Les Cahiers du GERAD ISSN: 0711–2440

Complexity of trust-region methods in the presence of unbounded Hessian approximations

Y. Diouane, M. Laghdaf Habiboullah, D. Orban

G–2024–43

Août 2024

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 in the presence of unbounded Hessian approximations

Youssef Diouane
Mohamed Laghdaf Habiboullah
Dominique Orban

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

youssef.diouane@polymtl.ca
mohamed.habiboullah@polymtl.ca
dominique.orban@gerad.ca

Août 2024
Les Cahiers du GERAD
G–2024–43

Abstract : We extend traditional complexity analyses of trust-region methods for unconstrained, possibly nonconvex, optimization. Whereas most complexity analyses assume uniform boundedness of the model Hessians, we work with potentially unbounded model Hessians. Boundedness is not guaranteed in practical implementations, in particular ones based on quasi-Newton updates such as PSB, BFGS and SR1. Our analysis is conducted for a family of trust-region methods that includes most known methods as special cases. We examine two regimes of Hessian growth: one bounded by a power of the number of successful iterations, and one bounded by a power of the number of iterations. This allows us to formalize and confirm the profound intuition of Powell [16], who studied convergence under a special case of our assumptions, but whose proof contained complexity arguments. Specifically, for $0 \leq p < 1$, we establish sharp $O(\epsilon^{-2/(1-p)})$ evaluation complexity to find an ϵ-stationary point when model Hessians are $O(k^p)$, where k is the iteration counter. For $p = 1$, which is the case studied by Powell, we establish a sharp $O(\exp(c\epsilon^{-2}))$ evaluation complexity for a certain constant $c > 0$. This is as Powell suspected and is far worse than other bounds surmised elsewhere in the literature. We establish similar bounds when model Hessians are $O(|\mathcal{S}_k|^p)$, where $|\mathcal{S}_k|$ is the number of iterations where the step was accepted, up to iteration k. To the best of our knowledge, ours is the first work to provide complexity bounds when model Hessians grow linearly with $|\mathcal{S}_k|$ or at most linearly with k, which covers multiple quasi-Newton approximations.

Keywords : Complexity analysis, quasi-Newton methods, trust-region, unconstrained optimization

Résumé : Nous étendons les analyses de complexité traditionnelles des méthodes de régions de confiance pour l'optimisation sans contrainte, possiblement non convexe. Alors que la plupart des analyses de complexité supposent que les hessiennes du modèle sont uniformément bornées, nous travaillons avec des hessiennes de modèle potentiellement non bornées. La bornitude n'est pas garantie dans les implémentations pratiques, en particulier celles basées sur des mises à jour quasi-Newton telles que PSB, BFGS et SR1. Notre analyse est menée pour une famille de méthodes de régions de confiance qui inclut la plupart des méthodes connues comme cas particuliers. Nous examinons deux régimes de croissance de la Hessienne : l'un borné par une puissance du nombre d'itérations réussies, et l'autre borné par une puissance du nombre d'itérations total. Cela nous permet de formaliser et de confirmer la profonde intuition de Powell (2010), qui a étudié la convergence sous un cas spécial de nos hypothèses, mais dont la preuve contenait des arguments de complexité. Plus précisément, pour $0 \leq p < 1$, nous établissons une complexité d'évaluation en $O(\epsilon^{-2/(1-p)})$ pour trouver un point ϵ-stationnaire lorsque les hessiennes du modèle sont en $O(k^p)$, où k est le compteur d'itération. Pour $p = 1$, qui est le cas étudié par Powell, nous établissons une complexité d'évaluation en $O(\exp(c\epsilon^{-2}))$ pour une certaine constante $c > 0$. Comme Powell le soupçonnait, cette borne est bien pire que d'autres bornes spéculées ailleurs dans la littérature. Nous établissons des bornes similaires lorsque les hessiennes du modèle sont en $O(|\mathcal{S}_k|^p)$, où $|\mathcal{S}_k|$ est le nombre d'itérations réussies jusqu'à l'itération k. À notre connaissance, ce travail est le premier à fournir des bornes de complexité lorsque les hessiennes du modèle croissent linéairement avec $|\mathcal{S}_k|$ ou au plus linéairement avec k, ce qui couvre plusieurs approximations quasi-Newton.

Acknowledgements: We express our sincere gratitude to Coralia Cartis for insightful discussions that improved our arguments, and in particular, for bringing [16] to our attention. Research supported by an NSERC Discovery grant.

