

Complexity analysis of trust-region methods under (α, L_0, L_1) -smoothness

Y. Diouane, M. L. Habiboullah, A. Khouna, D. Orban

G-2026-56

September 2026

La collection *Les Cahiers du GERAD* est constituée des travaux de recherche menés par nos membres. La plupart de ces documents de travail a été soumis à des revues avec comité de révision. Lorsqu'un document est accepté et publié, le pdf original est retiré si c'est nécessaire et un lien vers l'article publié est ajouté.

Citation suggérée : Y. Diouane, M. L. Habiboullah, A. Khouna, D. Orban (Septembre 2026). Complexity analysis of trust-region methods under (α, L_0, L_1) -smoothness, Rapport technique, Les Cahiers du GERAD G-2026-56, GERAD, HEC Montréal, Canada.

Avant de citer ce rapport technique, veuillez visiter notre site Web (<https://www.gerad.ca/fr/papers/G-2026-56>) afin de mettre à jour vos données de référence, s'il a été publié dans une revue scientifique.

The series *Les Cahiers du GERAD* consists of working papers carried out by our members. Most of these pre-prints have been submitted to peer-reviewed journals. When accepted and published, if necessary, the original pdf is removed and a link to the published article is added.

Suggested citation: Y. Diouane, M. L. Habiboullah, A. Khouna, D. Orban (September 2026). Complexity analysis of trust-region methods under (α, L_0, L_1) -smoothness, Technical report, Les Cahiers du GERAD G-2026-56, GERAD, HEC Montréal, Canada.

Before citing this technical report, please visit our website (<https://www.gerad.ca/en/papers/G-2026-56>) to update your reference data, if it has been published in a scientific journal.

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, 2026
– Bibliothèque et Archives Canada, 2026

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, 2026
– Library and Archives Canada, 2026

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

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

Complexity analysis of trust-region methods under (α, L_0, L_1) -smoothness

Youssef Diouane ^a

Mohamed L. Habiboullah ^a

Awa Khouna ^b

Dominique Orban ^a

^a GERAD & Department of Mathematics and Industrial Engineering, Polytechnique Montréal, Montréal (Qc), Canada, H3T 1J4

^b CIRRELT & SCALE-AI Chair in Data-Driven Supply Chains, Department of Mathematics and Industrial Engineering, Polytechnique Montréal, Montréal (Qc), Canada, H3T 1J4

youssef.diouane@polymtl.ca

mohamed-laghdaf-2.habiboullah@polymtl.ca

awa.khouna@polymtl.ca

dominique.orban@gerad.ca

September 2026

Les Cahiers du GERAD

G–2026–56

Copyright © 2026 Diouane, Habiboullah, Khouna, Orban

Les textes publiés dans la série des rapports de recherche *Les Cahiers du GERAD* n'engagent que la responsabilité de leurs auteurs. Les auteurs conservent leur droit d'auteur et leurs droits moraux sur leurs publications et les utilisateurs s'engagent à reconnaître et respecter les exigences légales associées à ces droits. Ainsi, les utilisateurs:

- Peuvent télécharger et imprimer une copie de toute publication du portail public aux fins d'étude ou de recherche privée;
- Ne peuvent pas distribuer le matériel ou l'utiliser pour une activité à but lucratif ou pour un gain commercial;
- Peuvent distribuer gratuitement l'URL identifiant la publication.

Si vous pensez que ce document enfreint le droit d'auteur, contactez-nous en fournissant des détails. Nous supprimerons immédiatement l'accès au travail et enquêterons sur votre demande.

The authors are exclusively responsible for the content of their research papers published in the series *Les Cahiers du GERAD*. Copyright and moral rights for the publications are retained by the authors and the users must commit themselves to recognize and abide the legal requirements associated with these rights. Thus, users:

- May download and print one copy of any publication from the public portal for the purpose of private study or research;
- May not further distribute the material or use it for any profit-making activity or commercial gain;
- May freely distribute the URL identifying the publication.

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Abstract : Generalized smoothness assumptions have attracted growing attention in recent years, motivated in part by machine learning problems in which the gradient of the objective function may not be Lipschitz continuous. Among the most prominent of these is the (α, L_0, L_1) -smoothness assumption. Existing methods that attain the best known complexity bounds require knowledge of, or upper bounds on, α , L_0 and L_1 , while the few parameter-agnostic methods available do not recover those bounds. In this paper, we show that trust-region methods attain the best known bounds for (α, L_0, L_1) -smooth objective functions without prior knowledge of these parameters. Establishing these results requires new analytical tools, beyond those used in classical trust-region complexity analyses. We further show that our working model assumption allows the use of general model Hessian approximations and, in particular, accommodates limited-memory quasi-Newton updates. Finally, we show that our complexity bound is sharp in the nonconvex setting.

Keywords : Complexity analysis; quasi-Newton methods; trust-region methods; generalized smoothness; unconstrained optimization; nonconvex optimization; convex optimization

Résumé : Les hypothèses de régularité généralisée suscitent un intérêt croissant depuis quelques années, notamment en raison de problèmes d'apprentissage automatique dans lesquels le gradient de la fonction objectif peut ne pas être lipschitzien. Parmi les plus en vue figure l'hypothèse de (α, L_0, L_1) -régularité. Les méthodes existantes qui atteignent les meilleures bornes de complexité connues exigent la connaissance de α , L_0 et L_1 , ou de bornes supérieures sur ces paramètres. À l'inverse, les rares méthodes qui s'en passent (dites agnostiques) n'atteignent pas ces bornes. Dans cet article, nous montrons que les méthodes de région de confiance atteignent les meilleures bornes connues pour les fonctions objectif (α, L_0, L_1) -régulières, sans connaître ces paramètres à l'avance. Pour établir ces résultats, il faut de nouveaux outils d'analyse, qui vont au-delà de ceux des analyses de complexité classiques des régions de confiance. Nous montrons aussi que notre hypothèse sur le modèle permet d'utiliser des approximations générales du hessien du modèle, et en particulier des mises à jour quasi-Newton à mémoire limitée. Enfin, nous montrons que notre borne de complexité est optimale (atteinte) dans le cas non convexe.

Mots clés : Analyse de complexité ; méthodes quasi-newtoniennes ; méthodes des régions de confiance ; régularité généralisée ; optimisation sans contraintes ; optimisation non convexe ; optimisation convexe

Acknowledgements: Research of R. Diouane and O. Orban is partially supported by an NSERC Discovery Grant. Research of M. L. Habiboullah partially supported by an FRQNT scholarship.

1 Introduction

We consider the unconstrained optimization problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x), \quad (1)$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is $\mathcal{C}^1$ and bounded below by f_{low} . We target to solve (1) by means of a trust-region method [9]. Given a tolerance $\epsilon > 0$, our goal is to compute an ϵ -approximate first-order stationary point x satisfying $\|\nabla f(x)\| \leq \epsilon$. The iteration complexity of an optimization method is the worst-case number of iterations required to reach such a point, while its evaluation complexity counts the associated objective and derivative evaluations. In this paper, we study the evaluation complexity of trust-region methods under generalized smoothness assumptions introduced by [8, 22], which we refer to as (α, L_0, L_1) -smoothness. Specifically, assuming that the Hessian of f exists, we suppose that there exist constants $L_0, L_1 \geq 0$ and $\alpha \in [0, 1]$ such that

$$\|\nabla^2 f(x)\| \leq L_0 + L_1 \|\nabla f(x)\|^\alpha, \quad \forall x \in \mathbb{R}^n. \quad (2)$$

The endpoint $\alpha = 1$ is commonly referred to as (L_0, L_1) -smoothness and has been studied extensively [10, 12, 14, 20–22].

Classical worst-case complexity analyses for smooth optimization usually rely on global Lipschitz continuity of the gradient, a convenient assumption that yields the standard sharp $\mathcal{O}(L\epsilon^{-2})$ complexity bound for nonconvex f , where L is the Lipschitz constant of the gradient [7, 11]. However, L may be large in practice, which can lead to a pessimistic bound. Condition (2) generalizes the classical setting, since ∇f is L -Lipschitz if f is, e.g., $(0, 0, L)$ -smooth. In the training of long short-term memory networks, L_0 is observed to be much smaller than L [21, 22], and other machine learning problems exhibit the same behavior [4]. These observations have motivated the study of complexity bounds under (α, L_0, L_1) -smoothness.

Contributions

This paper makes the following contributions:

- We develop a complexity analysis of trust-region methods under (α, L_0, L_1) -smoothness for nonconvex, convex, and strongly convex objectives. Unlike the classical analysis, our approach requires new mathematical tools to recover the best-known complexity bounds.
- We recover the best known complexity bounds in the nonconvex, convex, and strongly convex settings, matching the state of the art without requiring prior knowledge of (α, L_0, L_1) , except for those of accelerated methods in the convex case.
- We introduce a general class of (α, L_0, L_1) -smooth functions, which we call polynomial inverse problems and which includes phase retrieval as a special case. To the best of our knowledge, the literature contains a single explicit example for $\alpha \in (0, 1)$, namely phase retrieval itself.
- We show how limited-memory quasi-Newton (L-BFGS, L-SR1 and L-PSB) fit within our framework.
- We show that the $\epsilon^{\alpha-2}$ complexity bound is sharp for $(\alpha, 0, L_1)$ -smooth nonconvex objectives.

Related research

The $(1, L_0, L_1)$ -smoothness condition was introduced by Zhang et al. [22] to explain the empirical effectiveness of gradient clipping when the local smoothness varies significantly. They showed that clipping attains the best known rate, whereas fixed-step gradient descent requires the additional assumption that the gradient be uniformly bounded, and pays for it with a constant that degrades with that bound (Table 1). Zhang et al. [21] extended the condition to merely continuously differentiable functions,

and Koloskova et al. [14] allowed an arbitrary clipping radius while adding convex and strongly convex guarantees, at the cost of an additional Lipschitz assumption on the gradient. Gorbunov et al. [10] sharpened the strongly convex bound and combined clipping with acceleration in the convex case, at the price of a factor exponential in L_1 , which Vankov et al. [20] subsequently removed while identifying stepsizes that attain the best known rates for clipped, normalized and accelerated gradient methods (Table 1). In every one of these methods, the stepsize or the clipping radius is prescribed from L_0 and L_1 .

The general condition (2) for $\alpha \in (0, 1)$ was introduced by Chen et al. [8], whose normalized gradient descent is optimal at $\alpha = 1$, but whose bound for $\alpha < 1$ carries a factor that tends to infinity as $\alpha \uparrow 1$ (Table 2). Li et al. [15] works under a condition that strictly generalizes (2), at the cost of rates that remain suboptimal even at $\alpha = 1$, while Tyurin [19] attains the best known nonconvex and convex rates for $\alpha \leq 1$ through an integral-based stepsize rule, with no guarantee in the strongly convex case (Table 2). Both require knowledge of α , L_0 and L_1 .

Hessian approximations have also been considered: the Gradient-Normalized Smoothness framework of Semenov et al. [17] attains the best known nonconvex and convex rates for $(1, L_0, L_1)$ -smooth problems, but still selects its stepsize by grid search or adaptive tuning over (L_0, L_1) .

A few works dispense with prior knowledge of L_0 and L_1 , at the cost of either a worse rate or another requirement. In the nonconvex case, and only for $\alpha = 1$, the backtracking gradient method of Hübler et al. [12] is not optimal: its rate involves a uniform bound on the gradient norm established in the analysis (Table 1). Recently, Jerad [13] proposed a modified variant of the adaptive regularization method $\text{AR}p$, in which the gradient is evaluated at every trial point and only exact derivatives are allowed. Similarly to the trust-region method, their framework does not require prior knowledge of L_0 or L_1 , but yields a worse complexity bound than ours when $p = 1$ (Table 1). In the convex case, Vankov et al. [20] also analyze two methods that retain the best known rate without L_0 and L_1 , but substitute another quantity for them: a Polyak stepsize, which requires a lower bound on the optimal value, and a normalized gradient method, which requires an estimate of $R_0 = \|x_0 - x^*\|$. Takezawa et al. [18] propose an inexact Polyak stepsize that requires no such quantity, but their rate deteriorates from the optimal convex one to the nonconvex one, with an additional term measuring the quality of the lower bound used (Table 1). Finally, the convex bound of Jerad [13] also contains an additional logarithmic factor in ϵ^{-1} (Table 1). The best known rates are thus attained only by substituting some other problem-dependent quantity for L_0 and L_1 ; attaining them with none at all, simultaneously in the nonconvex, convex and strongly convex settings, is our main contribution.

We summarize the best known complexity bounds in Tables 1 and 2.

Notation

For a finite set $\mathcal{A}$, $|\mathcal{A}|$ denotes its cardinality. For a vector x and matrix B , $\|x\|$ and $\|B\|$ denote the Euclidean norm and its induced matrix norm, respectively. We denote by $\mathbb{S}^{n-1}$ the unit sphere in $\mathbb{R}^n$. In what follows, L -smoothness means that ∇f is Lipschitz continuous with constant L . Recall that a polynomial $P : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is said to be homogeneous of degree p if $P(tx) = t^p P(x)$ for every $x \in \mathbb{R}^n$ and every $t \in \mathbb{R}$. For $x \in \mathbb{R}^n$, let $x^{\otimes p}$ denote the p -fold tensor product of x , i.e., the order- p tensor with entries $(x^{\otimes p})_{i_1, \dots, i_p} := x_{i_1} \cdots x_{i_p}$, where $i_1, \dots, i_p$ range over $\{1, \dots, n\}$. Hence, for any order- p tensor A ,

$$\langle A, x^{\otimes p} \rangle = \sum_{i_1, \dots, i_p=1}^n A_{i_1, \dots, i_p} x_{i_1} \cdots x_{i_p}.$$

Finally, we use the convention that $0^0 = 1$.

Table 1: Iteration complexity under $(1, L_0, L_1)$ -smoothness. Here $\delta_0 = f(x_0) - f_{\text{low}}$, $R_0 = \|x_0 - x^*\|$, where x^* is a minimizer of f , $R \geq R_0$ is such that $\|x - x^*\| \leq R$ for all x with $f(x) \leq f(x_0)$ (see Assumption 4.1), M is an upper bound on the gradient norm, L is the Lipschitz constant of the gradient, $\mu > 0$ denotes the strong convexity constant, $\sigma^2 = f_{\text{low}} - l^* \geq 0$, where l^* is a lower bound on the optimal value f_{low} and $\epsilon > 0$ is the desired accuracy.

