leq |s_{k,1}^T B_k s_k| \leq \|s_{k,1}\| \|B_k s_{k,1}\| \leq \|B_k\| \|s_{k,1}\|^2,$$

where we used Cauchy-Schwarz in the second inequality and the consistency of the ℓ_2-norm for matrices in the third inequality,

$$m(s_k; x_k, B_k) \leq \varphi_{\text{cp}}(s_{k,1}; x_k) + \tfrac{1}{2} \|B_k\| \|s_{k,1}\|^2 + \psi(s_{k,1}; x_k),$$

which leads to

$$m(0; x_k, B_k) - m(s_k; x_k, B_k) \geq \xi_{\text{cp}}(\Delta_k; x_k, \nu_k) - \tfrac{1}{2} \|B_k\| \|s_{k,1}\|^2.$$

To satisfy Assumption 1, it is sufficient to show that there exists $\kappa_{\text{mdc}} \in (0, 1)$ such that

$$\xi_{\text{cp}}(\Delta_k; x_k, \nu_k) - \tfrac{1}{2} \|B_k\| \|s_{k,1}\|^2 \geq \kappa_{\text{mdc}} \xi_{\text{cp}}(\Delta_k; x_k, \nu_k),$$

i.e.,

$$(1 - \kappa_{\text{mdc}}) \xi_{\text{cp}}(\Delta_k; x_k, \nu_k) \geq \tfrac{1}{2} \|B_k\| \|s_{k,1}\|^2.$$

As $\xi_{\text{cp}}(\Delta_k; x_k, \nu_k) \geq \frac{1}{2} \nu_k^{-1} \|s_{k,1}\|^2$ by definition of $s_{k,1}$ and $\xi_{\text{cp}}(\Delta_k; x_k, \nu_k)$, it is also sufficient to show that there exists $\kappa_{\text{mdc}} \in (0, 1)$ such that

$$(1 - \kappa_{\text{mdc}}) \nu_k^{-1} \geq \|B_k\|. \tag{21}$$

Finally,

$$\|B_k\|\nu_k = \frac{1}{\alpha^{-1}\Delta_k^{-1}\|B_k\|^{-1} + 1 + \alpha^{-1}\Delta_k^{-1}} \le \frac{1}{\alpha^{-1}\Delta_{\max}^{-1}\|B_k\|^{-1} + 1 + \alpha^{-1}\Delta_{\max}^{-1}} \le \frac{1}{1 + \alpha^{-1}\Delta_{\max}^{-1}} \in (0,1). \quad (22)$$

We deduce from (22) that (21) holds, which is sufficient to satisfy Assumption 1. □

We begin the convergence analysis by showing that there still exists a Δ_{succ} as in [3, Theorem 3.4], despite our more general Assumption 2.

Theorem 1. *Let Model Assumption 2.1, Assumption 1 and Assumption 2 be satisfied and*

$$\Delta_{\text{succ}} := \frac{\kappa_{\text{mdc}}(1-\eta_2)}{2\kappa_{\text{ubd}}\alpha\beta^2} > 0.$$

If (1) satisfies the constraint qualification at x_k, (12) satisfies the constraint qualification at 0, x_k is not first-order stationary for (1), and $\Delta_k \le \Delta_{\text{succ}}$, then iteration k is very successful and $\Delta_{k+1} \ge \Delta_k$.

Proof. Lemma 2 of [6] guarantees that

$$\xi_{\text{cp}}(\Delta_k; x_k, \nu_k) \ge \tfrac{1}{2}\nu_k^{-1}\|s_{k,1}\|^2 \ge \tfrac{1}{2}(\alpha^{-1}\Delta_k^{-1} + \|B_k\|(1+\alpha^{-1}\Delta_k^{-1}))\|s_{k,1}\|^2 \ge \tfrac{1}{2}(\alpha^{-1}\Delta_k^{-1}(1+\|B_k\|))\|s_{k,1}\|^2. \quad (23)$$

If $\xi_{\text{cp}}(\Delta_k; x_k, \nu_k) = 0$, then $s_{k,1} = 0$, and x_k is first-order stationary with Proposition 1. If x_k is not first-order stationary, $s_{k,1} \ne 0$ according to Proposition 2. In this case, Assumption 1, Assumption 2, and (23) lead to

$$\begin{aligned}
|\rho_k - 1| &= \left|\frac{(f+h)(x_k+s_k) - m(s_k; x_k, B_k)}{m(0; x_k, B_k) - m(s_k; x_k, B_k)}\right| \\
&\le \frac{\kappa_{\text{ubd}}(1+\|B_k\|)\|s_k\|_2^2}{\kappa_{\text{mdc}}\xi_{\text{cp}}(\Delta_k; x_k, \nu_k)} \\
&\le \frac{\kappa_{\text{ubd}}(1+\|B_k\|)\beta^2\|s_{k,1}\|_2^2}{\frac{1}{2}\kappa_{\text{mdc}}\alpha^{-1}\Delta_k^{-1}(1+\|B_k\|)\|s_{k,1}\|^2} \\
&= \frac{2\kappa_{\text{ubd}}\beta^2\alpha\Delta_k}{\kappa_{\text{mdc}}}.
\end{aligned}$$

Thus, $\Delta_k \le \Delta_{\text{succ}}$ implies $\rho_k \ge \eta_2$ and iteration k is very successful. □

We set $\Delta_{\min} := \min(\Delta_0, \gamma_1\Delta_{\text{succ}})$, and we observe that $\Delta_k \ge \Delta_{\min}$ for all $k \in \mathbb{N}$. We use $\nu_k^{-1/2}\xi_{\text{cp}}(\Delta_k; x_k, \nu_k)^{1/2}$ as our criticality measure. Let $0 < \epsilon < 1$, and

$$\begin{aligned}
I(\epsilon) &:= \{k \in \mathbb{N} \mid \nu_k^{-1/2}\xi_{\text{cp}}(\Delta_k; x_k; \nu_k)^{1/2} > \epsilon\}, \\
S(\epsilon) &:= \{k \in I(\epsilon) \mid \rho_k \ge \eta_1\}, \\
U(\epsilon) &:= \{k \in I(\epsilon) \mid \rho_k < \eta_1\},
\end{aligned}$$

be the set of iterations, successful iterations, and unsuccessful iterations until the criticality measure drops below ϵ, respectively.

At iteration k of Algorithm 2.1, let σ_k be the number of successful iterations encountered so far:

$$\sigma_k = |\{j = 0, \dots, k \mid \rho_j \ge \eta_1\}|, \quad k \in \mathbb{N}. \quad (24)$$

We introduce an assumption allowing $\{B_k\}$ to be unbounded, as long as it is controlled by σ_k.