1 Introduction

We consider the solution of the nonconvex unconstrained problem

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

where $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable, by means of trust-region methods [2]. At each iteration of a trust-region method, a model of the objective is used to compute a step. In most cases, said model is a quadratic, and is required to match f up to first order at the current iterate. Although there are variants allowing inexact first-order information [2, §8.4], there is far more freedom in the model Hessian $B_k = B_k^T$ at iteration k. In particular, quasi-Newton approximations are a natural choice as they are known to yield fast local convergence under certain assumptions [4]. Convergence of trust-region methods, both local and global, has been studied extensively for the past few decades—see [2] and references therein—and their worst-case evaluation complexity has attracted much attention in the past decade and a half—see [1] and references therein. Although convergence of many trust-region schemes has been established under general assumptions, which allow for unbounded model Hessians, complexity has focused on cases where the latter remain uniformly bounded [1, 8], except for the work of Leconte and Orban [9], who study a trust-region method for nonsmooth optimization that does not reduce to a known classical method in the case of smooth optimization.

In practice, quasi-Newton updates, including PSB, BFGS, and SR1 are not guaranteed to result in uniformly bounded model Hessians. Our main contributions are to address the complexity of a family of trust-region methods when the model Hessians may be unbounded, and to expand upon the results of [9]. The family of methods includes most, if not all, known classical variants, and is parameterized by two scalars $\alpha \leq 1$ and $\beta \leq 1$ that control the presence of the gradient norm and the model Hessian norm in the trust-region radius.

Several authors show that usual quasi-Newton updates, including PSB, BFGS and SR1, grow at most linearly with k, e.g., [16] and [2, §8.4.1.2], although at the time of this writing, it is not known whether that bound is attained. However, in a typical implementation, those approximations are not updated when a step is rejected, only when a step is accepted. Thus, we also investigate the incidence on complexity of a weaker assumption; where the model Hessians grow at most linearly with the number of iterations in which a step is accepted, called *successful* iterations.

Specifically, we establish that if $\|B_k\| \leq \mu(1 + |\mathcal{S}_k|^p)$, where $0 \leq p < 1$ and $|\mathcal{S}_k|$ is the number of successful iterations up to iteration k, our family of trust-region methods may require as many as $O(\epsilon^{-2/(1-p)})$ iterations to identify an ϵ-stationary point. That is the same assumption and the same conclusion as Leconte and Orban [9], though with a better constant hidden inside the "O" term. However, we extend the complexity analysis to the case where $p = 1$, and establish that the complexity bound becomes $O(\exp(c_1\epsilon^{-2}))$ for a certain constant $c_1 > 0$. The leading term in those bounds is independent of α and β, and therefore the bounds apply to most know trust-region variants. The bounds are also shown to be sharp. Although it is not known whether quasi-Newton updates may indeed grow exactly linearly with $|\mathcal{S}_k|$, our analysis applies to them.

Attention then turns to the weaker assumption $\|B_k\| \leq \mu(1 + k^p)$, where again, $0 \leq p < 1$, which, among others, takes into account scenarios where B_k is also updated on unsuccessful iterations. We establish a complexity bound $O(\epsilon^{-2/(1-p)} + \epsilon^{(\alpha-1)/(1-p)})$. When $p = 1$, the complexity bound becomes $O(\exp(c_2\epsilon^{-2} + c_3\epsilon^{\alpha-1}))$, where $c_2 > 0$ and $c_3 > 0$ are constants. A choice $-1 \leq \alpha \leq 1$, which covers most known trust-region variants, results in bounds with leading terms of the same order of magnitude as under the previous assumption, and those bounds are shown to be sharp. The bound corresponding to $p = 1$ and $-1 \leq \alpha \leq 1$ cements a conjecture of Powell [16] described below. When $\alpha < -1$, we suspect that the bounds are not sharp and may be improved.

Powell [14] first investigated convergence of trust-region methods under the assumption $\|B_k\| \leq \mu(1 + \sum_{j=0}^{k-1} \|s_j\|)$, where $\mu > 0$ is a constant, and s_j is the step at iteration j. Among others, the PSB

quasi-Newton update [13] satisfies that condition [7]. The last paragraph of his paper mentions another researcher investigating the weaker condition $\|B_k\| \leq \mu(1+k)$, and Powell conjectured that convergence continues to hold under the weaker assumption. Powell [15] himself extended the convergence results of [14] under the weaker assumption, and Powell [16] further extended the previous analysis to a wide family of trust-region methods. In his conclusions, Powell [16] shared profound intuition on the number of iterations required to reduce the gradient norm, and described that number as "monstrous", hinting at potential exponential growth. To the best of our knowledge, the present complexity analysis is the first to confirm Powell's intuition.

The remainder of this paper is organized as follows. In Section 2, we introduce our family of trust-region methods along with key preliminary results. In Section 3, complexity bounds are derived in the case where the bound on model Hessians depends only on successful iterations. We also establish that the bounds are sharp. In Section 4, complexity bounds are derived in the case where the bound on model Hessians depends on the iteration counter. In Section 5, we investigate the performance in practice of different members of our family of trust-region methods, including classical schemes. Finally, we provide closing remarks in Section 6.