Class	Method	Parameters required				Complexity bound
		L_0	L_1	f_{low}	R_0	
Non-convex	[22, Theorem 6]	✓	✓	✗	✗	$\delta_0 (L_0 + L_1 M) \epsilon^{-2}$
	[21, Theorem 3.1]	✓	✓	✗	✗	$\delta_0 (L_0 \epsilon^{-2} + L_1^2 / L_0)$
	[14, Theorem 3.1], [20, Theorem 2.1]	✓	✓	✗	✗	$\delta_0 (L_0 \epsilon^{-2} + L_1 \epsilon^{-1})$
	[12, Theorem 4]	✗	✗	✗	✗	$\delta_0 (L_0 + \delta_0 L_1^2) \epsilon^{-2}$
	[13, Theorem 3.1]	✗	✗	✗	✗	$\delta_0 (L_0 + L_1) \log(\epsilon^{-1}) \epsilon^{-2}$
	Theorem 4.6	✗	✗	✗	✗	$\delta_0 (L_0 \epsilon^{-2} + L_1 \epsilon^{-1})$
Convex	[14, Theorem 2.3]	✓	✓	✗	✗	$L_0 R_0^2 \epsilon^{-1} + L_1 R_0^2 \sqrt{L \epsilon^{-1}}$
	[10, Theorem 3.3]	✓	✓	✗	✗	$L_0 R_0^2 \epsilon^{-1} + (L_1 R_0)^2$
	[10, Theorem 5.2]	✓	✓	✗	✗	$\sqrt{L_0 R_0^2 (1 + L_1 R_0^2 \exp(L_1 R_0))} \epsilon^{-1}$
	[20, Theorem 3.2]	✓	✓	✗	✗	$L_0 R_0^2 \epsilon^{-1} + L_1 R_0 \log(\delta_0 / \epsilon)$
	[20, Theorem 6.2]	✓	✓	✗	✗	$\sqrt{L_0 R_0^2 \epsilon^{-1} + (L_1 R_0)^2}^{2/3} \log(\delta_0 / \epsilon)$
	[10, Theorem 4.1], [20, Theorem 5.1]	✗	✗	✓	✗	$L_0 R_0^2 \epsilon^{-1} + (L_1 R_0)^2$
	[20, Theorem 4.1]	✗	✗	✗	✓	$L_0 R_0^2 \epsilon^{-1} + (L_1 R_0)^2$
	[18, Theorem 5]	✗	✗	✗	✗	$(L_0 R_0^2 + \sigma^2) \epsilon^{-2} + (L L_1^2 R_0^4 + L_1^2 L \sigma^4 / L_0^2) \epsilon^{-1}$
	[13, Theorem 3.2]	✗	✗	✗	✗	$\delta_0 (L_0 + L_1) \log(\epsilon^{-1}) \epsilon^{-1}$
	Theorem 4.9	✗	✗	✗	✗	$L_0 R^2 \epsilon^{-1} + L_1 R \log(\delta_0 / \epsilon)$
Strongly convex	[14, Theorem 2.5]	✓	✓	✗	✗	$L_0 / \mu \log(R_0^2 / \epsilon) + L_0 R_0 \min\{\sqrt{L / \mu}, L R_0\}$
	[10, Theorem D.3]	✓	✓	✗	✗	$L_0 / \mu \log(R_0^2 / \epsilon) + (L_1 R_0)^2$
	Theorem 4.11	✗	✗	✗	✗	$L_0 / \mu \log(\delta_0 / \epsilon) + L_1 (\delta_0 / \mu)^{1/2}$

2 Preliminaries

We make the following standard assumptions on the problem.

Problem Assumption 2.1. The objective function f is continuously differentiable and bounded below. We denote its infimum by f_{low} .

We recall the definition of (α, L_0, L_1) -smoothness. The case $\alpha = 1$ (2) was introduced by Zhang et al. [22] for twice continuously differentiable functions; Zhang et al. [21] observed that the integral form below is meaningful for merely continuously differentiable functions, and Chen et al. [8] extended it to general $\alpha \in [0, 1]$.

Definition 2.1 ([8, 21, 22]). Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be continuously differentiable, $\alpha \in [0, 1]$, and $L_0 > 0$ or $L_1 > 0$. We say that f is (α, L_0, L_1) -smooth if, for all $x, y \in \mathbb{R}^n$,

$$\|\nabla f(x) - \nabla f(y)\| \leq (L_0 + L_1 \int_0^1 \|\nabla f(x + t(y-x))\|^\alpha dt) \|x - y\|. \quad (3)$$

Definition 2.1 yields the following equivalent form as proved by Chen et al. [8, Lemma A.1].

$$\|\nabla f(x) - \nabla f(y)\| \leq (L_0 + L_1 \sup_{u \in [x, y]} \|\nabla f(u)\|^\alpha) \|x - y\|, \quad \forall x, y \in \mathbb{R}^n. \quad (4)$$

We make the following assumption on the objective function f .

Problem Assumption 2.2. We assume that f is (α, L_0, L_1) -smooth in the sense of Definition 2.1, for some $\alpha \in [0, 1]$ and $L_0, L_1 \geq 0$.

When $f \in \mathcal{C}^2$, Problem Assumption 2.2 is equivalent to (2).

Table 2: Iteration complexity under (α, L_0, L_1) -smoothness, where $\alpha \in (0, 1)$. The conventions are those of Table 1.

Class	Method	Parameters required			Complexity bound
		L_0	L_1	α	
Non-convex	[8, Theorem 2]	✓	✓	✓	$\delta_0(L_0(2^{\frac{\alpha^2}{1-\alpha}} + 1)\epsilon^{-2} + L_1 2^{\frac{\alpha^2}{1-\alpha}} 3^\alpha \epsilon^{\alpha-2})$
	[15, Theorem 5.2]	✓	✓	✓	$\delta_0(L_0 + L_1 \ \nabla f(x_0)\ ^\alpha) \epsilon^{-2}$
	[19, Theorem 5.1]	✓	✓	✓	$\delta_0(L_0 \epsilon^{-2} + L_1 \epsilon^{\alpha-2})$
	Theorem 4.6	✗	✗	✗	$\delta_0(L_0 \epsilon^{-2} + L_1 \epsilon^{\alpha-2})$
Convex	[15, Theorem 4.2]	✓	✓	✓	$(L_0 R_0^2 + L_1 R_0^2 \ \nabla f(x_0)\ ^\alpha) \epsilon^{-1}$
	[19, Theorem 8.1]	✓	✓	✓	$L_0 R_0^2 \epsilon^{-1} + L_1^{\frac{2}{2-\alpha}} R_0^2 \epsilon^{-\frac{2(1-\alpha)}{2-\alpha}}$
	Theorem 4.9	✗	✗	✗	$L_0 R^2 \epsilon^{-1} + L_1 R^{2-\alpha} (\epsilon^{\alpha-1} - \delta_0^{\alpha-1}) / (1 - \alpha)$
Strongly convex	[15, Theorem 4.3]	✓	✓	✓	$(L_0 + L_1 \ \nabla f(x_0)\ ^\alpha) / \mu \log(R_0^2 / \epsilon)$
	Theorem 4.11	✗	✗	✗	$L_0 / \mu \log(\delta_0 / \epsilon) + L_1 \mu^{\alpha/2-1} (\delta_0^{\alpha/2} - \epsilon^{\alpha/2}) / \alpha$

2.1 Examples

Several classes of functions satisfy Problem Assumption 2.2. For instance, a univariate polynomial of degree at least three is not globally L -smooth, yet it satisfies the $(1, L_0, L_1)$ -smoothness condition [22]. Functions such as $f(x) = \|x\|^{2n}$, $f(x) = \exp(a^T x)$, and the logistic loss also satisfy the $(1, L_0, L_1)$ -smoothness condition [10]. A class of machine learning problems is also known to satisfy $(1, L_0, L_1)$ -smoothness, including distributionally robust optimization [8], symmetric and asymmetric matrix factorization [4], and the training of neural networks [22]. By contrast, to the best of our knowledge, phase retrieval [8], which satisfies $(\frac{2}{3}, L_0, L_1)$ -smoothness, is the only explicit example previously exhibited for $0 < \alpha < 1$.

In the following, we introduce a generic class of (α, L_0, L_1) -smooth functions, which we call *polynomial inverse problems*, and which includes phase retrieval [8] as a special case.

Theorem 2.2 (Polynomial inverse problems). *Let $\Phi : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be a polynomial map of degree $p \geq 1$, let $y \in \mathbb{R}^m$ be fixed data, and consider $f(x) = \frac{1}{2} \|\Phi(x) - y\|^2$. Write $\Phi = \Phi_p + \Phi_{p-1} + \dots + \Phi_0$, where Φ_k is homogeneous of degree k .*

Define the leading homogeneous objective $f_{2p}(x) = \frac{1}{2} \|\Phi_p(x)\|^2$.

Assume the following noncriticality condition on the unit sphere,

$$\nabla f_{2p}(u) \neq 0, \quad \forall u \in \mathbb{S}^{n-1}.$$

Then there exist constants $L_0, L_1 > 0$ such that

$$\|\nabla^2 f(x)\| \leq L_0 + L_1 \|\nabla f(x)\|^\alpha, \quad \alpha = \frac{2p-2}{2p-1}, \quad \forall x \in \mathbb{R}^n. \quad (5)$$

For $p = 1$, $\alpha = 0$, which is consistent with f being a convex quadratic, hence L -smooth.

Proof. Since Φ has degree p , $f(x) = \frac{1}{2} \|\Phi(x) - y\|^2$ is a polynomial of degree $2p$ with leading homogeneous part $f_{2p}(x) = \frac{1}{2} \|\Phi_p(x)\|^2$. Hence,

$$\nabla f_{2p}(x) = O(\|x\|^{2p-1}) \quad \text{and} \quad \nabla^2 f_{2p}(x) = O(\|x\|^{2p-2}),$$

and therefore,

$$\nabla f(x) = \nabla f_{2p}(x) + O(\|x\|^{2p-2}) \quad \text{and} \quad \nabla^2 f(x) = \nabla^2 f_{2p}(x) + O(\|x\|^{2p-3}),$$

thus there exist $C_0, C_1 > 0$ such that

$$\|\nabla^2 f(x)\| \leq C_0 + C_1 \|x\|^{2p-2}, \quad \forall x \in \mathbb{R}^n. \quad (6)$$

Moreover, since f_{2p} is homogeneous of degree $2p$,

$$\|\nabla f_{2p}(x)\| = \|x\|^{2p-1} \left\| \nabla f_{2p} \left(\frac{x}{\|x\|} \right) \right\|,$$

Since $\mathbb{S}^{n-1}$ is compact and ∇f_{2p} is continuous, the noncriticality assumption ensures that $\mu := \min_{u \in \mathbb{S}^{n-1}} \|\nabla f_{2p}(u)\| > 0$, hence $\|\nabla f_{2p}(x)\| \geq \mu \|x\|^{2p-1}$ for every x .

Because $\nabla f(x) = \nabla f_{2p}(x) + O(\|x\|^{2p-2})$, there exist $C_2, C_3 > 0$ such that

$$\|\nabla f(x)\| \geq \|\nabla f_{2p}(x)\| - C_2 \|x\|^{2p-2} - C_3 \geq \mu \|x\|^{2p-1} - C_2 \|x\|^{2p-2} - C_3 \geq \frac{\mu}{2} \|x\|^{2p-1},$$

where the last inequality holds for all sufficiently large $\|x\|$ outside a compact set $K \subset \mathbb{R}^n$. Substituting this estimate into the Hessian bound (6) gives the desired inequality outside K . Enlarging L_0 , if necessary, extends the inequality to all $x \in \mathbb{R}^n$. $\square$

We conclude this section with a few examples of polynomial inverse problems.

Example 2.3. Let $\Phi(x) = (\langle A_1, x^{\otimes p} \rangle, \dots, \langle A_m, x^{\otimes p} \rangle)$, where $A_1, \dots, A_m$ are symmetric tensors of order p , and consider $f(x) = \frac{1}{2} \sum_{i=1}^m (\langle A_i, x^{\otimes p} \rangle - y_i)^2$. The noncriticality assumption holds whenever

$$\sum_{i=1}^m \langle A_i, u^{\otimes p} \rangle^2 > 0, \quad \forall u \in \mathbb{S}^{n-1}. \quad (7)$$

Then, by [Theorem 2.2](#),

$$\|\nabla^2 f(x)\| \leq L_0 + L_1 \|\nabla f(x)\|^\alpha, \quad \alpha = \frac{2p-2}{2p-1}.$$

The next two examples are the special cases $A_i = a_i a_i^T$ which corresponds to phase retrieval [8], and $A_i = Q_i$ symmetric positive semidefinite matrices, which corresponds to the case $p = 2$.

Example 2.4 (Phase retrieval [8]). This corresponds to the tensor sensing problem with $A_i = a_i^{\otimes 2} = a_i a_i^T$ for some vectors $a_1, \dots, a_m \in \mathbb{R}^n$. The noncriticality assumption holds whenever

$$\text{span}\{a_1, \dots, a_m\} = \mathbb{R}^n,$$

Therefore,

$$\|\nabla^2 f(x)\| \leq L_0 + L_1 \|\nabla f(x)\|^{2/3},$$

Example 2.5. We now consider the case $p = 2$ with $A_i = Q_i$ symmetric positive semidefinite matrices, so that $\Phi(x) = (x^T Q_1 x, \dots, x^T Q_m x)$. The noncriticality assumption holds whenever

$$\bigcap_{i=1}^m \ker(Q_i) = \{0\}.$$

Hence, according to [Theorem 2.2](#),

$$\|\nabla^2 f(x)\| \leq L_0 + L_1 \|\nabla f(x)\|^{2/3}.$$

Proof. Here $\langle Q_i, u^{\otimes 2} \rangle = u^T Q_i u$, so (7) reads $\sum_{i=1}^m (u^T Q_i u)^2 > 0$. If this failed for some $u \in \mathbb{S}^{n-1}$, then $u^T Q_i u = 0$ for every i , and $Q_i \succeq 0$ would give $Q_i u = 0$ for every i , that is, $u \in \bigcap_{i=1}^m \ker(Q_i) = \{0\}$, contradicting $\|u\| = 1$. $\square$

2.2 Properties of (α, L_0, L_1) -smooth functions

Before presenting the trust-region algorithm, we need the following characterization of (α, L_0, L_1) -smoothness that will be useful in the analysis of the algorithm. The proof is provided in [Appendix A](#).