Assumption 3. There are constants $\mu_1 > 0$, $\mu_2 > 0$ and $0 \le p < 1$ such that $\max_{0 \le j \le k} \|B_j\| \le \max(\mu_1,\, \mu_2 \sigma_k^p)$ for all $k \in \mathbb{N}$.

Clearly, Assumption 3 allows approximations that grow unbounded, though they must not grow too fast. It reduces to the bounded case when $p = 0$. The role of μ_1 is only to allow sufficiently large B_k in the early iterations without being constrained by σ_k^p. We may now establish a variant of [3, Lemma 3.6] based on Assumption 3.

Lemma 2. *Let Assumption 1 and Assumption 3 be satisfied. Assume that Algorithm 2.1 generates infinitely many successful iterations, that the step size $\nu_k := \alpha\Delta_k/(1 + \|B_k\|(1 + \alpha\Delta_k))$ is selected at each iteration, and that there exists $(f+h)_{\text{low}} \in \mathbb{R}$ such that $(f+h)(x_k) \ge (f+h)_{\text{low}}$ for all $k \in \mathbb{N}$. Let $\epsilon \in (0, 1)$. If either $\mu_1 \ge \mu_2 |S(\epsilon)|^p$, or $\mu_1 < \mu_2 |S(\epsilon)|^p < 1/(1 + \alpha\Delta_{\min})$, then*

$$|S(\epsilon)| \le \max(\mu_1(1 + \alpha^{-1}\Delta_{\min}^{-1}) + \alpha^{-1}\Delta_{\min}^{-1},\, 2\alpha^{-1}\Delta_{\min}^{-1}) \frac{(f+h)(x_0) - (f+h)_{\text{low}}}{\eta_1 \kappa_{\text{mdc}} \epsilon^2} = O(\epsilon^{-2}). \tag{25}$$

Otherwise,

$$|S(\epsilon)| \le \left(2\mu_2 (1 + \alpha^{-1}\Delta_{\min}^{-1}) \frac{(f+h)(x_0) - (f+h)_{\text{low}}}{\eta_1 \kappa_{\text{mdc}} \epsilon^2} \right)^{1/(1-p)} = O(\epsilon^{-2/(1-p)}). \tag{26}$$

Proof. Let $k \in S(\epsilon)$. We have

$$\begin{aligned}
(f+h)(x_k) - (f+h)(x_k + s_k) &\ge \eta_1 \kappa_{\text{mdc}} \xi_{\text{cp}}(\Delta_k; x_k, \nu_k) \\
&\ge \eta_1 \kappa_{\text{mdc}} \nu_k \epsilon^2 \\
&= \eta_1 \kappa_{\text{mdc}} \frac{1}{\alpha^{-1}\Delta_k^{-1} + \|B_k\|(1 + \alpha^{-1}\Delta_k^{-1})} \epsilon^2 \\
&\ge \eta_1 \kappa_{\text{mdc}} \frac{1}{\alpha^{-1}\Delta_{\min}^{-1} + \|B_k\|(1 + \alpha^{-1}\Delta_{\min}^{-1})} \epsilon^2.
\end{aligned}$$

We add together the above inequalities over all $k \in S(\epsilon)$ and use the assumption that $f + h$ is bounded below to obtain

$$\begin{aligned}
(f+h)(x_0) - (f+h)_{\text{low}} &\ge \eta_1 \kappa_{\text{mdc}} \epsilon^2 \sum_{k \in S(\epsilon)} \frac{1}{\alpha^{-1}\Delta_{\min}^{-1} + \|B_k\|(1 + \alpha^{-1}\Delta_{\min}^{-1})} \\
&\ge \eta_1 \kappa_{\text{mdc}} \epsilon^2 |S(\epsilon)| \min_{k \in S(\epsilon)} \frac{1}{\alpha^{-1}\Delta_{\min}^{-1} + \|B_k\|(1 + \alpha^{-1}\Delta_{\min}^{-1})} \\
&= \eta_1 \kappa_{\text{mdc}} \epsilon^2 |S(\epsilon)| \frac{1}{\max_{k \in S(\epsilon)}(\alpha^{-1}\Delta_{\min}^{-1} + \|B_k\|(1 + \alpha^{-1}\Delta_{\min}^{-1}))} \\
&= \eta_1 \kappa_{\text{mdc}} \epsilon^2 |S(\epsilon)| \frac{1}{\alpha^{-1}\Delta_{\min}^{-1} + (\max_{k \in S(\epsilon)} \|B_k\|)(1 + \alpha^{-1}\Delta_{\min}^{-1})} \\
&\ge \eta_1 \kappa_{\text{mdc}} \epsilon^2 |S(\epsilon)| \frac{1}{\alpha^{-1}\Delta_{\min}^{-1} + (\max(\mu_1, \mu_2 |S(\epsilon)|^p))(1 + \alpha^{-1}\Delta_{\min}^{-1})}
\end{aligned} \tag{27}$$

where we appealed to Assumption 3 in the last step.

Firstly, if $\mu_1 \ge \mu_2 |S(\epsilon)|^p$, the denominator of the last inequality can be bounded above by $\mu_1(1 + \alpha^{-1}\Delta_{\min}^{-1}) + \alpha^{-1}\Delta_{\min}^{-1}$. Secondly, if $\mu_1 < \mu_2 |S(\epsilon)|^p < 1/(1 + \alpha\Delta_{\min})$, it can be bounded above by $2\alpha^{-1}\Delta_{\min}^{-1}$. In both cases, it can be bounded above by the constant $\max(\mu_1(1 + \alpha^{-1}\Delta_{\min}^{-1}) + \alpha^{-1}\Delta_{\min}^{-1},\, 2\alpha^{-1}\Delta_{\min}^{-1})$, and

$$(f+h)(x_0) - (f+h)_{\text{low}} \ge \frac{\eta_1 \kappa_{\text{mdc}} \epsilon^2}{\max(\mu_1(1 + \alpha^{-1}\Delta_{\min}^{-1}) + \alpha^{-1}\Delta_{\min}^{-1},\, 2\alpha^{-1}\Delta_{\min}^{-1})} |S(\epsilon)|,$$

which establishes (25).