Notation

For a finite set $\mathcal{A}$, $|\mathcal{A}|$ denotes its cardinality. For a vector x and matrix B, $\|x\|$ and $\|B\|$ denote their ℓ_2 norm. We denote $\mathbb{N}_0$ the set of positive integers.

2 A family of trust-region methods

In this section, we describe a parameterized family of trust-region methods for (1), where the main novelty is the definition of a general trust-region radius in the subproblem.

2.1 Algorithm

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

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

where $B_k = B_k^T$ and $\Delta_k > 0$ is used to determine the trust-region radius. Specific choices of $\alpha \leq 1$ and $\beta \leq 1$ reduce (2) to known formulations. Namely, $\alpha = \beta = 0$ leads to the classical method [2], while $\alpha = 1$ and $\beta = 0$ reduces to the choice of the trust-region radius used by Curtis et al. [3]—note however, that the update of Δ_k differs in [3]. When $\beta = 0$, our scaled trust-region radius can be seen as a particular case of more general nonlinear step-size control mechanisms [8, 19]. Choosing $\beta \neq 0$ incorporates second-order information into the trust-region radius. We are not aware of other trust-region methods doing so, except for [9] in nonsmooth optimization. The parameters α and β are for generalization purposes, but also play a key role in our complexity analysis; a discussion on the values of these parameters, ensuring that the trust-region method enjoys favorable worst-case complexity bounds, is provided later.

The rest of the algorithm is standard. 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 both are in sufficient agreement, $x_k + s_k$ becomes the new iterate, and Δ_k is possibly increased. If the model turns out to predict poorly the actual decrease, the trial point is rejected and Δ_k is reduced. Algorithm 1 states the whole procedure.

Algorithm 1 A family of trust-region methods

Choose constants $0 < \eta_1 \leq \eta_2 < 1$, $0 < \gamma_1 \leq \gamma_2 < 1 \leq \gamma_3 \leq \gamma_4$ and $0 < \kappa_{\text{mdc}} \leq \frac{1}{2}$.
Choose $x_0 \in \mathbb{R}^n$ and $\Delta_0 > 0$.
Choose $(\alpha, \beta) \in (-\infty, 1]^2$.
for $k = 0, 1, \ldots$ **do**
  Choose $B_k = B_k^T \in \mathbb{R}^{n \times n}$.
  Compute $s_{k,\text{cp}}$ as in (3) and an approximate minimizer s_k of (2) satisfying (4).
  Compute the ratio

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

  If $\rho_k \geq \eta_1$, set $x_{k+1} := x_k + s_k$. Otherwise, set $x_{k+1} := x_k$.
  Update Δ_k according to

$$\Delta_{k+1} \in \begin{cases} [\gamma_3 \Delta_k, \gamma_4 \Delta_k] & \text{if } \rho_k \geq \eta_2, \quad \textit{(very successful iteration)} \\ [\gamma_2 \Delta_k, \Delta_k] & \text{if } \eta_1 \leq \rho_k < \eta_2, \quad \textit{(successful iteration)} \\ [\gamma_1 \Delta_k, \gamma_2 \Delta_k] & \text{if } \rho_k < \eta_1. \quad \textit{(unsuccessful iteration)} \end{cases}$$

Only an approximate solution of (2) is required; a step s_k should provide a decrease larger than or equal to a fraction of the decrease at the Cauchy point within the trust region. The Cauchy point $s_{k,\text{cp}}$ is defined as

$$s_{k,\text{cp}} := -t_k \nabla f(x_k), \quad \text{where} \quad t_k := \operatorname*{argmin}_{t \geq 0} \; m_k\left(-t \nabla f(x_k)\right) \text{ s.t. } \|t \nabla f(x_k)\| \leq \frac{\|\nabla f(x_k)\|^\alpha}{(1 + \|B_k\|)^\beta} \Delta_k. \quad (3)$$

The decrease required of s_k is stated as

$$m_k(0) - m_k(s_k) \geq 2\kappa_{\text{mdc}}(m_k(0) - m_k(s_{k,\text{cp}})), \quad (4)$$

where $0 < \kappa_{\text{mdc}} \leq \frac{1}{2}$.

In the remainder of this paper, we focus on the formulation (2), but setting the radius to

$$\frac{\min_{j=0,\ldots,k} \|\nabla f(x_j)\|^\alpha}{(1 + \max_{j=0,\ldots,k} \|B_j\|)^\beta} \Delta_k \quad \text{instead of} \quad \frac{\|\nabla f(x_k)\|^\alpha}{(1 + \|B_k\|)^\beta} \Delta_k,$$

leads to a similar complexity analysis.

We make the following standard assumption.