Theorem 2.6. *Let [Problem Assumptions 2.1](#) and [2.2](#) hold. Then, for all $x, s \in \mathbb{R}^n$, such that $\nabla f(x) \neq 0$, we have*

$$\|\nabla f(x)\| \leq \sqrt{8L_0(f(x) - f_{\text{low}})} + (2^{\alpha+3}L_1(f(x) - f_{\text{low}}))^{\frac{1}{2-\alpha}}, \quad (8)$$

$$\|\nabla f(x+s) - \nabla f(x)\| \leq (L_0 + L_1\|\nabla f(x)\|^\alpha)g(\|s\|, \|\nabla f(x)\|)\|s\|, \quad (9)$$

and

$$|f(x+s) - f(x) - \nabla f(x)^T s| \leq \frac{1}{2}(L_0 + L_1\|\nabla f(x)\|^\alpha)g(\|s\|, \|\nabla f(x)\|)\|s\|^2, \quad (10)$$

where $g : [0, \infty) \times (0, \infty) \rightarrow (0, \infty)$ is defined by

$$g(a, b) := \begin{cases} \left(1 + \frac{(1-\alpha)L_1a^\alpha}{(L_0 + L_1b^\alpha)^{1-\alpha}}\right)^{\frac{1}{1-\alpha}}, & \text{if } \alpha \in [0, 1) \text{ and } L_0 + L_1b^\alpha > 0, \\ \exp(L_1a), & \text{if } \alpha = 1. \end{cases} \quad (11)$$

[See proof on page 18.](#)

When $\alpha = 1$, the function g only depends on a and not on b , but we keep the notation $g(a, b)$ for consistency with the case $\alpha \in [0, 1)$.

The bounds in [Theorem 2.6](#) are tighter than those of Chen et al. [\[8\]](#) under the (α, L_0, L_1) -smoothness condition for $\alpha < 1$. Moreover, as $\alpha \rightarrow 1^-$, the function g converges to $\exp(L_1a)$, whereas the bound of Chen et al. [\[8\]](#) diverges to $+\infty$.

In our analysis, we will need the following properties. The proof of these properties is straightforward and is omitted for brevity.

Proposition 2.7. *For $\alpha \in [0, 1]$, let g be defined as in [Theorem 2.6](#). Then,*

1. g is continuous on $[0, \infty) \times (0, \infty)$ and $\mathcal{C}^1((0, \infty)^2)$.
2. For all fixed $b > 0$, the function $a \mapsto g(a, b)$ is nondecreasing on $[0, \infty)$.
3. For all fixed $a \geq 0$, the function $b \mapsto g(a, b)$ is nonincreasing on $(0, \infty)$.
4. For all $\alpha > 0$ and $b > 0$, $\lim_{a \rightarrow 0^+} g(a, b) = 1$.
5. For all $a \geq 0$ and $b > 0$, $g(a, b) \geq 1$.

In the next section, we present the trust-region algorithm that we analyze in this paper, and we state the main assumptions on the model Hessian.

3 Trust-region algorithm

3.1 Algorithm

At each iteration k , we compute a step s_k from the current iterate x_k as an (inexact) solution of the subproblem

$$\min_{s \in \mathbb{R}^n} m_k(s) \quad \text{s.t. } \|s\| \leq \Delta_k, \quad \text{where } m_k(s) := f(x_k) + \nabla f(x_k)^T s + \frac{1}{2} s^T B_k s, \quad (12)$$

where $B_k = B_k^T$ and $\Delta_k > 0$ is the trust-region radius.

Only an approximate solution of [\(12\)](#) is required; the step s_k is required to achieve at least a fixed fraction of the Cauchy decrease. The decrease required of s_k is stated as

$$m_k(0) - m_k(s_k) \geq \kappa_{\text{mdc}} \|\nabla f(x_k)\| \min \left\{ \frac{\|\nabla f(x_k)\|}{\|B_k\|}, \Delta_k \right\}, \quad (13)$$

where $0 < \kappa_{\text{mdc}} \leq \frac{1}{2}$, with the convention that $\frac{\|\nabla f(x_k)\|}{\|B_k\|} = +\infty$ if $B_k = 0$. In particular, if s_k solves exactly the subproblem (12) then it satisfies (13).

Once a trial step s_k has been determined, the decrease in f at $x_k + s_k$ is compared to the decrease predicted by the model. If the actual and predicted decreases are in sufficient agreement, $x_k + s_k$ becomes the new iterate, and Δ_k is possibly increased. If the model predicts the actual decrease poorly, the trial point is rejected and Δ_k is reduced. Algorithm 1 states the entire procedure.

Algorithm 1 Classical trust-region algorithm Cartis et al. [7, Algorithm 2.3.1].

1: Choose $x_0 \in \mathbb{R}^n$, $\Delta_0 > 0$, $0 < \eta_1 \leq \eta_2 < 1$, and $0 < \gamma_1 \leq \gamma_2 < 1 \leq \gamma_3$.

2: **for** $k = 0, 1, \dots$ **do**

3: Choose $B_k = B_k^T \in \mathbb{R}^{n \times n}$.

4: Compute s_k satisfying (13).

5: Set

$$\rho_k := \frac{f(x_k) - f(x_k + s_k)}{m_k(0) - m_k(s_k)}.$$

6: Set $x_{k+1} = x_k + s_k$ if $\rho_k \geq \eta_1$, and $x_{k+1} = x_k$ otherwise.

7: Choose

$$\Delta_{k+1} \in \begin{cases} [\Delta_k, \gamma_3 \Delta_k], & \rho_k \geq \eta_2, \\ [\gamma_2 \Delta_k, \Delta_k], & \eta_1 \leq \rho_k < \eta_2, \\ [\gamma_1 \Delta_k, \gamma_2 \Delta_k], & \rho_k < \eta_1. \end{cases}$$

8: **end for**

In Algorithm 1, no upper bound is imposed on the trust-region radius Δ_k , which may therefore become arbitrarily large.

We will repeatedly use the notation

$$\begin{aligned} \mathcal{S} &:= \{i \in \mathbb{N} \mid \rho_i \geq \eta_1\} && \text{(all successful iterations)} \\ \mathcal{S}_k &:= \{i \in \mathcal{S} \mid i \leq k\} && \text{(successful iterations until iteration } k) \\ \mathcal{U}_k &:= \{i \in \mathbb{N} \mid i \notin \mathcal{S}, i \leq k\} && \text{(unsuccessful iterations until iteration } k). \end{aligned}$$

We recall the following standard relation between the numbers of successful and unsuccessful iterations [7, Lemma 2.3.1].

Lemma 3.1. For all $k \in \mathbb{N}^*$, $|\mathcal{U}_{k-1}| \leq |\log_{\gamma_2}(\gamma_3)| |\mathcal{S}_{k-1}| + \log_{\gamma_2} \left(\frac{\Delta_k}{\Delta_0} \right)$.

Lemma 3.1 implies that the number of unsuccessful iterations is at most a constant times the number of successful iterations plus a logarithmic term that can be controlled using a lower bound on Δ_k . Unless otherwise stated, we will not explicitly keep track of the number of unsuccessful iterations in our analysis and will only focus on the number of successful iterations and the lower bound on Δ_k .

3.2 Model Hessian bounds

We make the following assumption.

Model Assumption 3.1. There exist $b_0, b_1 \geq 0$ such that for all $k \in \mathbb{N}$,

$$\|B_k\| \leq b_0 + b_1 \|\nabla f(x_k)\|^\alpha. \quad (14)$$

Model Assumption 3.1 allows for three standard choices: $B_k = 0$ and $B_k = \lambda I$ satisfy it with $b_1 = 0$, the former yielding normalized-gradient-type trust-region steps and the latter clipped-gradient steps [22], while $B_k = \nabla^2 f(x_k)$ satisfies it with $(b_0, b_1) = (L_0, L_1)$ whenever f is twice continuously differentiable, by (2).

Moreover, Model Assumption 3.1 is compatible with limited-memory quasi-Newton approximations as shown in the following proposition.

Proposition 3.2 (Limited-memory quasi-Newton updates). *Let $m \in \mathbb{N}$ be the memory size. At iteration k , let B_k be obtained from an initial matrix $B_{k,0}$ and a set of pairs $\{(s_i, y_i)\}_{i \in \mathcal{I}_k}$ stored at successful iterations, where $\mathcal{I}_k = \{i_0, \dots, i_{q_k-1}\} \subseteq \mathcal{S}$, $q_k := |\mathcal{I}_k| \leq m$, $s_i \neq 0$, and $y_i = \nabla f(x_i + s_i) - \nabla f(x_i)$. Assume that $B_{k,0}$ satisfies [Model Assumption 3.1](#). Then, for any of the following limited-memory quasi-Newton updates, B_k satisfies [Model Assumption 3.1](#).*

- (i) **Limited-memory BFGS** [5]. *Assume that $B_{k,0} \succ 0$ and every stored pair satisfies $y_{i_j} \neq 0$ and $s_{i_j}^T y_{i_j} \geq \omega_{\text{BFGS}} \|s_{i_j}\| \|y_{i_j}\|$, where $\omega_{\text{BFGS}} \in (0, 1)$. Set $B_k = B_{k,q_k}$, where, for $j = 0, \dots, q_k - 1$,*

$$B_{k,j+1} = B_{k,j} - \frac{B_{k,j} s_{i_j} s_{i_j}^T B_{k,j}}{s_{i_j}^T B_{k,j} s_{i_j}} + \frac{y_{i_j} y_{i_j}^T}{s_{i_j}^T y_{i_j}}. \quad (15)$$

- (ii) **Modified limited-memory BFGS**. *Assume that $B_{k,0} \succ 0$ and every stored pair satisfies $y_{i_j} \neq 0$ and $s_{i_j}^T y_{i_j} \geq \bar{\omega}_{\text{BFGS}} \|y_{i_j}\|^2$, where $\bar{\omega}_{\text{BFGS}} \in (0, 1)$. Set $B_k = B_{k,q_k}$ as in (15).*
- (iii) **Limited-memory SR1** [5]. *Set $z_{i_j} = y_{i_j} - B_{k,j} s_{i_j}$ and assume that every stored pair satisfies $z_{i_j} \neq 0$ and $|s_{i_j}^T z_{i_j}| \geq \omega_{\text{SR1}} \|s_{i_j}\| \|z_{i_j}\|$, where $\omega_{\text{SR1}} \in (0, 1)$. Set $B_k = B_{k,q_k}$, where, for $j = 0, \dots, q_k - 1$,*

$$B_{k,j+1} = B_{k,j} + \frac{(y_{i_j} - B_{k,j} s_{i_j})(y_{i_j} - B_{k,j} s_{i_j})^T}{(y_{i_j} - B_{k,j} s_{i_j})^T s_{i_j}}. \quad (16)$$

- (iv) **Modified limited-memory SR1** [1, 2]. *Set $z_{i_j} = y_{i_j} - B_{k,j} s_{i_j}$ and assume that every stored pair satisfies $z_{i_j} \neq 0$ and $|s_{i_j}^T z_{i_j}| \geq \bar{\omega}_{\text{SR1}} \|z_{i_j}\|^2$, where $\bar{\omega}_{\text{SR1}} \in (0, 1)$. Set $B_k = B_{k,q_k}$ as in (16).*
- (v) **Limited-memory PSB** [16]. *Set $z_{i_j} = y_{i_j} - B_{k,j} s_{i_j}$ and $B_k = B_{k,q_k}$, where, for $j = 0, \dots, q_k - 1$,*

$$B_{k,j+1} = B_{k,j} + \frac{z_{i_j} s_{i_j}^T + s_{i_j} z_{i_j}^T}{\|s_{i_j}\|^2} - \frac{z_{i_j}^T s_{i_j}}{\|s_{i_j}\|^4} s_{i_j} s_{i_j}^T. \quad (17)$$

- (vi) **Modified limited-memory PSB**. *Assume that every stored pair satisfies $\|s_{i_j}\| \geq \bar{\omega}_{\text{PSB}} \|y_{i_j} - B_{k,j} s_{i_j}\|$, where $\bar{\omega}_{\text{PSB}} > 0$. Set $B_k = B_{k,q_k}$ as in (17).*

See proof on page 19.

[Proposition 3.2](#) shows that the considered limited-memory quasi-Newton updates preserve [Model Assumption 3.1](#) whenever the initial matrix $B_{k,0}$ satisfies this assumption. Possible choices include $B_{k,0} = \lambda I$, or, when α is known, $B_{k,0} = \|\nabla f(x_k)\|^\alpha I$.

3.3 Preliminary results

Without loss of generality, we assume that $L_0 \geq b_0$ and $L_1 \geq b_1$, thereby avoiding the need to repeatedly take the maximum of the corresponding constants, where L_0, L_1, b_0 , and b_1 are defined in [Problem Assumption 2.2](#) and [Model Assumption 3.1](#).

We will repeatedly use

$$\psi(t) := \frac{t}{L_0 + L_1 t^\alpha}. \quad (18)$$

Note that ψ is nondecreasing on $(0, \infty)$. The following lemma states a consequence of the Cauchy decrease (13) in terms of ψ .

Lemma 3.3. *Let [Problem Assumption 2.1](#) and [Model Assumption 3.1](#) hold. For all $k \in \mathbb{N}$,*

$$m_k(0) - m_k(s_k) \geq \kappa_{\text{mdc}} \|\nabla f(x_k)\| \min \{\psi(\|\nabla f(x_k)\|), \Delta_k\}, \quad (19)$$

where ψ is defined in (18).

The following model-error estimate is a consequence of [Theorem 2.6](#) (10) and [Proposition 2.7](#) (5).

Lemma 3.4. Let [Problem Assumptions 2.1](#) and [2.2](#) and [Model Assumption 3.1](#) hold. For all $k \in \mathbb{N}$,

$$|f(x_k + s_k) - m_k(s_k)| \leq (L_0 + L_1 \|\nabla f(x_k)\|^\alpha) g(\|s_k\|, \|\nabla f(x_k)\|) \|s_k\|^2, \quad (20)$$

where g is defined in [Theorem 2.6](#).

4 Complexity analysis

The complexity analysis of [Algorithm 1](#) relies on the following lemma, which establishes the well-definedness and monotonicity of the function h that will be used to define a lower bound on Δ_k .