The last situation occurs when $\mu_2|S(\epsilon)|^p \geq \max(\mu_1,\, 1/(1+\alpha\Delta_{\min}))$. In this case, $\mu_2|S(\epsilon)|^p \geq \mu_1$ so that $\max(\mu_1,\, \mu_2|S(\epsilon)|^p) = \mu_2|S(\epsilon)|^p$, and $\mu_2|S(\epsilon)|^p(1+\alpha^{-1}\Delta_{\min}^{-1}) \geq (1+\alpha^{-1}\Delta_{\min}^{-1})/(1+\alpha\Delta_{\min}) = \alpha^{-1}\Delta_{\min}^{-1}$. By adding $\mu_2|S(\epsilon)|^p(1+\alpha^{-1}\Delta_{\min}^{-1})$ to both sides of the latter inequality and taking its reciprocal, we obtain

$$\frac{1}{2\mu_2|S(\epsilon)|^p(1+\alpha^{-1}\Delta_{\min}^{-1})} \leq \frac{1}{\mu_2|S(\epsilon)|^p(1+\alpha^{-1}\Delta_{\min}^{-1}) + \alpha^{-1}\Delta_{\min}^{-1}}.$$

The above combines with (27) to yield

$$(f+h)(x_0) - (f+h)_{\text{low}} \geq \frac{\eta_1 \kappa_{\text{mdc}} \epsilon^2}{2\mu_2(1+\alpha^{-1}\Delta_{\min}^{-1})}|S(\epsilon)|^{1-p},$$

which establishes (26). □

According to Lemma 2, there are two regimes. In the first, ϵ is large enough that $|S(\epsilon)|^p$ is small, and we recover the worst-case iteration complexity of the bounded Hessian scenario. In the second regime, ϵ is small enough that the number of successful iterations is significant and impacts the complexity bound. For instance, in this regime, we obtain a complexity bound of $O(\epsilon^{-5/2})$ for $p = \frac{1}{5}$ and $O(\epsilon^{-3})$ for $p = \frac{1}{3}$. In other words, the faster the growth of $\|B_k\|$, the worse the deterioration of the complexity bound.

A bound on the number of unsuccessful iteration is obtained using the technique of Cartis et al. [14].

Proposition 5 (3, Lemma 3.7). *Under the assumptions of Lemma 2,*

$$|U(\epsilon)| \leq \log_{\gamma_2}(\Delta_{\min}/\Delta_0) + |S(\epsilon)|\,|\log_{\gamma_2}(\gamma_4)|. \tag{28}$$

Proof. The proof is a minor modification of that of [3, Lemma 3.7]. We provide it for completeness. Let k_ϵ be the smallest integer satisfying $\nu_k^{-1/2}\xi_{\text{cp}}(\Delta_k; x_k, \nu_k)^{1/2} \leq \epsilon$. The update rule of Δ_k in Line 11 indicates that

$$\Delta_{\min} \leq \Delta_{k_\epsilon - 1} \leq \min(\Delta_0 \gamma_2^{|U(\epsilon)|}\gamma_4^{|S(\epsilon)|},\, \Delta_{\max}) \leq \Delta_0 \gamma_2^{|U(\epsilon)|}\gamma_4^{|S(\epsilon)|}.$$

As $0 < \gamma_2 < 1$, we take the logarithm of the above inequalities to obtain

$$|U(\epsilon)|\log(\gamma_2) + |S(\epsilon)|\log(\gamma_4) \geq \log(\Delta_{\min}/\Delta_0),$$

which leads to (28). □

Thus, under Assumption 3, Lemma 2 and Proposition 5 show that $\liminf \nu_k^{-1/2}\xi_{\text{cp}}(\Delta_k; x_k, \nu_k)^{1/2} = 0$.

4 Sharpness of the complexity bound

In this section, we show that the bound of Lemma 2 is attained using the techniques of Cartis et al. [14, Theorem 2.2.3]. To this end, for $0 < \epsilon \leq 1/2$, we explicitly construct $k_\epsilon = \lfloor \epsilon^{-2/(1-p)} \rfloor$ iterates of Algorithm 2.1 with $n = 1$ and $h = 0$, so that $\nu_k^{-1/2}\xi_{\text{cp}}(\Delta_k; x_k, \nu_k)^{1/2} > \epsilon$ for $k = 0, \ldots, k_\epsilon - 1$, and $\nu_{k_\epsilon}^{-1/2}\xi(\Delta_{k_\epsilon}; x_{k_\epsilon}, \nu_{k_\epsilon})^{1/2} = \epsilon$. Then, we invoke [14, Theorem A.9.2] to establish that there exists $f : \mathbb{R} \to \mathbb{R}$ in (1) that interpolates our iterates and satisfies our assumptions. The following result is a special case of [14, Theorem A.9.2].

Proposition 6 (Hermite interpolation with function and gradient evaluations). *Let k_ϵ be a positive integer, $\{f_k\}$, $\{g_k\}$ and $\{x_k\}$ be sequences of numbers given for $k \in \{0,\dots,k_\epsilon\}$. Assume that for $k \in \{0,\dots,k_\epsilon\}$, $s_k = x_{k+1} - x_k > 0$, and that for all $k \in \{0,\dots,k_\epsilon - 1\}$,*

$$|f_{k+1} - (f_k + g_k s_k)| \le \kappa_f s_k^2, \tag{29a}$$
$$|g_{k+1} - g_k| \le \kappa_f s_k, \tag{29b}$$

for some constant $\kappa_f \ge 0$. Then, there exists $f : \mathbb{R} \to \mathbb{R}$ continuously differentiable such that

$$f(x_k) = f_k \quad \text{and} \quad f'(x_k) = g_k.$$

In addition, if

$$|f_k| \le \kappa_f, \quad |g_k| \le \kappa_f \quad \text{and} \quad s_k \le \kappa_f,$$

then $|f|$ and $|f'|$ are bounded by a constant depending only on κ_f.

Proof. The result is a special case of [14, Theorem $A.9.2$] with $p = 1$. □

In the following, we use

$$0 < \epsilon \le 1/2, \tag{30a}$$
$$0 \le p < 1, \tag{30b}$$
$$k_\epsilon = \lfloor \epsilon^{-2/(1-p)} \rfloor, \tag{30c}$$
$$\alpha > 0, \tag{30d}$$
$$\beta \ge 2\alpha^{-1} + 1, \tag{30e}$$

and for all $k \in \{0,\dots,k_\epsilon\}$, we define the sequences

$$w_k := (k_\epsilon - k)/k_\epsilon, \tag{31a}$$
$$g_k := -\epsilon(1 + w_k). \tag{31b}$$

In addition, using the initial values

$$\Delta_0 := 1, \tag{32a}$$
$$B_0 := 1, \tag{32b}$$
$$s_0 := -g_0, \tag{32c}$$
$$x_0 := 0, \tag{32d}$$
$$f_0 := 8\epsilon^2 + \frac{4}{1-p}, \tag{32e}$$

we define, for all $k \in \{1,\dots,k_\epsilon\}$,

$$B_k := k^p, \tag{33a}$$
$$x_k := x_{k-1} + s_{k-1}, \tag{33b}$$
$$f_k := f_{k-1} + g_{k-1} s_{k-1}, \tag{33c}$$

and for all $k \in \{0,\dots,k_\epsilon\}$,

$$s_k := -B_k^{-1} g_k > 0, \tag{34a}$$
$$\nu_k := \frac{1}{\alpha^{-1}\Delta_k^{-1} + |B_k|(1 + \alpha^{-1}\Delta_k^{-1})}. \tag{34b}$$

Sequences (31), (33) and (34) may seem obscure without looking at [14, Theorem 2.2.3]. However, they will make more sense in Theorem 2 below. In particular, we aim to have iterates satisfying the assumptions of Proposition 6, along with $\nu_k^{-1/2}\xi_{\text{cp}}(\Delta_k; x_k, \nu_k)^{1/2} = |g_k| > \epsilon$ for $k \in \{0, \dots, k_\epsilon - 1\}$, and $|g_{k_\epsilon}| = \epsilon$.