Problem Assumption 2.1. *The objective function f is continuously differentiable with Lipschitz continuous gradient, i.e., there exists $L > 0$ such that for all x, $y \in \mathbb{R}^n$, $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$. Moreover, f is bounded below by a constant f_{low}.*

Below, we allow $\{B_k\}$ to be unbounded. We study the effect of two different bounds on $\{\|B_k\|\}$ on the worst-case complexity. The next section provides results useful to both analyses.

2.2 Preliminary results

Our first result clarifies the Cauchy decrease (4). The proof is similar to that of, e.g., [1, Lemma 2.3.2] with the radius replaced with that of (2).

Lemma 1. *Let Problem Assumption 2.1 hold. For all $k \in \mathbb{N}$,*

$$m_k(0) - m_k\left(s_k\right) \geq \kappa_{\text{mdc}} \|\nabla f(x_k)\| \min\left\{ \frac{\|\nabla f(x_k)\|}{1 + \|B_k\|}, \frac{\|\nabla f(x_k)\|^\alpha}{(1 + \|B_k\|)^\beta} \Delta_k \right\}. \quad (5)$$

The second result follows immediately from [2, Theorem 8.4.2] using Lipschitz continuity of ∇f.

Lemma 2. *Let Problem Assumption 2.1 hold. For all $k \in \mathbb{N}$,*

$$|f(x_k + s_k) - m_k(s_k)| \leq \kappa(1 + \|B_k\|)\frac{\|\nabla f(x_k)\|^{2\alpha}}{(1 + \|B_k\|)^{2\beta}}\Delta_k^2, \tag{6}$$

where $\kappa := \max\{L, 1\}/2$.

The quantity

$$a_k := \frac{\Delta_k(1 + \max_{j=0,\ldots,k}\|B_j\|)^{1-\beta}}{(\min_{j=0,\ldots,k}\|\nabla f(x_j)\|)^{1-\alpha}} \quad (k \in \mathbb{N}) \tag{7}$$

plays a central role in the analysis. We first establish that $\{a_k\}$ is uniformly bounded below. The proof proceeds by induction in the spirit of [3, Lemma 2.4].

Lemma 3. *Let Problem Assumption 2.1 hold. For all $k \in \mathbb{N}$,*

$$a_k \geq a_{\min} := \min\left\{a_0, \gamma_1, \frac{\gamma_1\kappa_{\text{mdc}}(1 - \eta_2)}{\kappa}\right\} > 0,$$

where κ is defined in Lemma 2. Equivalently, for all $k \in \mathbb{N}$,

$$\Delta_k \geq \frac{(\min\limits_{j=0,\ldots,k}\|\nabla f(x_j)\|)^{1-\alpha}}{(1 + \max_{j=0,\ldots,k}\|B_j\|)^{1-\beta}}\, a_{\min}.$$

Proof. First note that $\{a_k/\Delta_k\}$ is non-decreasing because $\alpha \leq 1$ and $\beta \leq 1$.

We proceed by induction. For $k = 0$, $a_0 \geq a_{\min}$ by definition. We assume that $a_k \geq a_{\min}$, and show that $a_{k+1} \geq a_{\min}$ as well.

There are two cases. Assume first that $a_k \geq \min\left\{1, \frac{\kappa_{\text{mdc}}(1-\eta_2)}{\kappa}\right\}$. The non-decreasing nature of $\{a_k/\Delta_k\}$ and the update of Δ_k in Algorithm 1 ensure that

$$a_{k+1} = \frac{a_{k+1}}{\Delta_{k+1}}\Delta_{k+1} \geq \frac{a_k}{\Delta_k}\gamma_1\Delta_k = \gamma_1 a_k \geq \min\left\{\gamma_1, \frac{\gamma_1\kappa_{\text{mdc}}(1 - \eta_2)}{\kappa}\right\} \geq a_{\min}.$$

Now, assume conversely that $a_k < \min\left\{1, \frac{\kappa_{\text{mdc}}(1-\eta_2)}{\kappa}\right\}$. It follows from the definition of ρ_k and Δ_k, together with Lemmas 1 and 2 that