Lemma 4.1 (Definition and monotonicity of h). Let $\alpha \in [0, 1]$, and consider the function g defined in [Theorem 2.6](#). Define

$$\Psi(a, b) := ag(a, b), \quad a, b > 0. \quad (21)$$

Then, for every $b > 0$, the mapping $\Psi(\cdot, b)$ is a bijection from $(0, \infty)$ onto $(0, \infty)$. Consequently, $h : (0, \infty)^2 \rightarrow (0, \infty)$

$$h(b, t) := (\Psi(\cdot, b))^{-1}(t), \quad (22)$$

is well-defined. Equivalently, $h(b, t)$ is the unique positive number satisfying

$$\Psi(h(b, t), b) = t. \quad (23)$$

Moreover, $h \in C^1((0, \infty)^2)$, is strictly increasing with respect to t , and is nondecreasing with respect to b . More precisely,

$$\partial_t h(b, t) = \frac{1}{\partial_a \Psi(h(b, t), b)} > 0 \quad \text{and} \quad \partial_b h(b, t) = -\frac{\partial_b \Psi(h(b, t), b)}{\partial_a \Psi(h(b, t), b)} \geq 0. \quad (24)$$

[See proof on page 20.](#)

Now we derive a lower bound on $h(b, t)$.

Lemma 4.2. For fixed $b > 0$, let $h(b, \cdot)$ be defined as in [\(22\)](#). Then, for all $t > 0$,

$$\frac{t}{1 + \frac{L_1 t^\alpha}{(L_0 + L_1 b^\alpha)^{1-\alpha}}} \leq h(b, t) \leq t.$$

[See proof on page 21.](#)

For every k such that $\nabla f(x_k) \neq 0$, define

$$\theta_k := \frac{\Delta_k}{h(\min_{j \leq k} \|\nabla f(x_j)\|, \kappa_{\text{mdc}}(1 - \eta_2) \psi(\min_{j \leq k} \|\nabla f(x_j)\|))}.$$

If $\nabla f(x_k) = 0$ for some k , then x_k is already a first-order stationary point. Otherwise, both arguments of h are positive, so θ_k is well defined by [Lemma 4.1](#). Next, we establish that $\{\theta_k\}_{k \in \mathbb{N}}$ is uniformly bounded below.

Lemma 4.3. Let [Problem Assumptions 2.1](#) and [2.2](#) and [Model Assumption 3.1](#) hold. For all $k \in \mathbb{N}$,

$$\theta_k \geq \theta_{\min} := \gamma_1 \min \left\{ 1, \frac{\Delta_0}{h(\|\nabla f(x_0)\|, \kappa_{\text{mdc}}(1 - \eta_2) \psi(\|\nabla f(x_0)\|))} \right\} > 0.$$

Moreover, if $L_0 \geq 1$ or $L_1 \geq 1$, then by choosing $\Delta_0 \geq \max\{1, \|\nabla f(x_0)\|\}$, we have $\theta_{\min} = \gamma_1$.

Proof. First, observe that for $k = 0$,

$$\theta_0 = \frac{\Delta_0}{h(\|\nabla f(x_0)\|, \kappa_{\text{mdc}}(1 - \eta_2)\psi(\|\nabla f(x_0)\|))} \geq \theta_{\min},$$

because $\gamma_1 \leq 1$.

Suppose, for contradiction, that $\theta_{k+1} < \theta_{\min}$ for some index, and let k be the smallest such index; by the case $k = 0$ above and minimality, $\theta_k \geq \theta_{\min}$. By Line 7, $\gamma_1 \Delta_k \leq \Delta_{k+1}$, which implies that,

$$\frac{\Delta_k}{h(\min_{j \leq k+1} \|\nabla f(x_j)\|, \kappa_{\text{mdc}}(1 - \eta_2)\psi(\min_{j \leq k+1} \|\nabla f(x_j)\|))} \leq \frac{\theta_{k+1}}{\gamma_1} < \frac{\theta_{\min}}{\gamma_1} \leq 1.$$

Applying the increasing mapping $\Psi(\cdot, \min_{j \leq k+1} \|\nabla f(x_j)\|)$ to both sides and using the definition of h , we obtain

$$\begin{aligned} \Psi(\Delta_k, \min_{j \leq k+1} \|\nabla f(x_j)\|) &< \kappa_{\text{mdc}}(1 - \eta_2)\psi(\min_{j \leq k+1} \|\nabla f(x_j)\|) \\ &\stackrel{(a)}{\leq} \kappa_{\text{mdc}}(1 - \eta_2)\psi(\|\nabla f(x_k)\|) \\ &\stackrel{(b)}{\leq} \psi(\|\nabla f(x_k)\|), \end{aligned} \tag{25}$$

where (a) follows from the fact that ψ is increasing and (b) follows from the fact that $\kappa_{\text{mdc}}(1 - \eta_2) \leq 1$.

On the other hand, it follows from the definition of ρ_k ,

$$\begin{aligned} |\rho_k - 1| &= \frac{|f(x_k + s_k) - m_k(s_k)|}{m_k(0) - m_k(s_k)} \\ &\stackrel{(a)}{\leq} \frac{(L_0 + L_1 \|\nabla f(x_k)\|^\alpha)g(\|s_k\|, \|\nabla f(x_k)\|)\|s_k\|^2}{\kappa_{\text{mdc}}\|\nabla f(x_k)\| \min\{\psi(\|\nabla f(x_k)\|), \Delta_k\}} \\ &\stackrel{(b)}{=} \frac{(L_0 + L_1 \|\nabla f(x_k)\|^\alpha)\Psi(\|s_k\|, \|\nabla f(x_k)\|)\|s_k\|}{\kappa_{\text{mdc}}\|\nabla f(x_k)\|\Delta_k} \\ &\stackrel{(c)}{\leq} \frac{\Psi(\Delta_k, \min_{j \leq k+1} \|\nabla f(x_j)\|)}{\kappa_{\text{mdc}}\psi(\|\nabla f(x_k)\|)} \stackrel{(d)}{\leq} 1 - \eta_2, \end{aligned}$$

Here, (a) follows from Lemma 3.4 together with the sharp form (19) of the Cauchy decrease. Step (b) uses the definition $\Psi(a, b) = a g(a, b)$ in the numerator, and the identity $\min\{\psi(\|\nabla f(x_k)\|), \Delta_k\} = \Delta_k$ in the denominator. The latter holds because $g \geq 1$ by Proposition 2.7 (5), so that, writing $b_{k+1} := \min_{j \leq k+1} \|\nabla f(x_j)\|$,

$$\Delta_k \leq \Delta_k g(\Delta_k, b_{k+1}) = \Psi(\Delta_k, b_{k+1}) \stackrel{(25)}{\leq} \psi(\|\nabla f(x_k)\|).$$

Step (c) follows from $\|s_k\| \leq \Delta_k$ together with the fact that Ψ is increasing with respect to its first argument and nonincreasing with respect to its second argument, and from the identity $(L_0 + L_1 \|\nabla f(x_k)\|^\alpha)/\|\nabla f(x_k)\| = 1/\psi(\|\nabla f(x_k)\|)$. Finally, (d) follows from inequality (25). Therefore, $\rho_k \geq \eta_2$, which implies that iteration k is very successful, and $\Delta_{k+1} \geq \Delta_k$. Thus, by the monotonicity of h established in Lemma 4.1 and of ψ , we have

$$\begin{aligned} &h(\min_{j \leq k+1} \|\nabla f(x_j)\|, \kappa_{\text{mdc}}(1 - \eta_2)\psi(\min_{j \leq k+1} \|\nabla f(x_j)\|)) \\ &\leq h(\min_{j \leq k} \|\nabla f(x_j)\|, \kappa_{\text{mdc}}(1 - \eta_2)\psi(\min_{j \leq k} \|\nabla f(x_j)\|)). \end{aligned}$$

Hence, $\theta_k \leq \theta_{k+1} < \theta_{\min}$, which contradicts the assumption that $\theta_k \geq \theta_{\min}$.

For the last statement, by [Lemma 4.2](#),

$$\begin{aligned} h(\|\nabla f(x_0)\|, \kappa_{\text{mdc}}(1 - \eta_2)\psi(\|\nabla f(x_0)\|)) &\leq \kappa_{\text{mdc}}(1 - \eta_2)\psi(\|\nabla f(x_0)\|) \\ &\leq \psi(\|\nabla f(x_0)\|) \\ &\leq \max\{1, \|\nabla f(x_0)\|\}, \end{aligned}$$

Then, if $L_0 \geq 1$, then $\psi(\|\nabla f(x_0)\|) \leq \|\nabla f(x_0)\|/L_0 \leq \|\nabla f(x_0)\|$. If instead $L_1 \geq 1$, then $\psi(\|\nabla f(x_0)\|) \leq \|\nabla f(x_0)\|^{1-\alpha}/L_1 \leq \|\nabla f(x_0)\|^{1-\alpha} \leq \max\{1, \|\nabla f(x_0)\|\}$, since $1 - \alpha \in [0, 1]$.

Hence, if $\Delta_0 \geq \max\{1, \|\nabla f(x_0)\|\}$, then $\theta_{\min} = \gamma_1$. $\square$

In general, $\theta_{\min}$ may depend on the initial gradient norm $\|\nabla f(x_0)\|$. This dependence can be avoided by choosing Δ_0 sufficiently large, as indicated in [Lemma 4.3](#).

The following lemma provides a model-decrease estimate.

Lemma 4.4. *Let [Problem Assumptions 2.1](#) and [2.2](#) and [Model Assumption 3.1](#) hold. For all $k \in \mathbb{N}$,*

$$\begin{aligned} m_k(0) - m_k(s_k) &\geq \\ &\kappa_{\text{mdc}}\|\nabla f(x_k)\|h(\min_{j \leq k} \|\nabla f(x_j)\|, \kappa_{\text{mdc}}(1 - \eta_2)\psi(\min_{j \leq k} \|\nabla f(x_j)\|))\theta_{\min}. \end{aligned}$$

Proof. [Lemma 3.3](#) gives

$$\begin{aligned} m_k(0) - m_k(s_k) &\geq \kappa_{\text{mdc}}\|\nabla f(x_k)\| \min\{\psi(\|\nabla f(x_k)\|), \Delta_k\} \\ &\stackrel{(a)}{\geq} \kappa_{\text{mdc}}\|\nabla f(x_k)\| \min\left\{\kappa_{\text{mdc}}(1 - \eta_2)\psi(\min_{j \leq k} \|\nabla f(x_j)\|), \Delta_k\right\} \\ &\stackrel{(b)}{\geq} \kappa_{\text{mdc}}\|\nabla f(x_k)\| \min\left\{h(\min_{j \leq k} \|\nabla f(x_j)\|, \kappa_{\text{mdc}}(1 - \eta_2)\psi(\min_{j \leq k} \|\nabla f(x_j)\|))\theta_{\min}, \Delta_k\right\} \\ &\stackrel{(c)}{=} \kappa_{\text{mdc}}\|\nabla f(x_k)\|h(\min_{j \leq k} \|\nabla f(x_j)\|, \kappa_{\text{mdc}}(1 - \eta_2)\psi(\min_{j \leq k} \|\nabla f(x_j)\|))\theta_{\min}, \end{aligned}$$

where (a) follows from the fact that $\kappa_{\text{mdc}}(1 - \eta_2) \leq 1$ and that ψ is nondecreasing, (b) follows from $h(b, t) \leq t$ and $\theta_{\min} \leq 1$, and (c) follows from [Lemma 4.3](#). $\square$

The lower bound in [Lemma 4.4](#) is key to the complexity analysis in [Section 4.1](#).

If [Algorithm 1](#) has finitely many successful iterations, it identifies a first-order critical point in finite time.

Theorem 4.5. *Let [Problem Assumptions 2.1](#) and [2.2](#) and [Model Assumption 3.1](#) hold. If [Algorithm 1](#) generates finitely many successful iterations, then $x_k = x^*$ for all sufficiently large k , and $\nabla f(x^*) = 0$.*

Proof. Let k_f be the last successful iteration. Then $x_k = x_{k_f+1}$ for all $k \geq k_f + 1$. Set $x^* := x_{k_f+1}$. Suppose, by contradiction, that $\|\nabla f(x_k)\| \geq \nu > 0$ for all $k \in \mathbb{N}$. By [Lemma 4.3](#) and the monotonicity of h and ψ ,

$$\Delta_k \geq \theta_{\min}h(\nu, \kappa_{\text{mdc}}(1 - \eta_2)\psi(\nu)).$$

Thus, Δ_k is bounded away from zero for all $k \geq k_f + 1$. On the other hand, all iterations after $k_f + 1$ are unsuccessful, so the update mechanism of [Algorithm 1](#) yields $\Delta_k \rightarrow 0$, a contradiction. Hence, $\liminf_{k \rightarrow \infty} \|\nabla f(x_k)\| = \|\nabla f(x^*)\| = 0$. $\square$

4.1 Complexity on general objectives

Let $0 < \epsilon < 1$ and k_ϵ be the first iteration of [Algorithm 1](#) with $\|\nabla f(x_{k_\epsilon})\| > \epsilon$, and $\|\nabla f(x_{k_\epsilon+1})\| \leq \epsilon$. Define

$$\mathcal{S}(\epsilon) := \mathcal{S}_{k_\epsilon-1} = \{k \in \mathcal{S} \mid k < k_\epsilon\}.$$

Theorem 4.6. *Let [Problem Assumptions 2.1](#) and [2.2](#) and [Model Assumption 3.1](#) hold. Assume that [Algorithm 1](#) generates infinitely many successful iterations. Let $\theta_{\min}$ be as in [Lemma 4.3](#). Define*

$$\kappa_0 := \frac{f(x_0) - f_{\text{low}}}{\eta_1 \kappa_{\text{mdc}}^2 (1 - \eta_2) \theta_{\min}}. \quad (26)$$

Then, for all $\alpha \in [0, 1]$,

$$|\mathcal{S}(\epsilon)| \leq L_1 (1 + (\kappa_{\text{mdc}}(1 - \eta_2))^\alpha) \kappa_0 \epsilon^{\alpha-2} + L_0 \kappa_0 \epsilon^{-2}. \quad (27)$$

Proof. If $\mathcal{S}(\epsilon) = \emptyset$, then the theorem holds trivially as $|\mathcal{S}(\epsilon)| = 0$.