First, Lemma 3 establishes bounds on f_k.

Lemma 3. *Using the parameters in (30) and the sequences defined in (31), (33), and (34), the following properties hold for the sequence $\{f_k\}$:*

1. *for all $k \in \{1, \dots, k_\epsilon\}$,*
$$f_k < f_{k-1}, \tag{35}$$
2. *for all $k \in \{0, \dots, k_\epsilon\}$,*
$$0 \le f_0 - f_k \le 4\epsilon^2 \left(2 + \frac{k^{(1-p)}}{1-p}\right) \le 8\epsilon^2 + \frac{4}{1-p}, \tag{36}$$
3. *for all $k \in \{0, \dots, k_\epsilon\}$,*
$$f_k \ge 0. \tag{37}$$

Proof. First, we notice that for all $k \in \{0, \dots, k_\epsilon\}$, $g_k < 0$ and $s_k > 0$. By combining these observations and the definition of f_k, we deduce that $f_k < f_{k-1}$ for all $k \in \{0, \dots, k_\epsilon\}$, and in particular

$$f_0 - f_k \ge 0.$$

Inequalities (36) hold for $k = 0$ and for $k = 1$ because $f_0 - f_1 = -g_0 s_0 = 4\epsilon^2$. For all $k \in \{2, \dots, k_\epsilon\}$,

$$\begin{aligned}
f_0 - f_k &= -\sum_{i=0}^{k-1} g_i s_i \\
&= -g_0 s_0 + \sum_{i=1}^{k-1} g_i^2 i^{-p} \\
&= 4\epsilon^2 + \sum_{i=1}^{k-1} \epsilon^2 (1 + w_i)^2 i^{-p} \\
&= \epsilon^2 \left(4 + \sum_{i=1}^{k-1} (1 + w_i)^2 i^{-p}\right).
\end{aligned}$$

Now,

$$\begin{aligned}
\sum_{i=1}^{k-1} (1 + w_i)^2 i^{-p} &\le \sum_{i=1}^{k-1} 4 i^{-p} && \text{because } 1 + w_i \le 2 \\
&\le 4\left(1 + \sum_{i=2}^{k-1} i^{-p}\right) \\
&\le 4\left(1 + \sum_{i=2}^{k-1} \int_{i-1}^{i} t^{-p} dt\right) && \text{because } i^{-p} = \int_{i-1}^{i} i^{-p} dt \le \int_{i-1}^{i} t^{-p} dt \\
&\le 4\left(1 + \int_{1}^{k-1} t^{-p} dt\right) \\
&\le 4\left(1 + \int_{1}^{k} t^{-p} dt\right)
\end{aligned}$$

$$= 4\left(1 + \frac{k^{1-p} - 1}{1-p}\right)$$
$$\leq 4\left(1 + \frac{k^{1-p}}{1-p}\right).$$

This results in

$$f_0 - f_k \leq 4\epsilon^2 + 4\epsilon^2\left(1 + \frac{k^{1-p}}{1-p}\right) = 8\epsilon^2 + 4\frac{\epsilon^2 k^{1-p}}{1-p}. \tag{38}$$

Finally, since $k \leq k_\epsilon = \lfloor \epsilon^{-2/(1-p)} \rfloor \leq \epsilon^{-2/(1-p)}$, we have, for all $k \leq k_\epsilon$,

$$\epsilon^2 k^{(1-p)} \leq 1. \tag{39}$$

We combine (38) and (39) to obtain (36). The value of f_0 and (36) then allows us to establish (37). □

Now, Lemma 4 establishes a bound for $|g_{k+1} - g_k|$.

Lemma 4. *Using the parameters in (30) and the sequences defined in (31), (32) and (34), we have that, for all $k \in \{0, \ldots, k_\epsilon\}$,*

$$|g_{k+1} - g_k| \leq s_k. \tag{40}$$

Proof. For $k \in \{0, \ldots, k_\epsilon - 1\}$,

$$|g_{k+1} - g_k| = |-\epsilon(1 + w_{k+1}) + \epsilon(1 + w_k)| = \epsilon / k_\epsilon. \tag{41}$$

Since $p < 1$ and $k < k_\epsilon$, we have $k^p / k_\epsilon \leq 1 \leq 1 + w_k$. We multiply the latter inequality by ϵk^{-p} to obtain $\epsilon / k_\epsilon \leq k^{-p}\epsilon(1 + w_k)$, which leads to $|g_{k+1} - g_k| \leq s_k$ using (41). □

The following result uses Lemma 3 and Lemma 4 to apply Proposition 6.

Proposition 7. *Using the parameters in (30) and the sequences defined in (31), (32) and (34), there exists $f : \mathbb{R} \to \mathbb{R}$ differentiable such that*

$$f(x_k) = f_k, \quad f'(x_k) = g_k. \tag{42}$$

Proof. We can see that $s_k > 0$ and, by definition of f_k,

$$|f_{k+1} - (f_k + g_k s_k)| = 0.$$

Lemma 4 shows that

$$|g_{k+1} - g_k| \leq s_k.$$

Using Lemma 3, we know that for all $k \in \{0, \ldots, k_\epsilon\}$, $f_k \geq 0$, and since $\{f_k\}$ is decreasing, we have

$$|f_k| \leq f_0.$$

In addition,

$$|g_k| \leq 2\epsilon \leq 1 \quad \text{and} \quad s_k \leq |g_k| \leq 1.$$

The result follows from Proposition 6. □

For the following lemma, we define the sequence $\{s_{k,1}\}$ such that for all $k \in \{0, \ldots, k_\epsilon\}$,

$$s_{k,1} := -\nu_k g_k. \tag{43}$$

Lemma 5. *Using the parameters in* (30) *and the sequences defined in* (31), (32) *and* (34)*, we establish that, for all* $k \in \{0, \dots, k_\epsilon\}$,

$$|s_k| \le \min(\Delta_k, \beta |s_{k,1}|). \tag{44}$$

Proof. On the one hand, we have

$$|s_k| = \epsilon \frac{(1 + w_k)}{B_k} \le 2\epsilon \le 1 \le \Delta_k. \tag{45}$$