$$\begin{aligned}
|\rho_k - 1| = \frac{|f(x_k + s_k) - m_k(s_k)|}{m_k(0) - m_k(s_k)} &\leq \frac{\kappa(1 + \|B_k\|)^{1-2\beta}\|\nabla f(x_k)\|^{2\alpha}\Delta_k^2}{\kappa_{\text{mdc}}\|\nabla f(x_k)\| \min\left\{\frac{\|\nabla f(x_k)\|}{1+\|B_k\|}, \frac{\|\nabla f(x_k)\|^{\alpha}}{(1+\|B_k\|)^{\beta}}\Delta_k\right\}} \\
&\leq \frac{\kappa(1 + \|B_k\|)^{1-\beta}\|\nabla f(x_k)\|^{\alpha-1}\Delta_k^2}{\kappa_{\text{mdc}}\frac{(1+\|B_k\|)^{\beta}}{\|\nabla f(x_k)\|^{\alpha}} \min\left\{\frac{\|\nabla f(x_k)\|}{1+\|B_k\|}, \frac{\|\nabla f(x_k)\|^{\alpha}}{(1+\|B_k\|)^{\beta}}\Delta_k\right\}} \\
&= \frac{\kappa(1 + \|B_k\|)^{1-\beta}\|\nabla f(x_k)\|^{\alpha-1}\Delta_k^2}{\kappa_{\text{mdc}} \min\left\{\frac{\|\nabla f(x_k)\|^{1-\alpha}}{(1+\|B_k\|)^{1-\beta}}, \Delta_k\right\}}.
\end{aligned}$$

Because $\alpha - 1 \leq 0$ and $1 - \beta \geq 0$,

$$\begin{aligned}
|\rho_k - 1| &\leq \frac{\kappa(1 + \max_{j=0,\ldots,k}\|B_j\|)^{1-\beta}(\min_{j=0,\ldots,k}\|\nabla f(x_j)\|)^{\alpha-1}\Delta_k^2}{\kappa_{\text{mdc}} \min\left\{\frac{(\min_{j=0,\ldots,k}\|\nabla f(x_j)\|)^{1-\alpha}}{(1+\max_{j=0,\ldots,k}\|B_j\|)^{1-\beta}}, \Delta_k\right\}} \\
&= \frac{\kappa\left((1 + \max_{j=0,\ldots,k}\|B_j\|)^{1-\beta}(\min_{j=0,\ldots,k}\|\nabla f(x_j)\|)^{\alpha-1}\Delta_k\right)^2}{\kappa_{\text{mdc}} \min\{1, a_k\}} = \frac{\kappa a_k}{\kappa_{\text{mdc}}} \leq 1 - \eta_2.
\end{aligned}$$

Therefore, $\rho_k \geq \eta_2$, which implies that iteration k is very successful, and $\Delta_{k+1} > \Delta_k$. Thus, since $\{a_k/\Delta_k\}$ is non-decreasing,

$$a_{k+1} = \frac{a_{k+1}}{\Delta_{k+1}}\Delta_{k+1} > \frac{a_k}{\Delta_k}\Delta_k = a_k \geq a_{\min}. \qquad \square$$

Lemmas 1 and 3 allow us to quantify the model decrease in terms of $a_{\min}$.

Lemma 4. *Let Problem Assumption 2.1 hold. For all $k \in \mathbb{N}$, Algorithm 1 generates s_k satisfying*

$$m_k(0) - m_k(s_k) \geq \kappa_{\text{mdc}} \frac{\min_{j=0,\ldots,k} \|\nabla f(x_j)\|^2}{1 + \max_{j=0,\ldots,k} \|B_j\|} a_{\min}. \tag{8}$$

Proof. Lemmas 1 and 3 and the fact that $a_{\min} \leq \gamma_1 < 1$ combine to give

$$\begin{aligned}
m_k(0) - m_k(s_k) &\geq \kappa_{\text{mdc}} \|\nabla f(x_k)\| \min\left\{ \frac{\|\nabla f(x_k)\|}{1 + \|B_k\|}, \frac{\|\nabla f(x_k)\|^\alpha}{(1 + \|B_k\|)^\beta} \Delta_k \right\} \\
&\geq \kappa_{\text{mdc}} \min_{j=0,\ldots,k} \|\nabla f(x_j)\| \min\left\{ \frac{\min_{j=0,\ldots,k} \|\nabla f(x_j)\|}{1 + \max_{j=0,\ldots,k} \|B_j\|}, \frac{(\min_{j=0,\ldots,k} \|\nabla f(x_j)\|)^\alpha}{(1 + \max_{j=0,\ldots,k} \|B_j\|)^\beta} \Delta_k \right\} \\
&\geq \kappa_{\text{mdc}} \frac{(\min_{j=0,\ldots,k} \|\nabla f(x_j)\|)^2}{1 + \max_{j=0,\ldots,k} \|B_j\|} \min\{1, a_k\} \\
&\geq \kappa_{\text{mdc}} \frac{(\min_{j=0,\ldots,k} \|\nabla f(x_j)\|)^2}{1 + \max_{j=0,\ldots,k} \|B_j\|} \min\{1, a_{\min}\} = \kappa_{\text{mdc}} \frac{\min_{j=0,\ldots,k} \|\nabla f(x_j)\|^2}{1 + \max_{j=0,\ldots,k} \|B_j\|} a_{\min}.
\end{aligned}$$

$\square$

We conclude this section with a technical result.