Otherwise, let $\ell \in \mathcal{S}(\epsilon)$. By [Lemma 4.4](#), we have

$$\begin{aligned} f(x_\ell) - f(x_\ell + s_\ell) &\geq \eta_1 (m_\ell(0) - m_\ell(s_\ell)) \\ &\geq \eta_1 \kappa_{\text{mdc}} \|\nabla f(x_\ell)\| h(\min_{j \leq \ell} \|\nabla f(x_j)\|, \kappa_{\text{mdc}}(1 - \eta_2) \psi(\min_{j \leq \ell} \|\nabla f(x_j)\|)) \theta_{\min} \\ &\stackrel{(a)}{\geq} \eta_1 \kappa_{\text{mdc}} \epsilon h(\epsilon, \kappa_{\text{mdc}}(1 - \eta_2) \psi(\epsilon)) \theta_{\min}, \end{aligned}$$

where (a) follows from the fact that ψ is increasing and so is h with respect to both of its arguments, and that $\|\nabla f(x_j)\| \geq \epsilon$ for all $j = 0, \dots, \ell$. Using [Lemma 4.2](#),

$$\begin{aligned} f(x_\ell) - f(x_\ell + s_\ell) &\geq \eta_1 \kappa_{\text{mdc}} \epsilon \frac{\frac{\kappa_{\text{mdc}}(1 - \eta_2) \epsilon}{L_0 + L_1 \epsilon^\alpha}}{1 + \frac{L_1 (\kappa_{\text{mdc}}(1 - \eta_2) \epsilon / (L_0 + L_1 \epsilon^\alpha))^\alpha}{(L_0 + L_1 \epsilon^\alpha)^{1-\alpha}}} \theta_{\min} \\ &\geq \eta_1 \kappa_{\text{mdc}}^2 \epsilon^2 \theta_{\min} (1 - \eta_2) \frac{1}{L_0 + L_1 \epsilon^\alpha + L_1 (\kappa_{\text{mdc}}(1 - \eta_2) \epsilon)^\alpha}, \end{aligned}$$

We now sum the above inequality over $\ell \in \mathcal{S}(\epsilon)$. The sequence $\{f(x_k)\}$ is nonincreasing over all iterations, since $x_{k+1} = x_k$ at an unsuccessful iteration and $f(x_{k+1}) < f(x_k)$ at a successful one. Hence the decreases $f(x_\ell) - f(x_{\ell+1})$, $\ell \in \mathcal{S}(\epsilon)$, telescope into at most $f(x_0) - f_{\text{low}}$, and we obtain

$$f(x_0) - f_{\text{low}} \geq \frac{\eta_1 \kappa_{\text{mdc}}^2 \epsilon^2 \theta_{\min} (1 - \eta_2) |\mathcal{S}(\epsilon)|}{L_0 + L_1 (1 + (\kappa_{\text{mdc}}(1 - \eta_2))^\alpha) \epsilon^\alpha},$$

which gives the desired result. $\square$

Several remarks on [Theorem 4.6](#) are in order. Firstly, the constants and the bound are well-defined for every $\alpha \in [0, 1]$, where α is the parameter in [Problem Assumption 2.2](#). Secondly, the bound in [Theorem 4.6](#) recovers the best known gradient-descent complexity bound without requiring any estimate of the problem parameters; see [Table 1](#). Thirdly, the bound in [Theorem 4.6](#) is of order $O(\epsilon^{-2})$ when $\alpha = 0$ and of order $O(L_0 \epsilon^{-2} + L_1 \epsilon^{-1})$ when $\alpha = 1$, which matches the best known dependence under $(1, L_0, L_1)$ -smoothness.

[Theorem 4.6](#) bounds the number of *successful* iterations. The total number of iterations, including unsuccessful ones, is bounded in the following theorem.

Theorem 4.7. *Let the assumptions of [Theorem 4.6](#) hold, assume $k_\epsilon \geq 1$, and let*

$$c_\alpha := \frac{a}{1 + \frac{L_1 a^\alpha}{(L_0 + L_1)^{1-\alpha}}}, \quad a := \frac{\kappa_{\text{mdc}}(1 - \eta_2)}{L_0 + L_1}, \quad \kappa_0 := \frac{f(x_0) - f_{\text{low}}}{\eta_1 \kappa_{\text{mdc}}^2 (1 - \eta_2) \theta_{\min}}.$$

Then,

$$k_\epsilon \leq (1 + |\log_{\gamma_2}(\gamma_3)|) \kappa_0 [L_1 (1 + (\kappa_{\text{mdc}}(1 - \eta_2))^\alpha) \epsilon^{\alpha-2} + L_0 \epsilon^{-2}] \\ + \log_{\gamma_2} \left(\frac{\theta_{\min} c_\alpha \epsilon}{\Delta_0} \right).$$

Proof. Since the iterations $0, \dots, k_\epsilon - 1$ are either successful or unsuccessful, $k_\epsilon = |\mathcal{S}(\epsilon)| + |\mathcal{U}_{k_\epsilon-1}|$. By [Lemma 3.1](#),

$$k_\epsilon \leq (1 + |\log_{\gamma_2}(\gamma_3)|) |\mathcal{S}(\epsilon)| + \log_{\gamma_2} \left(\frac{\Delta_{k_\epsilon}}{\Delta_0} \right).$$

Hence, by [Lemma 4.3](#), the fact that $\min_{j \leq k_\epsilon} \|\nabla f(x_j)\| \geq \epsilon$ and the monotonicity of h and ψ ,

$$\Delta_{k_\epsilon} \geq \theta_{\min} h(\epsilon, \kappa_{\text{mdc}}(1 - \eta_2) \psi(\epsilon)).$$

Moreover, because $\epsilon < 1$, we have

$$\psi(\epsilon) \geq \frac{\epsilon}{L_0 + L_1},$$

and therefore, by [Lemma 4.2](#),

$$h(\epsilon, \kappa_{\text{mdc}}(1 - \eta_2) \psi(\epsilon)) \geq c_\alpha \epsilon.$$

The result now follows by the decreasing property of $\log_{\gamma_2}$, since $\gamma_2 \in (0, 1)$, and [Theorem 4.6](#). $\square$

4.2 Complexity for convex and strongly convex objectives

In this section, we derive complexity results of [Algorithm 1](#) for both convex and strongly convex functions under the (α, L_0, L_1) -smoothness assumption.

Convexity leads to improved complexity bounds. We focus on successful iterations, since [Lemma 3.1](#) then provides a bound on the unsuccessful iterations.

For the convex case, we consider the optimality measure

$$\delta_k := f(x_k) - f_{\text{low}}, \tag{28}$$

where f_{low} is the infimum of [Problem Assumption 2.1](#).

Let $0 < \epsilon < 1$ and, since $\{\delta_k\}_{k \in \mathbb{N}}$ is nonincreasing, let $\hat{k}_\epsilon$ be the first iteration of [Algorithm 1](#) with $\delta_{\hat{k}_\epsilon} > \epsilon$, and $\delta_{\hat{k}_\epsilon+1} \leq \epsilon$. Define

$$\hat{\mathcal{S}}(\epsilon) := \mathcal{S}_{\hat{k}_\epsilon-1} = \{k \in \mathcal{S} \mid k < \hat{k}_\epsilon\}.$$

We impose the following condition, which is standard in the analysis of convex optimization algorithms, and is also used in [\[6\]](#).

Problem Assumption 4.1. The function f is convex, and there exists $R > 0$ such that $\|x - x^*\| \leq R$, for every x satisfying $f(x) \leq f(x_0)$, where x^* is a minimizer of f .

The following property from [\[6\]](#) relates the objective gap to the gradient norm.

Lemma 4.8 ([\[6\]](#), Lemma 2.4). *Assume that [Problem Assumption 4.1](#) holds. Then, for all $k \in \mathbb{N}$, $\delta_k \leq R \|\nabla f(x_k)\|$.*

We now state the main complexity result for convex objectives satisfying (α, L_0, L_1) -smoothness.

Theorem 4.9. *Let [Problem Assumptions 2.1](#), [2.2](#) and [4.1](#) and [Model Assumption 3.1](#) hold. Assume that [Algorithm 1](#) generates infinitely many successful iterations. Let $\theta_{\min}$ be as in [Lemma 4.3](#). Define*

$$\begin{aligned}\hat{\kappa}_0 &:= \frac{R^2}{\eta_1 \kappa_{\text{mdc}}^2 (1 - \eta_2) \theta_{\min}}, \quad \hat{\kappa}_\alpha := \frac{L_1 (1 + (\kappa_{\text{mdc}} (1 - \eta_2))^\alpha)}{R^\alpha (1 - \alpha)} \hat{\kappa}_0, \\ \hat{\kappa}_1 &:= \frac{L_1 (1 + \kappa_{\text{mdc}} (1 - \eta_2))}{R} \hat{\kappa}_0, \quad \hat{\kappa}_2 := L_0 \hat{\kappa}_0.\end{aligned}$$

Then, if $\alpha \in [0, 1)$,

$$|\hat{\mathcal{S}}(\epsilon)| \leq \hat{\kappa}_\alpha \epsilon^{\alpha-1} + \hat{\kappa}_2 \epsilon^{-1} - \hat{\kappa}_\alpha \delta_0^{\alpha-1}. \quad (29)$$

If $\alpha = 1$,

$$|\hat{\mathcal{S}}(\epsilon)| \leq \hat{\kappa}_1 \log(\delta_0/\epsilon) + \hat{\kappa}_2 \epsilon^{-1}. \quad (30)$$

Proof. If $\hat{\mathcal{S}}(\epsilon) = \emptyset$, then the theorem holds trivially as $|\hat{\mathcal{S}}(\epsilon)| = 0$. Otherwise, consider some $\ell \in \hat{\mathcal{S}}(\epsilon)$. By [Lemma 4.4](#), we have

$$\begin{aligned}f(x_\ell) - f(x_\ell + s_\ell) &\geq \eta_1 (m_\ell(0) - m_\ell(s_\ell)) \\ &\geq \eta_1 \kappa_{\text{mdc}} \|\nabla f(x_\ell)\| h(\min_{j \leq \ell} \|\nabla f(x_j)\|, \kappa_{\text{mdc}} (1 - \eta_2) \psi(\min_{j \leq \ell} \|\nabla f(x_j)\|)) \theta_{\min}.\end{aligned}$$

Using [Lemma 4.2](#), we have

$$\begin{aligned}&f(x_\ell) - f(x_\ell + s_\ell) \\ &\geq \frac{\frac{\eta_1 \kappa_{\text{mdc}}^2 \|\nabla f(x_\ell)\| (1 - \eta_2) \min_{j \leq \ell} \|\nabla f(x_j)\|}{L_0 + L_1 \min_{j \leq \ell} \|\nabla f(x_j)\|^\alpha}}{1 + \frac{L_1 (\kappa_{\text{mdc}} (1 - \eta_2) \min_{j \leq \ell} \|\nabla f(x_j)\| / (L_0 + L_1 \min_{j \leq \ell} \|\nabla f(x_j)\|^\alpha))^\alpha}{(L_0 + L_1 \min_{j \leq \ell} \|\nabla f(x_j)\|^\alpha)^{1-\alpha}}} \theta_{\min} \\ &\geq \eta_1 \kappa_{\text{mdc}}^2 (1 - \eta_2) \frac{\min_{j \leq \ell} \|\nabla f(x_j)\|^2}{L_0 + L_1 (1 + (\kappa_{\text{mdc}} (1 - \eta_2))^\alpha) \min_{j \leq \ell} \|\nabla f(x_j)\|^\alpha} \theta_{\min}.\end{aligned}$$

Since δ_j is nonincreasing, [Lemma 4.8](#) implies that

$$\delta_j \leq R \|\nabla f(x_j)\| \implies \delta_\ell = \min_{j=0, \dots, \ell} \delta_j \leq R \min_{j=0, \dots, \ell} \|\nabla f(x_j)\|.$$

Thus, recalling (28), we obtain

$$\delta_\ell - \delta_{\ell+1} = f(x_\ell) - f(x_\ell + s_\ell) \geq \frac{\eta_1 \kappa_{\text{mdc}}^2 (1 - \eta_2) (\delta_\ell / R)^2}{L_0 + L_1 (1 + (\kappa_{\text{mdc}} (1 - \eta_2))^\alpha) (\delta_\ell / R)^\alpha} \theta_{\min}.$$

We set $\omega := \eta_1 \kappa_{\text{mdc}}^2 (1 - \eta_2) \theta_{\min}$ and $L_\alpha := L_1 (1 + (\kappa_{\text{mdc}} (1 - \eta_2))^\alpha)$ and define the function $\varphi : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ as $\varphi(t) := \frac{t^2}{L_0 + L_\alpha t^\alpha}$, so that $\hat{\kappa}_0 = R^2 / \omega$. Therefore, we have $\delta_\ell - \delta_{\ell+1} \geq \omega \varphi(\delta_\ell / R)$. The function φ is increasing. Hence, for every $t \in [\delta_{\ell+1}, \delta_\ell]$,

$$\omega \leq \frac{\delta_\ell - \delta_{\ell+1}}{\varphi(\delta_\ell / R)} \leq \int_{\delta_{\ell+1}}^{\delta_\ell} \frac{1}{\varphi(t/R)} dt = \int_{\delta_{\ell+1}}^{\delta_\ell} \left(\frac{L_0 R^2}{t^2} + \frac{L_\alpha R^{2-\alpha}}{t^{2-\alpha}} \right) dt.$$

We sum this inequality over $\ell \in \hat{\mathcal{S}}(\epsilon)$. Since $\{\delta_k\}$ is nonincreasing over all iterations, the intervals $[\delta_{\ell+1}, \delta_\ell]$, $\ell \in \hat{\mathcal{S}}(\epsilon)$, are pairwise disjoint and contained in $[\delta_{\hat{\kappa}_\epsilon}, \delta_0]$, so that

$$\omega |\hat{\mathcal{S}}(\epsilon)| \leq \int_{\delta_{\hat{\kappa}_\epsilon}}^{\delta_0} \left(\frac{L_0 R^2}{t^2} + \frac{L_\alpha R^{2-\alpha}}{t^{2-\alpha}} \right) dt.$$

Hence, if $\alpha \in [0, 1)$, then using the fact that $\delta_{\widehat{\kappa}_\epsilon} > \epsilon$,

$$\omega|\widehat{\mathcal{S}}(\epsilon)| \leq L_0 R^2 \epsilon^{-1} + \frac{L_\alpha R^{2-\alpha}}{1-\alpha} (\epsilon^{\alpha-1} - \delta_0^{\alpha-1})$$

If $\alpha = 1$, then

$$\omega|\widehat{\mathcal{S}}(\epsilon)| \leq L_0 R^2 \epsilon^{-1} + L_\alpha R \log(\delta_0/\epsilon). \quad \square$$

We next specialize the analysis to strongly convex objectives.

Problem Assumption 4.2. There exists $\mu > 0$ such that,

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2} \mu \|y - x\|^2 \quad \text{for all } x, y \in \mathbb{R}^n.$$

Strong convexity yields a sharper relation between the distance to the global minimum and the gradient norm, which yields a better complexity bound.

Lemma 4.10 (6, Lemma 2.6). *Assume that Problem Assumption 4.2 holds. Then, for all $k \in \mathbb{N}$, $\delta_k \leq \frac{1}{2\mu} \|\nabla f(x_k)\|^2$.*

We now state the main complexity result for strongly convex objectives.