On the other hand, since $B_k^{-1} \le 1$ and $\Delta_k \ge 1$,

$$2\alpha^{-1} + 1 \ge \alpha^{-1}\Delta_k^{-1}(B_k^{-1} + 1) + 1,$$

so that

$$1 \le \frac{2\alpha^{-1} + 1}{\alpha^{-1}\Delta_k^{-1}(B_k^{-1} + 1) + 1} \le \frac{\beta}{\alpha^{-1}\Delta_k^{-1}(B_k^{-1} + 1) + 1}.$$

We multiply the above inequality by B_k^{-1} to obtain

$$B_k^{-1} \le \frac{\beta B_k^{-1}}{\alpha^{-1}\Delta_k^{-1}(B_k^{-1} + 1) + 1} = \frac{\beta}{\alpha^{-1}\Delta_k^{-1} + B_k(1 + \alpha^{-1}\Delta_k^{-1})} = \beta \nu_k^{-1},$$

and, by multiplying by $|g_k|$, we deduce that

$$|s_k| = B_k^{-1}|g_k| \le \beta \nu_k^{-1}|g_k| = \beta |s_{k,1}|. \tag{46}$$

We combine (45) and (46) to obtain (44). □

The following theorem finally establishes the main result of this section.

Theorem 2 (Slow convergence of Algorithm 2.1)**.** *Algorithm 2.1 applied to* (1) *with model* m_k *satisfying* Model Assumption 2.1, Assumption 1, Assumption 2 *and using Hessian approximations* $\{B_k\}$ *satisfying* Assumption 3 *may require as many as* $O(\epsilon^{-2/(1-p)})$ *iterations to produce an iterate* x_{k_ϵ} *such that*

$$\nu_{k_\epsilon}^{-1/2} \xi_{\text{cp}}(\Delta_{k_\epsilon}; x_{k_\epsilon}, \nu_{k_\epsilon})^{1/2} \le \epsilon. \tag{47}$$

Proof. The proof consists in constructing $f : \mathbb{R} \to \mathbb{R}$ by interpolation, as in [14, Theorem 2.2.3]. Let $n = 1$, $h = 0$, $\ell = -\infty$, $u = +\infty$. We use the parameters in (30) and the sequences defined in (31), (32) and (34). We invoke Proposition 7 to obtain $f : \mathbb{R} \to \mathbb{R}$ differentiable and bounded such that $f(x_k) = f_k$ and $f'(x_k) = g_k$. Our goal is to show that $\{x_k\}$, $\{s_k\}$, $\{f_k\}$ and $\{g_k\}$ satisfy all our assumptions and are generated by Algorithm 2.1 applied to f with $x_0 = 0$ and with the special value of $\{B_k\}$ in (32b) and (33a).

We proceed by choosing $0 \le k \le k_\epsilon$ such that $\Delta_k \ge 1$, which holds at least for $k = 0$, and going through the steps of Algorithm 2.1 at iteration k to check that it generates the iterates defined in (31), (32) and (34).

In Line 5, ν_k in (34b) is as large as allowed.

In Line 6, Lemma 1 indicates that $s_{k,1}$ in (43) is a global minimizer of (12b) with $\psi = 0$. As $1 + w_k \le 2$ and $|B_k| \ge 1$, we observe that

$$|s_{k,1}| = |\nu_k g_k| = \frac{\epsilon(1 + w_k)}{\alpha^{-1}\Delta_k^{-1} + |B_k|(1 + \alpha^{-1}\Delta_k^{-1})} \le 2\epsilon \le 1 \le \Delta_k,$$

which implies that $s_{k,1}$ is a solution of (13) because the condition $|s_{k,1}| \le \Delta_k$ is already satisfied.

In Line 8, let $m_k(\cdot; x_k, B_k)$ be defined as in (11). $m_k(\cdot; x_k, B_k)$ satisfies Model Assumption 2.1, and using Lemma 1, we have that s_k in (34a) with $\psi = 0$ and $B = B_k$ is its global minimizer. Lemma 5 shows that

$$|s_k| \le \min(\Delta_k, \beta |s_{k,1}|),$$

which also implies that s_k is a solution of (11).

In Line 9, we compute

$$\begin{aligned}\rho_k &= \frac{f_k - f_{k+1}}{m(0; x_k, B_k) - m(s_k; x_k, B_k)} \\ &= \frac{f_k - f_{k+1}}{f_k - f_k - g_k s_k - B_k s_k^2/2} \\ &= \frac{f_k - f_{k+1}}{g_k^2 B_k^{-1}/2} \\ &= \frac{-g_k s_k}{g_k^2 B_k^{-1}/2} \\ &= \frac{B_k^{-1} g_k^2}{g_k^2 B_k^{-1}/2} \\ &= 2.\end{aligned} \tag{48}$$

In Line 10, $\rho_k = 2$ implies that $x_{k+1} = x_k + s_k$, and in Line 11, we can set $\Delta_{k+1} = \min(\gamma_3 \Delta_k, \Delta_{\max}) \ge \Delta_k \ge 1$.

Now, either $\nu_k^{-1/2} \xi_{\mathrm{cp}}(\Delta_k; x_k, \nu_k)^{1/2} > \epsilon$, and we perform the next iteration of Algorithm 2.1, or $\nu_k^{-1/2} \xi_{\mathrm{cp}}(\Delta_k; x_k, \nu_k)^{1/2} \le \epsilon$, which stops the algorithm. We have shown that $s_{k,1}$ is a solution of (13), thus

$$\xi_{\mathrm{cp}}(\Delta_k; x_k, \nu_k) = f_k - (f_k + g_k s_{k,1}) = -g_k s_{k,1} = \nu_k g_k^2, \tag{49}$$

and

$$\nu_k^{-1/2} \xi_{\mathrm{cp}}(\Delta_k; x_k, \nu_k)^{1/2} = |g_k|. \tag{50}$$

Therefore, for all $k \in \{0, \dots, k_\epsilon - 1\}$, $\nu_k^{-1/2} \xi_{\mathrm{cp}}(\Delta_k; x_k, \nu_k)^{1/2} > \epsilon$, and $\nu_{k_\epsilon}^{-1/2} \xi_{\mathrm{cp}}(\Delta_{k_\epsilon}; x_{k_\epsilon}, \nu_{k_\epsilon})^{1/2} = \epsilon$, so that Algorithm 2.1 performs exactly k_ϵ iterations to generate x_{k_ϵ} satisfying (47).

To finish the proof, we must verify that Assumption 1, Assumption 2 and Assumption 3 hold. Assumption 1 is satisfied thanks to Proposition 4. Assumption 2 is satisfied with $\kappa_{\mathrm{ubd}} = \frac{1}{2}$ because

$$|f_{k+1} - m(s_k; x_k, B_k)| = |f_{k+1} - f_k - g_k s_k - \tfrac{1}{2} B_k s_k^2| = \tfrac{1}{2} B_k s_k^2 \le \tfrac{1}{2}(1 + B_k) s_k^2.$$

Finally, our choice of B_k allows Assumption 3 to be satisfied because all iterations are successful and $\sigma_k = k$. □

5 Numerical verification of the bound

We construct $f : \mathbb{R} \to \mathbb{R}$ satisfying the properties of the function in the proof of Theorem 2. The construction follows the formula used in the proof of [14, Theorem A.9.2], and we use similar notation.