Lemma 5. *Let $\mu > 0$ and $p > 0$. For any $k_1 \in \mathbb{N}$ and $k_2 \in \mathbb{N}$ such that $k_1 < k_2$,*

$$\sum_{k=k_1}^{k_2} \frac{1}{1 + \mu(1 + (k+1)^p)} \geq \frac{(k_1+1)^p}{(1 + \mu(1 + (k_1+1)^p))} \int_{k_1+1}^{k_2+2} \frac{1}{t^p} \mathrm{d}t.$$

Proof. Define $\phi : \mathbb{R}_+ \to \mathbb{R}$, $\phi(x) := x/(1 + \mu(1+x))$. Because ϕ is non-decreasing and $\int_k^{k+1} \frac{1}{t^p} \mathrm{d}t \leq \int_k^{k+1} \frac{1}{k^p} \mathrm{d}t = \frac{1}{k^p}$,

$$\begin{aligned}
\sum_{k=k_1}^{k_2} \frac{1}{1 + \mu(1 + (k+1)^p)} &= \sum_{k=k_1}^{k_2} \frac{1}{(k+1)^p} \phi((k+1)^p) \\
&\geq \phi((k_1+1)^p) \sum_{k=k_1+1}^{k_2+1} \frac{1}{k^p} \\
&\geq \phi((k_1+1)^p) \sum_{k=k_1+1}^{k_2+1} \int_k^{k+1} \frac{1}{t^p} \mathrm{d}t.
\end{aligned}$$

$\square$

In the next sections, we derive worst-case complexity analyses for Algorithm 1 that allows for potentially unbounded model Hessians B_k. We will repeatedly use the notation

$$\mathcal{S} := \{i \in \mathbb{N} \mid \rho_i \geq \eta_1\} \qquad \text{(all successful iterations)} \tag{9a}$$

$$\mathcal{S}_k := \{i \in \mathcal{S} \mid i \leq k\} \qquad \text{(successful iterations until iteration } k\text{)} \tag{9b}$$

$$\mathcal{U}_k := \{i \in \mathbb{N} \mid i \notin \mathcal{S},\ i \leq k\} \qquad \text{(unsuccessful iterations until iteration } k\text{)}. \tag{9c}$$

3 Complexity when the bound on $\|B_k\|$ depends only on successful iterations

Our assumption on the growth of the model Hessians is as follows.

Model Assumption 3.1. *There exist $\mu > 0$ and $0 \leq p \leq 1$ such that, for all $k \in \mathcal{S}$,*

$$\max_{j=0,\ldots,k} \|B_j\| \leq \mu(1 + |\mathcal{S}_k|^p). \tag{10}$$

Several comments on Model Assumption 3.1 are in order. Model Assumption 3.1 was first suggested and studied by Leconte and Orban [9], who required (10) to hold for all $k \in \mathbb{N}$. To the best of our knowledge, theirs was the first complexity analysis to allow for potentially unbounded model Hessians. Here, we require instead that (10) hold for all $k \in \mathcal{S}$, but note that that the maximum covers all indices $j = 0, \ldots, k$, which include unsuccessful iterations, if any. That is not particularly restrictive as in practice one typically updates the model Hessian approximation after a successful iteration only. In Section 4, we consider a bound that involves k instead of $|\mathcal{S}_k|$, which may be suitable if one updates the Hessian on unsuccessful iterations, but that leads to a worse complexity bound.

The special case $p = 0$, which corresponds to bounded model Hessians, is the only one considered in the literature for complexity analyses, e.g., [1–3]. Leconte and Orban [9] studied complexity when $0 \leq p < 1$ in the context of a trust-region method for nonsmooth optimization, although their results also apply to smooth optimization.

We believe that the present work is the first to address the case where $p = 1$; an important case as it encompasses known bounds on existing quasi-Newton approximations such as SR1. Indeed, Conn et al. [2, §8.4.1.2] show that the SR1 Hessian approximation satisfies

$$\|B_{k+1}\| \leq \|B_k\| + \mu_f,$$

where μ_f is a constant related to f and k is the index of a successful iteration for which the SR1 update is well defined. They provide a similar bound for the BFGS updates when f is convex. Powell [16] establishes a similar bound for a quasi-Newton update based on the PSB formula, where the bound on $\|B_k\|$ depends linearly on the number of successful iterations $|\mathcal{S}_k|$. Even though it is not currently known whether those bounds on SR1, BFGS and PSB are tight, Model Assumption 3.1 includes them when $p = 1$.

In Model Assumption 3.1, p is not allowed to take a value larger than 1. In fact, when $p > 1$, global convergence is no longer guaranteed as the series $\sum_{k=1}^{\infty} 1/(1 + |\mathcal{S}_k|^p)$ becomes convergent [18]. Finally, due to the non-decreasing nature of $\{|\mathcal{S}_k|\}_{k\in\mathbb{N}}$, (10) is equivalent to

$$\|B_k\| \leq \mu(1 + |\mathcal{S}_k|^p) \quad \text{for all } k \in \mathbb{N}. \tag{11}$$

If Algorithm 1 generates only a finite number of successful iterations, a first-order critical point is identified after a finite number of iterations. The following result parallels [2, Theorem 6.4.4].