Theorem 4.11. *Let Problem Assumptions 2.1, 2.2 and 4.2 and Model Assumption 3.1 hold. Assume that Algorithm 1 generates infinitely many successful iterations. Let $\theta_{\min}$ be as in Lemma 4.3. Define*

$$\widehat{\kappa}_0 := \frac{1}{2\mu\eta_1\kappa_{\text{mdc}}^2(1-\eta_2)\theta_{\min}}, \quad \widehat{\kappa}_\alpha := L_1(1 + (\kappa_{\text{mdc}}(1-\eta_2))^\alpha)\widehat{\kappa}_0,$$

Then if $\alpha = 0$,

$$|\widehat{\mathcal{S}}(\epsilon)| \leq (L_0 + 2L_1)\widehat{\kappa}_0 \log(\delta_0/\epsilon). \quad (31)$$

Otherwise, if $\alpha \in (0, 1]$,

$$|\widehat{\mathcal{S}}(\epsilon)| \leq L_0\widehat{\kappa}_0 \log(\delta_0/\epsilon) + \frac{2^{1+\alpha/2}\widehat{\kappa}_\alpha\mu^{\alpha/2}}{\alpha} (\delta_0^{\alpha/2} - \epsilon^{\alpha/2}). \quad (32)$$

Proof. If $\widehat{\mathcal{S}}(\epsilon) = \emptyset$, then the theorem holds trivially as $|\widehat{\mathcal{S}}(\epsilon)| = 0$. Otherwise, consider some $\ell \in \widehat{\mathcal{S}}(\epsilon)$. By Lemmas 4.2 and 4.4, we have

$$f(x_\ell) - f(x_\ell + s_\ell) \geq \frac{\eta_1\kappa_{\text{mdc}}^2(1-\eta_2) \min_{j \leq \ell} \|\nabla f(x_j)\|^2}{L_0 + L_1(1 + (\kappa_{\text{mdc}}(1-\eta_2))^\alpha) \min_{j \leq \ell} \|\nabla f(x_j)\|^\alpha} \theta_{\min}.$$

Furthermore, due to the decreasing property of δ_k and the application of Lemma 4.10, we conclude that for all $j \in \{0, \dots, \ell\}$,

$$\delta_j \leq \frac{1}{2\mu} \|\nabla f(x_j)\|^2 \implies \delta_\ell = \min_{j=0, \dots, \ell} \delta_j \leq \frac{1}{2\mu} \min_{j=0, \dots, \ell} \|\nabla f(x_j)\|^2.$$

Thus, recalling (28), we obtain

$$\delta_\ell - \delta_{\ell+1} = f(x_\ell) - f(x_\ell + s_\ell) \geq \frac{2\eta_1\kappa_{\text{mdc}}^2(1-\eta_2)\mu\delta_\ell}{L_0 + L_1(1 + (\kappa_{\text{mdc}}(1-\eta_2))^\alpha)\sqrt{2\mu\delta_\ell}^\alpha} \theta_{\min}.$$

Note that the map $m \mapsto m^2/(L_0 + L_\alpha m^\alpha)$ is increasing on $(0, \infty)$, which is what licenses substituting the lower bound $\min_{j \leq \ell} \|\nabla f(x_j)\| \geq \sqrt{2\mu\delta_\ell}$ in the display above. We set $\omega := \eta_1\kappa_{\text{mdc}}^2(1-\eta_2)\theta_{\min}$ and $L_\alpha := L_1(1 + (\kappa_{\text{mdc}}(1-\eta_2))^\alpha)$ and define the function $\varphi: \mathbb{R}_+ \rightarrow \mathbb{R}_+$ as $\varphi(t) := \frac{t}{L_0 + L_\alpha(\sqrt{2\mu}t)^\alpha}$, so that

$\hat{\kappa}_0 = 1/(2\mu\omega)$. Therefore, we have $\delta_\ell - \delta_{\ell+1} \geq 2\mu\omega \varphi(\delta_\ell)$. The function φ is increasing. Hence, for every $t \in [\delta_{\ell+1}, \delta_\ell]$,

$$2\mu\omega \leq \frac{\delta_\ell - \delta_{\ell+1}}{\varphi(\delta_\ell)} \leq \int_{\delta_{\ell+1}}^{\delta_\ell} \frac{1}{\varphi(t)} dt = \int_{\delta_{\ell+1}}^{\delta_\ell} \left(\frac{L_0}{t} + \frac{L_\alpha (2\mu)^{\alpha/2}}{t^{1-\alpha/2}} \right) dt.$$

We sum this inequality over $\ell \in \widehat{\mathcal{S}}(\epsilon)$. As above, $\{\delta_k\}$ is nonincreasing over all iterations, so the intervals $[\delta_{\ell+1}, \delta_\ell]$, $\ell \in \widehat{\mathcal{S}}(\epsilon)$, are pairwise disjoint and contained in $[\delta_{\widehat{\kappa}_\epsilon}, \delta_0]$, whence

$$2\mu\omega |\widehat{\mathcal{S}}(\epsilon)| \leq \int_{\delta_{\widehat{\kappa}_\epsilon}}^{\delta_0} \left(\frac{L_0}{t} + \frac{L_\alpha (2\mu)^{\alpha/2}}{t^{1-\alpha/2}} \right) dt.$$

If $\alpha = 0$, using $\delta_{\widehat{\kappa}_\epsilon} > \epsilon$, we have

$$2\mu\omega |\widehat{\mathcal{S}}(\epsilon)| \leq (L_0 + 2L_1) \log(\delta_0/\epsilon).$$

Otherwise, if $\alpha \in (0, 1]$, then

$$2\mu\omega |\widehat{\mathcal{S}}(\epsilon)| \leq L_0 \log(\delta_0/\epsilon) + \frac{2^{1+\alpha/2} L_\alpha \mu^{\alpha/2}}{\alpha} (\delta_0^{\alpha/2} - \epsilon^{\alpha/2}),$$

which gives the desired result. $\square$

Several remarks on [Theorems 4.9](#) and [4.11](#) are in order. For $\alpha = 0$, which corresponds to the L -smooth case, where $L = L_0 + L_1$, we recover the standard $O(L\epsilon^{-1})$ convex and $O(L \log(\epsilon^{-1}))$ strongly convex bounds. For $\alpha \in (0, 1)$, our bounds are $O(L_0\epsilon^{-1} + L_1\epsilon^{\alpha-1})$ and $O(L_0 \log(\epsilon^{-1}) + L_1)$, respectively, which improves the dependence on L_1 in [\[15\]](#); see [Table 2](#). The strongly convex bound appears to be new. For $\alpha = 1$, the convex bound $O(L_0\epsilon^{-1} + L_1 \log(\epsilon^{-1}))$ matches the order obtained for gradient descent in [\[20\]](#), but without requiring prior knowledge of L_0 or L_1 . In the convex case, our constants in [Tables 1](#) and [2](#) differ slightly from those in [\[10, 14, 15, 19, 20\]](#) with respect to the radius R : these works state their bounds in terms of $R_0 = \|x_0 - x^*\|$, whereas ours involve the radius $R \geq R_0$ of the sublevel set $\{x : f(x) \leq f(x_0)\}$ around x^* ([Problem Assumption 4.1](#)). This is because, unlike the gradient methods analyzed in those works, trust-region methods do not guarantee that $\|x_k - x^*\|$ is nonincreasing; they only guarantee that $f(x_k)$ is, so that the iterates remain in this sublevel set. In the strongly convex case, our constants in [Tables 1](#) and [2](#) also differ from those in [\[10, 14, 15\]](#) because we measure convergence by the objective gap rather than the distance to a solution. Despite this difference, the resulting complexity bounds have the same order in ϵ , while the displayed L_1 -dependent term is linear in L_1 , whereas Gorbunov et al. [\[10\]](#)'s is quadratic.

4.3 Sharpness of the complexity bound

Let $\alpha \in [0, 1]$ and let $\epsilon \in (0, \frac{1}{2}]$ denote the desired accuracy. Define

$$g_\epsilon := 2\epsilon, \quad \bar{k}_\epsilon := \lfloor 2g_\epsilon^{\alpha-2} \rfloor + 1 = \lfloor 2^{\alpha-1} \epsilon^{\alpha-2} \rfloor + 1. \quad (33)$$

Our goal is to construct a continuously differentiable, bounded-below, $(\alpha, 0, 1)$ -smooth function $f_{\alpha, \epsilon} : \mathbb{R} \rightarrow \mathbb{R}$ for which the trust-region method requires exactly $\bar{k}_\epsilon$ iterations to produce $x_{\bar{k}_\epsilon}$ satisfying $|f'_{\alpha, \epsilon}(x_{\bar{k}_\epsilon})| \leq \epsilon$.

Set $\vartheta_\alpha := \alpha^{1/(1-\alpha)}$ if $\alpha \in [0, 1)$ and $\vartheta_1 := e^{-1}$. For $k \in \{0, \dots, \bar{k}_\epsilon\}$, define

$$g_k := \begin{cases} -g_\epsilon, & 0 \leq k < \bar{k}_\epsilon, \\ -\vartheta_\alpha g_\epsilon, & k = \bar{k}_\epsilon. \end{cases} \quad (34)$$

Because $\vartheta_\alpha \leq e^{-1} < \frac{1}{2}$, we have $|g_k| = 2\epsilon > \epsilon$ for $k = 0, \dots, \bar{k}_\epsilon - 1$, whereas $|g_{\bar{k}_\epsilon}| = 2\vartheta_\alpha \epsilon < \epsilon$.

For every $k = 0, \dots, \bar{k}_\epsilon - 1$, define

$$B_k := |g_k|^\alpha = g_\epsilon^\alpha. \quad (35)$$

Thus, [Model Assumption 3.1](#) holds with $b_0 = 0$ and $b_1 = 1$.

Set $\Delta_0 \geq 1$ and, for every $k = 0, \dots, \bar{k}_\epsilon - 1$, define

$$x_0 := 0, \quad s_k := -B_k^{-1} g_k = g_\epsilon^{1-\alpha}, \quad x_{k+1} := x_k + s_k. \quad (36)$$

Consequently, $x_k = k g_\epsilon^{1-\alpha}$ for $k = 0, \dots, \bar{k}_\epsilon$.

Finally, prescribe the function values at the iterates by

$$f_k := \begin{cases} \frac{g_\epsilon^{2-\alpha}}{2-\alpha} + (\bar{k}_\epsilon - 1 - k) g_\epsilon^{2-\alpha}, & 0 \leq k < \bar{k}_\epsilon, \\ \frac{g_\epsilon^{2-\alpha}}{2-\alpha} \vartheta_\alpha^{2-\alpha}, & k = \bar{k}_\epsilon. \end{cases} \quad (37)$$

The next lemma establishes basic properties of $\{f_k\}$.

Lemma 4.12. *The sequence $\{f_k\}$ defined in (37) is non-increasing and satisfies $f_k \in [0, f_0]$ for every $k = 0, \dots, \bar{k}_\epsilon$, where $\bar{k}_\epsilon$ is defined in (33). Moreover, $f_0 \leq 3$.*

Proof. For every $k = 0, \dots, \bar{k}_\epsilon - 2$, $f_{k+1} - f_k = -g_\epsilon^{2-\alpha} < 0$. For the final pair, $f_{\bar{k}_\epsilon} - f_{\bar{k}_\epsilon-1} = \frac{g_\epsilon^{2-\alpha}}{2-\alpha} (\vartheta_\alpha^{2-\alpha} - 1) < 0$, because $\vartheta_\alpha < 1$. Hence, $\{f_k\}$ is non-increasing. Since every term in (37) is nonnegative, we obtain $0 \leq f_k \leq f_0$ for every $k = 0, \dots, \bar{k}_\epsilon$.

Finally, using (33),

$$f_0 \leq \frac{g_\epsilon^{2-\alpha}}{2-\alpha} + 2g_\epsilon^{\alpha-2} g_\epsilon^{2-\alpha} = \frac{g_\epsilon^{2-\alpha}}{2-\alpha} + 2 \leq 3,$$

where the last inequality follows from $g_\epsilon = 2\epsilon \leq 1$ and $2 - \alpha \geq 1$. $\square$

The next theorem shows that the complexity bound of order $\epsilon^{\alpha-2}$ is sharp for (α, L_0, L_1) -smooth objectives with $L_0 = 0$.

Theorem 4.13. *Let $\alpha \in [0, 1]$ and $0 < \epsilon \leq \frac{1}{2}$. [Algorithm 1](#), with model Hessians satisfying [Model Assumption 3.1](#) with $b_0 = 0$ and $b_1 = 1$, applied to a continuously differentiable, bounded-below, $(\alpha, 0, 1)$ -smooth function $f: \mathbb{R} \rightarrow \mathbb{R}$, may require $\bar{k}_\epsilon$ iterations to produce $x_{\bar{k}_\epsilon}$ such that $|f'(x_{\bar{k}_\epsilon})| \leq \epsilon$, where $\bar{k}_\epsilon$ is defined in (33).*

Proof. Set $\ell_{\alpha,\epsilon} := g_\epsilon^{1-\alpha}$, $z_{\alpha,\epsilon} := (\bar{k}_\epsilon - 1)\ell_{\alpha,\epsilon}$, $c_{\alpha,\epsilon} := g_\epsilon^{2-\alpha}/(2-\alpha)$ and set $w_{\alpha,\epsilon} := z_{\alpha,\epsilon} + g_\epsilon^{1-\alpha}/(1-\alpha)$, with the convention that $w_{\alpha,\epsilon} = +\infty$ when $\alpha = 1$. Consider

$$f_{\alpha,\epsilon}(x) := \begin{cases} c_{\alpha,\epsilon} + g_\epsilon(z_{\alpha,\epsilon} - x), & x \leq z_{\alpha,\epsilon}, \\ \frac{1}{2-\alpha} [g_\epsilon^{1-\alpha} - (1-\alpha)(x - z_{\alpha,\epsilon})]^{\frac{2-\alpha}{1-\alpha}}, & z_{\alpha,\epsilon} < x < w_{\alpha,\epsilon}, \quad \alpha < 1, \\ 0, & x \geq w_{\alpha,\epsilon}, \quad \alpha < 1, \\ g_\epsilon e^{z_{\alpha,\epsilon} - x}, & x > z_{\alpha,\epsilon}, \quad \alpha = 1. \end{cases} \quad (38)$$

We now show that $f_{\alpha,\epsilon}$ is continuously differentiable and $(\alpha, 0, 1)$ -smooth. Suppose first that $\alpha \in [0, 1)$. On the intervals $I_{\alpha,1} = (-\infty, z_{\alpha,\epsilon}]$, $I_{\alpha,2} = [z_{\alpha,\epsilon}, w_{\alpha,\epsilon}]$, and $I_{\alpha,3} = [w_{\alpha,\epsilon}, +\infty)$, the function is respectively affine, polynomial-type, and constant. At the first junction, the adjacent pieces satisfy

$f_{\alpha,\epsilon}(z_{\alpha,\epsilon}) = c_{\alpha,\epsilon}$, $f'_{\alpha,\epsilon}(z_{\alpha,\epsilon}) = -g_\epsilon$. At the second junction, they satisfy $f_{\alpha,\epsilon}(w_{\alpha,\epsilon}) = f'_{\alpha,\epsilon}(w_{\alpha,\epsilon}) = 0$. Thus, $f_{\alpha,\epsilon} \in \mathcal{C}^1(\mathbb{R})$. Moreover, on $I_{\alpha,1}$ and $I_{\alpha,3}$ we have $|f''_{\alpha,\epsilon}| = 0$, while on $I_{\alpha,2}$, $|f''_{\alpha,\epsilon}(x)| = |f'_{\alpha,\epsilon}(x)|^\alpha$. Hence, $f_{\alpha,\epsilon}$ is $(\alpha, 0, 1)$ -smooth on each $I_{\alpha,i}$. Applying [Lemma A.3](#) first at $z_{\alpha,\epsilon}$ on $(-\infty, w_{\alpha,\epsilon}]$, and then at $w_{\alpha,\epsilon}$ on $\mathbb{R}$, shows that $f_{\alpha,\epsilon}$ is $(\alpha, 0, 1)$ -smooth on $\mathbb{R}$.