We use again the parameters (30), and the sequences (31)–(34). Define the cubic Hermite interpolant

$$\pi_k(\tau) := c_{k,0} + c_{k,1}\tau + c_{k,2}\tau^2 + c_{k,3}\tau^3, \tag{51}$$

where, for all $k \in \{0, \dots, k_\epsilon\}$, $c_{k,0} = f_k$, $c_{k,1} = g_k$, and $c_{k,2}$, $c_{k,3}$ solve

$$\begin{bmatrix} s_k^2 & s_k^3 \\ 2s_k & 3s_k^2 \end{bmatrix} \begin{bmatrix} c_{k,2} \\ c_{k,3} \end{bmatrix} = \begin{bmatrix} f_{k+1} - (f_k + g_k s_k) \\ g_{k+1} - g_k \end{bmatrix} = \begin{bmatrix} 0 \\ g_{k+1} - g_k \end{bmatrix}. \tag{52}$$

We use the additional conditions $f_{-1} = f_0$, $g_{-1} = 0$, $f_{k_\epsilon+1} = f_{k_\epsilon}$, $g_{k_\epsilon+1} = g_{k_\epsilon}$, and $x_{-1} = -s_{-1}$, where $s_{-1} = 1$, which allows (29) to hold with $\kappa_f = 1$, because $|f_0 - (f_{-1} + g_{-1}s_{-1})| = 0$, and $|g_0 - g_{-1}| = |g_0| = \epsilon(1 + w_0) = 2\epsilon \le 1 = s_{-1}$ since $\epsilon \le 1/2$. Finally,

$$f(x) \coloneqq \begin{cases} f_0 & \text{if } x \le x_{-1} \\ \pi_k(x - x_k) & \text{if } x \in (x_k; x_{k+1}] \text{ for } k \in \{-1, \dots, k_\epsilon\} \\ f_{k_\epsilon} & \text{if } x > x_{k_\epsilon} + s_{k_\epsilon}. \end{cases} \tag{53}$$

By construction, f is a piecewise polynomial of degree 3. We have $\pi_k(0) = f_k$, $\pi_k'(0) = g_k$, $\pi_k(s_k) = f_{k+1}$ thanks to the definition of f in (33c) and the first line of (52), and $\pi_k'(s_k) = g_{k+1}$ with the second line of (52). Thus, $f : \mathbb{R} \to \mathbb{R}$ is continuously differentiable over $(x_{-1}, x_{k_\epsilon+1})$.

We minimize f using Algorithm 2.1 as implemented in [5], without nonsmooth regularizer, and with starting point $x_0 = 0$. Inside TR, we set $B_k = k^p$ so that $\{B_k\}$ grows unbounded and Assumption 3 holds, because $\rho_k = 2$ in (48) so that all iterations are very successful. In Line 8, we use the analytical solution $s_k = -B_k^{-1}\nabla f(x_k)$ of (17) given by Lemma 1 in order to avoid rounding errors occurring in a subproblem solver for (15). This expression of s_k satisfies the trust-region constraint by construction thanks to Lemma 5. The modified TR implementation is available from `https://github.com/geoffroyleconte/RegularizedOptimization.jl/tree/unbounded`.

We set $p = 1/10$, $\alpha = \beta = 10^{+16}$, $\gamma_3 = 3$, $\Delta_{\max} = 10^3$ and $\epsilon = 1/10$, so that $k_\epsilon = 166$. We observe that TR converges in precisely 166 iterations. With $\epsilon = 1/20$, we obtain the convergence of TR in precisely $k_\epsilon = 778$ iterations.

In order to make the oscillations of f' clearly visible, Figure 2 shows plots of f and f' over $[0, x_{k_\epsilon+1}]$ with $\epsilon = 1/3$. Table 1 shows the theoretical values of $\nu_k^{-1/2}\xi_{\text{cp}}(\Delta_k; x_k, \nu_k)^{1/2} = |g_k|$ according to (50). TR converges in 11 iterations and produces the logs in Figure 1 that align with these theoretical values. Note that $\rho_k = 2$, as predicted by (48), and therefore, that each iteration is successful.

Table 1: Rounded theoretical values of $\nu_k^{-1/2}\xi_{\text{cp}}(\Delta_k; x_k, \nu_k)^{1/2}$ for $\epsilon = 1/3$.

k	0	1	2	3	4	5	6	7	8	9	10	11
$\nu_k^{-1/2}\xi_{\text{cp}}(\Delta_k; x_k, \nu_k)^{1/2}$	0.67	0.64	0.61	0.58	0.55	0.52	0.48	0.45	0.42	0.39	0.36	0.33

```
outer    inner      f(x)      h(x) √ξcp/√ν      √ξ         ρ        Δ      ‖x‖      ‖s‖     ‖B_k‖
    1        1  5.3e+00  0.0e+00 6.7e-01 4.7e-01  2.0e+00 1.0e+00 0.0e+00 6.7e-01 1.0e+00
    2        1  4.9e+00  0.0e+00 6.4e-01 4.5e-01  2.0e+00 3.0e+00 6.7e-01 6.4e-01 1.0e+00
    3        1  4.5e+00  0.0e+00 6.1e-01 4.1e-01  2.0e+00 9.0e+00 1.3e+00 5.7e-01 1.1e+00
    4        1  4.1e+00  0.0e+00 5.8e-01 3.9e-01  2.0e+00 2.7e+01 1.9e+00 5.2e-01 1.1e+00
    5        1  3.8e+00  0.0e+00 5.5e-01 3.6e-01  2.0e+00 8.1e+01 2.4e+00 4.7e-01 1.1e+00
    6        1  3.6e+00  0.0e+00 5.2e-01 3.4e-01  2.0e+00 2.4e+02 2.9e+00 4.4e-01 1.2e+00
    7        1  3.4e+00  0.0e+00 4.8e-01 3.1e-01  2.0e+00 7.3e+02 3.3e+00 4.1e-01 1.2e+00
    8        1  3.2e+00  0.0e+00 4.5e-01 2.9e-01  2.0e+00 1.0e+03 3.7e+00 3.7e-01 1.2e+00
    9        1  3.0e+00  0.0e+00 4.2e-01 2.7e-01  2.0e+00 1.0e+03 4.1e+00 3.4e-01 1.2e+00
   10        1  2.8e+00  0.0e+00 3.9e-01 2.5e-01  2.0e+00 1.0e+03 4.4e+00 3.2e-01 1.2e+00
   11        1  2.7e+00  0.0e+00 3.6e-01 2.3e-01  2.0e+00 1.0e+03 4.7e+00 2.9e-01 1.3e+00
   12        1  2.6e+00  0.0e+00 3.3e-01                   1.0e+03 5.0e+00 2.6e-01 1.3e+00
TR: terminating with √ξcp/√ν = 0.3333333333333333
"Execution stats: first-order stationary"
```

Figure 1: TR logs with $\epsilon = 1/3$. *outer* denotes the iteration number, *inner* is the number of iterations performed by the subsolver to solve (15) with the model in (11), $\sqrt{\xi cp}/\sqrt{\nu}$ is $\nu_k^{-1/2}\xi_{\text{cp}}(\Delta_k; x_k, \nu_k)^{1/2}$, $\sqrt{\xi}$ is the numerator of (16), $\|s\|$ is $\|s_k\|$, and the remaining columns refer unambiguously to data used in Algorithm 2.1.