Theorem 1. Let Problem Assumption 2.1 and Model Assumption 3.1 be satisfied. If Algorithm 1 generates finitely many successful iterations, then $x_k = x^*$ for all sufficiently large k where $\nabla f(x^*) = 0$.

Proof. Assume by contradiction that there exists $\nu > 0$ such that $\|\nabla f(x_k)\| \geq \nu$ for all $k \in \mathbb{N}$, and let k_f be the last successful iteration. Necessarily, $x_k = x_{k_f}$ for all $k \geq k_f$, and hence, $\{x_k\} \to x^* := x_{k_f}$. By Lemma 3, for all $k \geq k_f$,

$$\Delta_k \geq \frac{(\min_{j=0,\ldots,k} \|\nabla f(x_j)\|)^{1-\alpha}}{(1 + \max_{j=0,\ldots,k} \|B_j\|)^{1-\beta}} a_{\min} \geq \frac{\nu^{1-\alpha}}{(1 + \mu(1 + |\mathcal{S}_k|^p))^{1-\beta}} a_{\min} = \frac{\nu^{1-\alpha}}{(1 + \mu(1 + |\mathcal{S}_{k_f}|^p))^{1-\beta}} a_{\min} > 0,$$

where we used the fact that $\mathcal{S}_k = \mathcal{S}_{k_f}$. Thus, Δ_k is bounded away from zero. However, the mechanism of Algorithm 1 ensures that Δ_k decreases on unsuccessful iterations and converges to 0, which is a contradiction, and shows that $\liminf_{k\to\infty} \|\nabla f(x_k)\| = \|\nabla f(x^*)\| = 0$. □

Let $\epsilon > 0$ and k_ϵ be the first iteration of Algorithm 1 such that $\|\nabla f(x_{k_\epsilon})\| \leq \epsilon$. Define

$$\mathcal{S}(\epsilon) := \mathcal{S}_{k_\epsilon - 1} = \{k \in \mathcal{S} \mid k < k_\epsilon\}, \tag{12a}$$
$$\mathcal{U}(\epsilon) := \mathcal{U}_{k_\epsilon - 1} = \{k \in \mathbb{N} \mid k \notin \mathcal{S} \text{ and } k < k_\epsilon\}. \tag{12b}$$

The next theorem states a bound on the cardinality of $\mathcal{S}(\epsilon)$ when infinitely many successful iterations are generated. The first part of the theorem generalizes and strenghthens Leconte and Orban [9, Lemma 2]. The second part is new and provides a rigorous answer to Powell's intuition [16, §4].

Theorem 2. Let Problem Assumption 2.1 and Model Assumption 3.1 hold. Assume that Algorithm 1 generates infinitely many successful iterations. Let $\kappa_1 := (f(x_0) - f_{\text{low}})/(\eta_1 \kappa_{\text{mdc}} a_{\min}) > 0$, μ and p be as in Model Assumption 3.1, and $a_{\min}$ be as in Lemma 3. If $0 \leq p < 1$,

$$|\mathcal{S}(\epsilon)| \leq \left[(1-p)\,(1+2\mu)\,\kappa_1\,\epsilon^{-2} + 1\right]^{1/(1-p)} - 1 = O\left(\epsilon^{-2/(1-p)}\right). \tag{13}$$

If $p = 1$,

$$|\mathcal{S}(\epsilon)| \leq \exp\left((1+2\mu)\kappa_1\,\epsilon^{-2}\right) - 1. \tag{14}$$

Proof. Let $\ell \in \mathcal{S}(\epsilon)$. Lemma 4 and Model Assumption 3.1 imply

$$\begin{aligned} f(x_\ell) - f(x_\ell + s_\ell) &\geq \eta_1\,(m_\ell(0) - m_\ell(s_\ell)) \\ &\geq \eta_1 \kappa_{\text{mdc}} a_{\min} \frac{\min_{j=0,\ldots,\ell} \|\nabla f(x_j)\|^2}{1 + \max_{j=0,\ldots,\ell} \|B_j\|} \\ &\geq \eta_1 \kappa_{\text{mdc}} a_{\min} \epsilon^2 \frac{1}{1 + \mu(1 + |\mathcal{S}_\ell|^p)}. \end{aligned}$$