When $\alpha = 1$, consider $I_{1,1} = (-\infty, z_{1,\epsilon}]$ and $I_{1,2} = [z_{1,\epsilon}, +\infty)$. The function is affine on $I_{1,1}$ and exponential on $I_{1,2}$, and the two pieces satisfy $f_{1,\epsilon}(z_{1,\epsilon}) = g_\epsilon$, $f'_{1,\epsilon}(z_{1,\epsilon}) = -g_\epsilon$. Therefore, $f_{1,\epsilon} \in \mathcal{C}^1(\mathbb{R})$. Furthermore, $|f''_{1,\epsilon}| = 0$ on $I_{1,1}$ and $|f''_{1,\epsilon}| = |f'_{1,\epsilon}|$ on $I_{1,2}$. An application of [Lemma A.3](#) at $z_{1,\epsilon}$ shows that $f_{1,\epsilon}$ is $(1, 0, 1)$ -smooth on $\mathbb{R}$.

We now verify that the sequences defined in (34)–(37) constitute a valid run of [Algorithm 1](#). From (35) and (36), $B_k = g_\epsilon^\alpha$, $s_k = \ell_{\alpha,\epsilon}$, and $x_k = k\ell_{\alpha,\epsilon}$. In particular, $x_{\bar{k}_\epsilon-1} = z_{\alpha,\epsilon}$, and direct substitution in (38) gives $f_{\alpha,\epsilon}(x_k) = f_k$ and $f'_{\alpha,\epsilon}(x_k) = g_k$.

The quadratic model is $m_k(s) = f_k - g_\epsilon s + \frac{1}{2}g_\epsilon^\alpha s^2$. Its unique minimizer is $s_k = g_\epsilon^{1-\alpha}$, and its predicted reduction is $\frac{1}{2}g_\epsilon^{2-\alpha}$. Because $g_\epsilon \leq 1$, $|s_k| \leq 1$. Starting from $\Delta_0 \geq 1$, an induction shows that s_k lies in the trust region at every iteration. Indeed, for $k = 0, \dots, \bar{k}_\epsilon - 2$, the objective is affine on $[x_k, x_{k+1}]$, and hence $\rho_k = 2$.

At the final iteration, direct evaluation of (38) gives

$$\rho_{\bar{k}_\epsilon-1} = \frac{2}{2-\alpha} (1 - \vartheta_\alpha^{2-\alpha}) \geq 1.$$

Indeed, $\vartheta_\alpha^{2-\alpha} \leq \alpha/2$ for $\alpha \in [0, 1)$, while $\vartheta_1 = e^{-1} < 1/2$. Thus, every iteration is very successful, $\Delta_{k+1} \geq \Delta_k$, and the induction is complete. Since s_k is the exact trust-region minimizer, the required model-decrease condition also holds.

Finally, (34) gives $|f'_{\alpha,\epsilon}(x_k)| = g_\epsilon = 2\epsilon > \epsilon$ for $k < \bar{k}_\epsilon$, whereas $|f'_{\alpha,\epsilon}(x_{\bar{k}_\epsilon})| = \vartheta_\alpha g_\epsilon < \epsilon$. Thus, $\bar{k}_\epsilon$ is the first successful termination index. Moreover, by (33), $\bar{k}_\epsilon > 2g_\epsilon^{\alpha-2} = 2^{\alpha-1}\epsilon^{\alpha-2}$, which proves the result. $\square$

5 Discussion

Several perspectives remain open. Our analysis is deterministic and concerns first-order stationarity, extending it to stochastic objectives would require additional assumptions. Moreover, sharpness is established only for the nonconvex case, leaving open whether the convex and strongly convex bounds are optimal. Finally, extending the analysis to broader generalized smoothness conditions, in particular $\alpha > 1$, remains an interesting direction for future work.

A Proofs

Proof of [Theorem 2.6](#). We split the proof into two parts.

Proof of (8) Let $\nu := \|\nabla f(x)\|$. Otherwise, set $d := -\nabla f(x)/\nu$, $q := L_0 + 2^\alpha L_1 \nu^\alpha > 0$ and $r := \nu/(2q)$. First claim is that $\|\nabla f(x + td)\| \leq 2\nu$ for all $t \in [0, r]$. Indeed, if $t_0 \leq r$ were the first time such that $\|\nabla f(x + t_0 d)\| = 2\nu$, then (4), applied on the segment $[x, x + t_0 d]$ on which $\|\nabla f\| \leq 2\nu$ by minimality of t_0 , gives $\|\nabla f(x + t_0 d) - \nabla f(x)\| \leq (L_0 + L_1(2\nu)^\alpha)t_0 \leq qr = \nu/2$, which would imply $\|\nabla f(x + t_0 d)\| \leq 3\nu/2$, a contradiction. Thus, $\|\nabla f(x + td) - \nabla f(x)\| \leq qt$ for every $t \in [0, r]$. Since $\langle \nabla f(x), d \rangle = -\nu$,

$$f(x + rd) - f(x) \leq \int_0^r (-\nu + qt) dt = -\frac{3\nu^2}{8q} \leq -\frac{\nu^2}{4q}.$$

As $f(x + rd) \geq f_{\text{low}}$, denoting $\delta_x := f(x) - f_{\text{low}}$, we obtain $\nu^2 \leq 4L_0\delta_x + 2^{\alpha+2}L_1\delta_x\nu^\alpha$. We use the implication $z^2 \leq a + bz^\alpha \Rightarrow z \leq \sqrt{2a} + (2b)^{1/(2-\alpha)}$, valid for $z, a, b \geq 0$ and $\alpha \in [0, 2)$. Indeed, either

$a \geq bz^\alpha$, in which case $z^2 \leq 2a$ and $z \leq \sqrt{2a}$, or $a < bz^\alpha$, in which case $z^2 \leq 2bz^\alpha$, that is, $z^{2-\alpha} \leq 2b$ and $z \leq (2b)^{1/(2-\alpha)}$. In both cases z is at most the sum of the two terms. Applying it here, we obtain

$$\|\nabla f(x)\| \leq \sqrt{8L_0\delta_x} + (2^{\alpha+3}L_1\delta_x)^{\frac{1}{2-\alpha}}.$$

Proof of (9) and (10) For $\alpha = 1$, the result follows from [20, Lemmas 2.5 and 2.6]. Assume now that $\alpha \in [0, 1)$. If $s = 0$, the result is trivial. For $t \in [0, 1]$, define $\chi(t) := \|s\| \int_0^t (L_0 + L_1 \|\nabla f(x + \tau s)\|^\alpha) d\tau$. By (α, L_0, L_1) -smoothness applied to x and $x + ts$,

$$\begin{aligned} \|\nabla f(x + ts) - \nabla f(x)\| &\leq \|ts\| \int_0^1 (L_0 + L_1 \|\nabla f(x + \tau ts)\|^\alpha) d\tau \\ &= \|s\| \int_0^t (L_0 + L_1 \|\nabla f(x + \tau s)\|^\alpha) d\tau = \chi(t), \end{aligned}$$

where the equality follows from the change of variable $\tau' := t\tau$. Since $\chi(0) = 0$, we have

$$\begin{aligned} \chi'(t) &= \|s\| (L_0 + L_1 \|\nabla f(x + ts)\|^\alpha) \leq \|s\| (L_0 + L_1 (\|\nabla f(x)\| + \chi(t))^\alpha) \\ &\leq \|s\| (L_0 + L_1 \|\nabla f(x)\|^\alpha + L_1 \chi(t)^\alpha) = \lambda + \beta \chi(t)^\alpha, \end{aligned}$$

where $\lambda := (L_0 + L_1 \|\nabla f(x)\|^\alpha) \|s\|$ and $\beta := L_1 \|s\|$. Here, we used $(a + b)^\alpha \leq a^\alpha + b^\alpha$, valid for $a, b \geq 0$ and $\alpha \in [0, 1]$. Integrating the preceding inequality gives $\chi(t) \leq \lambda t + \beta \int_0^t \chi(\tau)^\alpha d\tau$. For any $\varrho > 0$, define $\zeta_\varrho(t) := \lambda + \varrho + \beta \int_0^t \chi(\tau)^\alpha d\tau$. Since $t \leq 1$, we have $\chi(t) \leq \zeta_\varrho(t)$, and hence $\zeta'_\varrho(t) = \beta \chi(t)^\alpha \leq \beta \zeta_\varrho(t)^\alpha$. Because $\zeta_\varrho(t) > 0$,

$$\frac{d}{dt} \zeta_\varrho(t)^{1-\alpha} = (1-\alpha) \zeta_\varrho(t)^{-\alpha} \zeta'_\varrho(t) \leq (1-\alpha) \beta.$$

Integrating from 0 to 1 yields $\zeta_\varrho(1)^{1-\alpha} \leq (\lambda + \varrho)^{1-\alpha} + (1-\alpha)\beta$. Since $\chi(1) \leq \zeta_\varrho(1)$, letting $\varrho \downarrow 0$ gives

$$\chi(1) \leq (\lambda^{1-\alpha} + (1-\alpha)\beta)^{\frac{1}{1-\alpha}}.$$

Substituting the definitions of λ and β , we obtain

$$\begin{aligned} \chi(1) &\leq \|s\| \left((L_0 + L_1 \|\nabla f(x)\|^\alpha)^{1-\alpha} + (1-\alpha)L_1 \|s\|^\alpha \right)^{\frac{1}{1-\alpha}} \\ &= (L_0 + L_1 \|\nabla f(x)\|^\alpha) g(\|s\|, \|\nabla f(x)\|) \|s\|, \end{aligned}$$

where the final identity follows from (11) and $L_0 + L_1 \|\nabla f(x)\|^\alpha > 0$. This proves (9). Finally, the fundamental theorem of calculus gives

$$\begin{aligned} |f(x + s) - f(x) - \nabla f(x)^T s| &\leq \int_0^1 \|\nabla f(x + ts) - \nabla f(x)\| \|s\| dt \\ &\leq (L_0 + L_1 \|\nabla f(x)\|^\alpha) g(\|s\|, \|\nabla f(x)\|) \|s\|^2 \int_0^1 t dt \\ &= \frac{1}{2} (L_0 + L_1 \|\nabla f(x)\|^\alpha) g(\|s\|, \|\nabla f(x)\|) \|s\|^2, \end{aligned}$$

where we used (9) with ts in place of s and the monotonicity of $g(\cdot, \|\nabla f(x)\|)$. $\square$

Proof of Proposition 3.2. We consider each update separately.

(i) Limited-memory BFGS Iterating over at most m stored pairs and applying Burdakov et al. [5, Lemma 5] and Lemma A.1 shows that

$$\|B_k\| \leq \|B_{k,0}\| + \frac{mb_m}{\omega_{\text{BFGS}}}$$

where b_m is the bound defined in Lemma A.1, and hence B_k satisfies Model Assumption 3.1.

(ii) Modified limited-memory BFGS Using the same argument, one get that,

$$\|B_k\| \leq \|B_{k,0}\| + \frac{m}{\bar{\omega}_{\text{BFGS}}}$$

and the conclusion follows directly.

(iii) Limited-memory SR1 Let $z_{i_j} := y_{i_j} - B_{k,j}s_{i_j}$. Then,

$$\|B_{k,j+1}\| \leq \|B_{k,j}\| + \frac{\|z_{i_j}\|^2}{|z_{i_j}^T s_{i_j}|} \leq \|B_{k,j}\| + \frac{\|z_{i_j}\|}{\omega_{\text{SR1}}\|s_{i_j}\|} \leq \left(1 + \frac{1}{\omega_{\text{SR1}}}\right) \|B_{k,j}\| + \frac{\|y_{i_j}\|}{\omega_{\text{SR1}}\|s_{i_j}\|}.$$

Iterating over at most m stored pairs and applying [Lemma A.1](#) shows that

$$\|B_k\| \leq \left(1 + \frac{1}{\omega_{\text{SR1}}}\right)^m \|B_{k,0}\| + \left(\left(1 + \frac{1}{\omega_{\text{SR1}}}\right)^m - 1\right) b_m,$$

where b_m is defined in [Lemma A.1](#), and hence B_k satisfies [Model Assumption 3.1](#).

(iv) Modified limited-memory SR1 Using the same argument as in Aravkin et al. [1],

$$\|B_k\| \leq \|B_{k,0}\| + \frac{m}{\bar{\omega}_{\text{SR1}}}.$$

(v) Limited-memory PSB By the bound derived in Powell [16, Equation 2.13], iterating over at most m stored pairs and applying [Lemma A.1](#) shows that

$$\|B_k\| \leq \|B_{k,0}\| + 3mb_m,$$

where b_m is defined in [Lemma A.1](#), and hence B_k satisfies [Model Assumption 3.1](#).