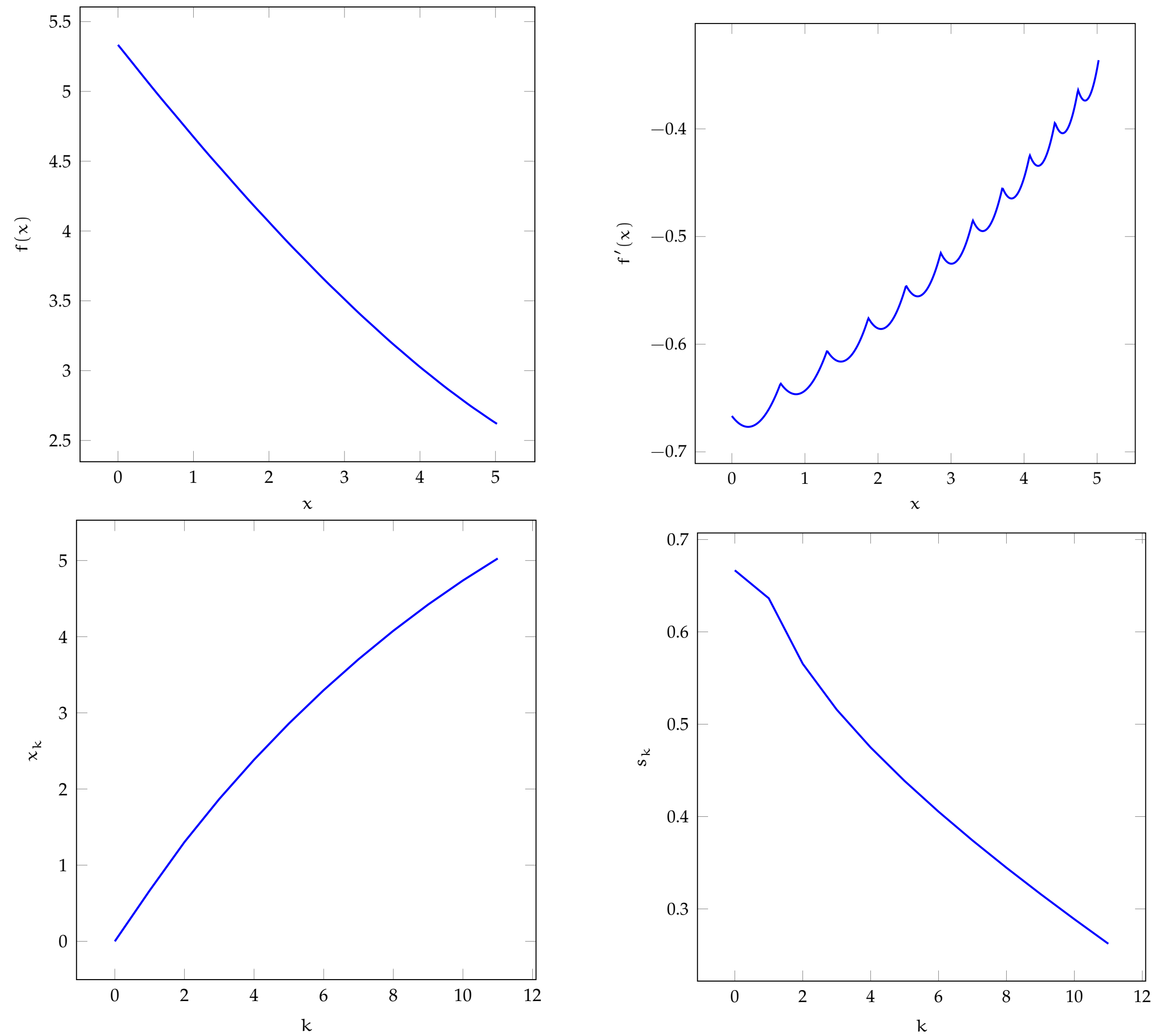

Figure 2: Illustration of example (53) **with** $\epsilon = 1/3$. **Top row: values of** f **(left) and of** f' **(right) for** $x \in [0, x_{k_\epsilon+1}]$. **Bottom row: iterates** x_k **(left) and steps** s_k **(right) for** $k \in [0, k_\epsilon + 1]$.

The code to run this experiment is available at `https://github.com/geoffroyleconte/docGL/blob/master/regularized-opt/test-unbounded-hess.jl`. By making similar changes to the algorithm TRDH [26], which can be found at the same URL, we obtain the same number of iterations.

6 Discussion

We have shown that it is possible to establish convergence and sharp worst-case evaluation complexity of Algorithm 2.1 in the presence of unbounded Hessian approximations B_k, provided they do not grow too fast—c.f., Assumption 3. We established that the complexity bound can be attained, and we gave an example of function for which it was attained, both theoretically and numerically.

When $p \geq 1$ in Assumption 3 or the growth of $\|B_k\|$ is not governed by the number of successful iterations, it may still be possible to establish convergence in the sense that $\liminf \nu_k^{-1/2} \xi_{\text{cp}}(\Delta_k; x_k, \nu_k) = 0$

as in [15, §8.4.1.2], where the main assumption is that

$$\sum_{k=0}^{\infty} \frac{1}{1 + \max_{0 \le j \le k} \|B_j\|} = \infty.$$

However, it is unclear at the time of this writing whether a sharp worst-case evaluation complexity bound holds for such more general cases.

A possible extension of the present work would be to analyze the worst-case evaluation complexity of ARp-type methods in the presence of potentially unbounded model Hessians.

References

[1] A. Aravkin, R. Baraldi, and D. Orban. A Levenberg-Marquardt method for nonsmooth regularized least squares. Cahier du GERAD G–2023–58, GERAD, Montréal, QC, Canada, 2022.

[2] A. Aravkin, R. Baraldi, G. Leconte, and D. Orban. Corrigendum: A proximal quasi-Newton trust-region method for nonsmooth regularized optimization. Cahier du GERAD G–2021–12-SM, GERAD, Montréal QC, Canada, 2023.

[3] A. Y. Aravkin, R. Baraldi, and D. Orban. A proximal quasi-Newton trust-region method for nonsmooth regularized optimization. SIAM J. Optim., 32(2):900–929, 2022.

[4] R. Baraldi and D. P. Kouri. A proximal trust-region method for nonsmooth optimization with inexact function and gradient evaluations. Math. Program., 201(1):559–598, 2022.

[5] R. Baraldi and D. Orban. RegularizedOptimization.jl: Algorithms for regularized optimization. https://github.com/JuliaSmoothOptimizers/RegularizedOptimization.jl, February 2022.

[6] J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math. Program., 146(1):459–494, 2014.

[7] R. G. Carter. Safeguarding hessian approximations in trust region algorithms. Technical Report TR87-12, Department of Computational and Applied Mathematics, Rice University, Houston, TX, USA, 1987.

[8] C. Cartis, N. I. M. Gould, and Ph. L. Toint. On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization problems. SIAM J. Optim., 20 (6):2833–2852, 2010.

[9] C. Cartis, N. I. M. Gould, and Ph. L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part II: Worst-case function- and derivative-evaluation complexity. Math. Program., 130 (2):295–319, 2011.

[10] C. Cartis, N. I. M. Gould, and Ph. L. Toint. On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming. SIAM J. Optim., 21(4):1721–1739, 2011.

[11] C. Cartis, N. I. M. Gould, and Ph. L. Toint. Complexity bounds for second-order optimality in unconstrained optimization. J. Complexity, 28(1):93–108, 2012.

[12] C. Cartis, N. I. M. Gould, and Ph. L. Toint. Sharp worst-case evaluation complexity bounds for arbitrary-order nonconvex optimization with inexpensive constraints. SIAM J. Optim., 30(1):513–541, 2020.

[13] C. Cartis, N. I. M. Gould, and Ph. L. Toint. Strong evaluation complexity bounds for arbitrary-order optimization of nonconvex nonsmooth composite functions. arXiv preprint arXiv:2001.10802, 2020.

[14] C. Cartis, N. I. M. Gould, and Ph. L. Toint. Evaluation Complexity of Algorithms for Nonconvex Optimization. Number 30 in MOS-SIAM Series on Optimization. SIAM, Philadelphia, USA, 2022.

[15] A. R. Conn, N. I. M. Gould, and Ph. L. Toint. Trust-Region Methods. Number 1 in MOS-SIAM Series on Optimization. SIAM, Philadelphia, USA, 2000.

[16] F. E. Curtis, D. P. Robinson, and M. Samadi. A trust region algorithm with a worst-case iteration complexity of $O(\epsilon^{-3/2})$ for nonconvex optimization. Math. Program., 162(1):1–32, 2017.

[17] J. Dennis, S. Li, and R. Tapia. A unified approach to global convergence of trust region methods for nonsmooth optimization. Math. Program., (68):319—-346, 1995.

[18] J. E. Dennis, Jr. and J. J. Moré. Quasi-Newton methods, motivation and theory. SIAM Rev., 19(1): 46–89, 1977.

[19] J.-P. Dussault, T. Migot, and D. Orban. Scalable adaptive cubic regularization methods. Math. Program., 2023.

[20] A. V. Fiacco and G. P. McCormick. Nonlinear Programming: Sequential Unconstrained Minimization Techniques. J. Wiley and Sons, Chichester, England, 1968. Reprinted as Classics in Applied Mathematics, SIAM, Philadelphia, USA, 1990.

[21] R. Fletcher. An algorithm for solving linearly constrained optimization problems. Math. Program., (2): 133–165, 1972.

[22] A. Forsgren, P. E. Gill, and M. H. Wright. Interior methods for nonlinear optimization. SIAM Rev., 44 (4):525–597, 2002.

[23] M. Fukushima and H. Mine. A generalized proximal point algorithm for certain non-convex minimization problems. International Journal of Systems Science, 12(8):989–1000, 1981.

[24] G. N. Grapiglia, J. Yuan, and Y. Yuan. Nonlinear stepsize control algorithms: Complexity bounds for first- and second-order optimality. J. Optim. Theory and Applics., (171):980—997, 2016.

[25] D. Kim, S. Sra, and I. S. Dhillon. A scalable trust-region algorithm with application to mixed-norm regression. In ICML, pages 519–526, 2010.

[26] G. Leconte and D. Orban. The indefinite proximal gradient method. Cahier du GERAD G–2023–37, GERAD, Montréal QC, Canada, 2023.

[27] P.-L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. SIAM J. Numer. Anal., 16(6):964–979, 1979.

[28] S. Lotfi, T. Bonniot de Ruisselet, D. Orban, and A. Lodi. Stochastic damped L-BFGS with controlled norm of the Hessian approximation. 2020. OPT2020 Conference on Optimization for Machine Learning.

[29] J. M. Martínez and A. C. Moretti. A trust region method for minimization of nonsmooth functions with linear constraints. Math. Program., (76):431–449, 1997.

[30] J. M. Martínez and M. Raydan. Cubic-regularization counterpart of a variable-norm trust-region method for unconstrained minimization. Journal of Global Optimization, 68(2):367–385, 2017.

[31] Y. Nesterov and A. Nemirovskii. Interior-Point Polynomial Algorithms in Convex Programming. SIAM, Philadelphia, USA, 1994.

[32] Y. Nesterov and B. Polyak. Cubic regularization of Newton method and its global performance. Math. Program., 108(1):177–205, 2006.

[33] M. J. D. Powell. A new algorithm for unconstrained optimization. In J. B. Rosen, O. L. Mangasarian, and K. Ritter, editors, Nonlinear Programming, pages 31–65. Academic Press, 1970.

[34] M. J. D. Powell. Convergence properties of a class of minimization algorithms. In O. L. Mangasarian, R. R. Meyer, and S. M. Robinson, editors, Nonlinear Programming 2, pages 1–27. Academic Press, 1975.

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

[36] L. Qi and J. Sun. A trust region algorithm for minimization of locally Lipschitzian functions. Math. Program., (66):25–43, 1994.

[37] R. Rockafellar and R. Wets. Variational Analysis, volume 317. Springer Verlag, 1998.

[38] C. W. Royer and S. J. Wright. Complexity analysis of second-order line-search algorithms for smooth nonconvex optimization. SIAM J. Optim., 28(2):1448–1477, 2018.

[39] Ph. L. Toint. Global convergence of a class of trust-region methods for nonconvex minimization in Hilbert space. IMA J. Numer. Anal., 8(2):231–252, 04 1988.

[40] Y.-X. Yuan. Conditions for convergence of trust region algorithms for nonsmooth optimization. Math. Program., (31):220–228, 1985.