(vi) Modified limited-memory PSB Let $z_{i_j} := y_{i_j} - B_{k,j}s_{i_j}$. Then,

$$\|B_{k,j+1} - B_{k,j}\| \leq \frac{2\|z_{i_j}\|}{\|s_{i_j}\|} + \frac{|z_{i_j}^T s_{i_j}|}{\|s_{i_j}\|^2} \leq \frac{3\|z_{i_j}\|}{\|s_{i_j}\|}.$$

Iterating over at most m stored pairs and using the safeguard implies,

$$\|B_k\| \leq \|B_{k,0}\| + \frac{3m}{\bar{\omega}_{\text{PSB}}}.$$

□

Proof of Lemma 4.1. Fix $b > 0$. By [Proposition 2.7 \(2\)](#), the mapping $a \mapsto g(a, b)$ is nondecreasing and satisfies $g(a, b) \geq 1$. Hence, for $0 < a_1 < a_2$, $\Psi(a_2, b) = a_2 g(a_2, b) \geq a_2 g(a_1, b) > a_1 g(a_1, b) = \Psi(a_1, b)$. Thus, $\Psi(\cdot, b)$ is strictly increasing.

Moreover, $\lim_{a \downarrow 0} \Psi(a, b) = \lim_{a \downarrow 0} a g(a, b) = 0$. For $\alpha > 0$ this follows from [Proposition 2.7 \(4\)](#); for $\alpha = 0$, the convention $0^0 = 1$ gives $g(a, b) = 1 + L_1/(L_0 + L_1)$, which is constant in a , so the limit is again 0. On the other hand, since $\Psi(a, b) = a g(a, b) \geq a$, we have $\lim_{a \uparrow \infty} \Psi(a, b) = +\infty$. Since $\Psi(\cdot, b)$ is continuous, it is a bijection from $(0, \infty)$ onto $(0, \infty)$. Consequently, h is well-defined by [\(22\)](#).

According to [Proposition 2.7 \(1\)](#), $\Psi \in \mathcal{C}^1((0, \infty)^2)$. Furthermore, the monotonicity of $g(\cdot, b)$ in [Proposition 2.7 \(2\)](#) implies $\partial_a g(a, b) \geq 0$, and hence

$$\partial_a \Psi(a, b) = g(a, b) + a \partial_a g(a, b) \geq g(a, b) \geq 1. \quad (39)$$

Define $F(a, b, t) := \Psi(a, b) - t$. For every $b, t > 0$, we have $F(h(b, t), b, t) = 0$, while $\partial_a F(h(b, t), b, t) > 0$ by [\(39\)](#). The implicit function theorem therefore yields $h \in \mathcal{C}^1((0, \infty)^2)$.

Differentiating (23) with respect to t gives

$$\partial_a \Psi(h(b, t), b) \partial_t h(b, t) = 1,$$

which proves the first identity in (24) because $\partial_a \Psi(h(b, t), b) > 0$ by (39).

Similarly, differentiation with respect to b gives

$$\partial_a \Psi(h(b, t), b) \partial_b h(b, t) + \partial_b \Psi(h(b, t), b) = 0.$$

By Proposition 2.7 (3), the mapping $b \mapsto g(a, b)$ is nonincreasing, so $\partial_b \Psi(a, b) = a \partial_b g(a, b) \leq 0$. This proves the second identity in (24). $\square$

Proof of Lemma 4.2. Fix $b > 0$ and $t > 0$. Since $g(h(b, t), b) \geq 1$ according to Proposition 2.7 and $t = h(b, t)g(h(b, t), b)$, we immediately obtain $h(b, t) \leq t$. For brevity, write $g := g(h(b, t), b)$.

Assume first that $0 \leq \alpha < 1$. Let $\mu := L_0 + L_1 b^\alpha$, $\nu := L_1/\mu^{1-\alpha}$. Then $g^{1-\alpha} = 1 + (1-\alpha)\nu h(b, t)^\alpha$, so

$$1 + \nu t^\alpha = 1 + \nu h(b, t)^\alpha g^\alpha = 1 + \frac{g - g^\alpha}{1 - \alpha}.$$

By the concavity of $x \mapsto x^\alpha$, $g^\alpha \leq \alpha g + 1 - \alpha$, hence $g \leq 1 + \nu t^\alpha$. Since $t = h(b, t)g$, it follows that

$$h(b, t) \geq \frac{t}{1 + \nu t^\alpha}.$$

For $\alpha = 1$, $g = e^{L_1 h}$ and $t = h e^{L_1 h}$. Using $e^x \leq 1 + x e^x$ for $x \geq 0$ gives $e^{L_1 h} \leq 1 + L_1 t$, and therefore

$$h = \frac{t}{e^{L_1 h}} \geq \frac{t}{1 + L_1 t}.$$

$\square$

Lemma A.1. Let Problem Assumptions 2.1 and 2.2 hold. Assume $k \in \mathcal{S}$, $s_k \neq 0$ and define $y_k := \nabla f(x_k + s_k) - \nabla f(x_k)$. Then, for every fixed $r > 0$,

$$\frac{\|y_k\|}{\|s_k\|} \leq b_m := \inf_{r>0} \max \left\{ L_0 + L_1 \bar{\nu}_r^\alpha, \frac{2g_{\max}}{r} \right\},$$

where

$$\bar{\nu}_r := \begin{cases} [(g_{\max} + L_0 r)^{1-\alpha} + (1-\alpha)L_1 r]^{\frac{1}{1-\alpha}}, & \alpha \in [0, 1), \\ \left(g_{\max} + \frac{L_0}{L_1}\right) e^{L_1 r} - \frac{L_0}{L_1}, & \alpha = 1, L_1 > 0, \\ g_{\max} + L_0 r, & \alpha = 1, L_1 = 0, \end{cases}$$

where $g_{\max} := \sqrt{8L_0(f(x_0) - f_{\text{low}})} + (2^{\alpha+3}L_1(f(x_0) - f_{\text{low}}))^{\frac{1}{2-\alpha}}$.

Proof. Let $r > 0$ be fixed. Fix $k \in \mathcal{S}$, and define $\phi(t) := \|\nabla f(x_k + t s_k)\|$ for $t \in [0, 1]$. By Theorem 2.6 (8), we have $\max\{\phi(0), \phi(1)\} \leq g_{\max}$. By (α, L_0, L_1) -smoothness,

$$\phi(t) \leq \phi(0) + \int_0^t (L_0 + L_1 \phi(\tau)^\alpha) \|s_k\| d\tau.$$

Assume first that $\|s_k\| \leq r$. For $\alpha \in [0, 1)$, we have

$$\phi(t) \leq g_{\max} + L_0 r + \int_0^t L_1 \|s_k\| \phi(\tau)^\alpha d\tau.$$

Applying [Lemma A.2](#) with $u = \phi$, $c = g_{\max} + L_0 r$, $f(\tau) = L_1 \|s_k\|$, and $w(z) = z^\alpha$,

$$\phi(t) \leq [(g_{\max} + L_0 r)^{1-\alpha} + (1-\alpha)L_1 t \|s_k\|]^{\frac{1}{1-\alpha}} \leq \bar{v}_r.$$

For $\alpha = 1$ and $L_1 > 0$, applying [Lemma A.2](#) with $u = \phi$, $c = \phi(0)$, $f(\tau) = \|s_k\|$, and $w(z) = L_0 + L_1 z$, gives

$$\phi(t) \leq \left(\phi(0) + \frac{L_0}{L_1} \right) e^{L_1 t \|s_k\|} - \frac{L_0}{L_1} \leq \bar{v}_r.$$

If $\alpha = 1$ and $L_1 = 0$, then directly $\phi(t) \leq \phi(0) + L_0 t \|s_k\| \leq g_{\max} + L_0 r = \bar{v}_r$.

Therefore,

$$\|y_k\| \leq \|s_k\| \sup_{t \in [0,1]} (L_0 + L_1 \phi(t)^\alpha) \leq (L_0 + L_1 \bar{v}_r^\alpha) \|s_k\|.$$

If $\|s_k\| > r$, the endpoint bounds give $\|y_k\| \leq \|\nabla f(x_k + s_k)\| + \|\nabla f(x_k)\| \leq 2g_{\max}$. Combining the two cases proves the result for every fixed $r > 0$. Taking the infimum over $r > 0$ gives the desired result. $\square$

Lemma A.2 (Bihari-LaSalle inequality [\[3\]](#)). *Let $T > 0$, let $u, f : [0, T] \rightarrow [0, \infty)$ be continuous, and let $w : [0, \infty) \rightarrow [0, \infty)$ be continuous and nondecreasing, with $w(r) > 0$ for every $r > 0$. Assume that, for all $t \in [0, T]$,*

$$u(t) \leq c + \int_0^t f(s)w(u(s)) ds,$$

where $c \geq 0$. Then

$$u(t) \leq G^{-1} \left(G(c) + \int_0^t f(s) ds \right), \quad G(r) := \int_{r_0}^r \frac{d\tau}{w(\tau)},$$

for all $t \in [0, T]$ for which the right-hand side is well-defined, where $r_0 > 0$ is arbitrary.

Lemma A.3. *Let $I \subseteq \mathbb{R}$ be an interval, $a \in I$, $I_- := \{x \in I : x \leq a\}$ and $I_+ := \{x \in I : x \geq a\}$. Let $\alpha \in [0, 1]$ and $L_0, L_1 \geq 0$. Suppose that $f \in \mathcal{C}^1(I)$ is (α, L_0, L_1) -smooth on I_- and I_+ . Then f is (α, L_0, L_1) -smooth on I .*

Proof. Consider $x \in I_-$ and $y \in I_+$. By the triangle inequality and generalized smoothness on I_- and I_+ ,

$$\begin{aligned} |f'(y) - f'(x)| &\leq |f'(y) - f'(a)| + |f'(a) - f'(x)| \\ &\leq (L_0 + L_1 \sup_{u \in [a, y]} |f'(u)|^\alpha) |y - a| + (L_0 + L_1 \sup_{u \in [x, a]} |f'(u)|^\alpha) |a - x| \\ &\leq (L_0 + L_1 \sup_{u \in [x, y]} |f'(u)|^\alpha) |y - x|. \end{aligned}$$

Hence, f is (α, L_0, L_1) -smooth on I . $\square$

References

- [1] A. Y. Aravkin, R. Baraldi, and D. Orban. [A proximal quasi-Newton trust-region method for nonsmooth regularized optimization](#). Technical Report arXiv:2103.15993v1, 2021.
- [2] J. Bignon, D. Orban, and P. Raynaud. [A framework around limited-memory partitioned quasi-newton methods](#). Cahier du GERAD G-2023-17, GERAD, 2023.
- [3] I. Bihari. [A generalization of a lemma of Bellman and its application to uniqueness problems of differential equations](#). Acta Math. Hungar., 7(1-2):81-94, 1956.
- [4] A. Bodard and P. Patrinos. [Escaping saddle points without Lipschitz smoothness: the power of nonlinear preconditioning](#). Advances in Neural Information Processing Systems, 38:124173-124208, 2026.

- [5] O. Burdakov, L. Gong, S. Zirkin, and Y. X.Yuan. [On efficiently combining limited-memory and trust-region techniques](#). *Math. Program. Comp.*, 9:101–134, 2017.
- [6] C. Cartis, N. I. Gould, and Ph. L. Toint. [Evaluation complexity of adaptive cubic regularization methods for convex unconstrained optimization](#). *Optim. Method Softw.*, 27(2):197–219, 2012.
- [7] C. Cartis, N. I. Gould, and Ph. L. Toint. [Evaluation Complexity of Algorithms for Nonconvex Optimization](#). Number 30 in MOS-SIAM Series on Optimization. SIAM, 2022.
- [8] Z. Chen, Y. Zhou, Y. Liang, and Z. Lu. [Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization](#). In *International Conference on Machine Learning*, pages 5396–5427, 2023.
- [9] A. R. Conn, N. I. M. Gould, and Ph. L. Toint. [Trust-Region Methods](#). Number 1 in MPS-SIAM Series on Optimization. SIAM, Philadelphia, USA, 2000.
- [10] E. Gorbunov, N. Tupitsa, S. Choudhury, A. Aliev, P. Richtárik, S. Horváth, and M. Takáč. [Methods for convex \$\(L_0, L_1\)\$ -smooth optimization: Clipping, acceleration, and adaptivity](#). In *International Conference on Learning Representations*, 2025.
- [11] G. Grapiglia, J. Yuan, and Y. Yuan. [On the convergence and worst-case complexity of trust-region and regularization methods for unconstrained optimization](#). *Math. Program.*, 152:491–520, 2015.
- [12] F. Hübler, J. Yang, X. Li, and N. He. [Parameter-agnostic optimization under relaxed smoothness](#). In *International Conference on Artificial Intelligence and Statistics*, 2024.
- [13] S. Jerad. [Fast adaptive tensor methods under local smoothness](#). *arXiv:2606.30225*, 2026.
- [14] A. Koloskova, H. Hendriks, and S. U. Stich. [Revisiting gradient clipping: Stochastic bias and tight convergence guarantees](#). In *International Conference on Machine Learning*, 2023.
- [15] H. Li, J. Qian, Y. Tian, A. Rakhlin, and A. Jadbabaie. [Convex and non-convex optimization under generalized smoothness](#). In *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc., 2023.
- [16] M. Powell. [On the convergence of a wide range of trust region methods for unconstrained optimization](#). *IMA J. Numer. Anal.*, 30(1):289–301, 2010.
- [17] A. Semenov, M. Jaggi, and N. Doikov. [Gradient-normalized smoothness for optimization with approximate hessians](#). In *International Conference on Learning Representations*, 2026.
- [18] Y. Takezawa, H. Bao, R. Sato, K. Niwa, and M. Yamada. [Polyak meets parameter-free clipped gradient descent](#). In *Advances in Neural Information Processing Systems*, volume 37, 2024.
- [19] A. Tyurin. [Toward a unified theory of gradient descent under generalized smoothness](#). In *International Conference on Machine Learning*, 2025.
- [20] D. Vankov, A. Rodomanov, A. Nedich, L. Sankar, and S. U. Stich. [Optimizing \$\(L_0, L_1\)\$ -smooth functions by gradient methods](#). In *International Conference on Learning Representations*, 2025.
- [21] B. Zhang, J. Jin, C. Fang, and L. Wang. [Improved analysis of clipping algorithms for non-convex optimization](#). *Advances in Neural Information Processing Systems*, 33:15511–15521, 2020.
- [22] J. Zhang, T. He, S. Sra, and A. Jadbabaie. [Why gradient clipping accelerates training: a theoretical justification for adaptivity](#). In *International Conference on Learning Representations*, 2020.