We sum the above inequality over all $\ell \in \mathcal{S}(\epsilon)$, use a telescoping argument, and obtain

$$f(x_0) - f_{\text{low}} \geq \eta_1 \kappa_{\text{mdc}} a_{\min} \epsilon^2 \sum_{\ell \in \mathcal{S}(\epsilon)} \frac{1}{1 + \mu(1 + |\mathcal{S}_\ell|^p)} = \eta_1 \kappa_{\text{mdc}} a_{\min} \epsilon^2 \sum_{k=0}^{|\mathcal{S}(\epsilon)|-1} \frac{1}{1 + \mu(1 + |\mathcal{S}_{\varphi(k)}|^p)},$$

where φ is an increasing map from $\{0, \ldots, |\mathcal{S}_\epsilon| - 1\}$ to $\mathcal{S}_\epsilon$. Hence, by definition of φ and $\mathcal{S}_{\varphi(k)}$,

$$\left|\mathcal{S}_{\varphi(k+1)}\right| = \left|\mathcal{S}_{\varphi(k)}\right| + 1 \quad \text{and} \quad \left|\mathcal{S}_{\varphi(0)}\right| = 1.$$

In other words, $\left|\mathcal{S}_{\varphi(k)}\right| = k + 1$, and

$$f(x_0) - f_{\text{low}} \geq \eta_1 \kappa_{\text{mdc}} a_{\min} \epsilon^2 \sum_{k=0}^{|\mathcal{S}(\epsilon)|-1} \frac{1}{1 + \mu(1 + (k+1)^p)}. \tag{15}$$

Lemma 5 with $k_1 = 0$ and $k_2 = |\mathcal{S}(\epsilon)| - 1$ gives

$$\sum_{k=0}^{|\mathcal{S}(\epsilon)|-1} \frac{1}{1 + \mu(1 + (k+1)^p)} \geq \frac{1}{(1+2\mu)} \int_1^{|\mathcal{S}(\epsilon)|+1} \frac{1}{t^p} \mathrm{d}t. \tag{16}$$

We combine (15) with (16) and obtain

$$f(x_0) - f_{\text{low}} \geq \frac{\eta_1 \kappa_{\text{mdc}} a_{\min} \epsilon^2}{(1+2\mu)} \int_1^{|\mathcal{S}(\epsilon)|+1} \frac{1}{t^p} \mathrm{d}t.$$

There are two cases:

- If $0 \leq p < 1$,
$$f(x_0) - f_{\text{low}} \geq \frac{\eta_1 \kappa_{\text{mdc}} a_{\min} \epsilon^2}{(1+2\mu)} \frac{(|\mathcal{S}(\epsilon)|+1)^{1-p} - 1}{1-p},$$
which provides (13).
- If $p = 1$,
$$f(x_0) - f_{\text{low}} \geq \eta_1 \kappa_{\text{mdc}} a_{\min} \epsilon^2 \left[\frac{1}{(1+2\mu)} \log \left(\frac{|\mathcal{S}(\epsilon)|+1}{1} \right) \right],$$
which establishes (14). □

When $0 \leq p < 1$ in Model Assumption 3.1, Theorem 2 improves the constant in the complexity bound of [9, Lemma 2]. In particular, the bound given in [9] diverges as $p \to 1$, whereas (13) suggests that the complexity becomes exponential, which (14) confirms.

Note that, although the performance of Algorithm 1 might depend on α and β, if the parameter γ_1 is selected such that $a_0 \geq \gamma_1$, then $a_{\min} \neq a_0$ in Lemma 3, and (13)–(14) no longer depend on α and β. The number of unsuccessful iterations, on the other hand, always depends on α and β.

Theorem 3. Let Problem Assumption 2.1 and Model Assumption 3.1 hold. Assume that Algorithm 1 generates infinitely many successful iterations. Then

$$|\mathcal{U}(\epsilon)| \leq |\log_{\gamma_2}(\gamma_4)||\mathcal{S}(\epsilon)| + (1-\alpha)\log_{\gamma_2}(\epsilon) + (\beta-1)\log_{\gamma_2}(1+\mu(1+|\mathcal{S}(\epsilon)|^p)) + \log_{\gamma_2}\left(\frac{a_{\min}}{\Delta_0}\right), \quad (17)$$

where μ and p are defined in Model Assumption 3.1 and $|\mathcal{S}(\epsilon)|$ is as in Theorem 2.

Proof. As in [1, Lemma 2.3.1], the mechanism of Algorithm 1 guarantees that for all $k \in \mathbb{N}$,

$$|\mathcal{U}_k| \leq |\log_{\gamma_2}(\gamma_4)|\,|\mathcal{S}_k| + \log_{\gamma_2}\left(\frac{\Delta_k}{\Delta_0}\right).$$

Hence, Lemma 3 yields

$$\begin{aligned}|\mathcal{U}(\epsilon)| &\leq |\log_{\gamma_2}(\gamma_4)||\mathcal{S}(\epsilon)| + \log_{\gamma_2}\left(\frac{(\min_{j=0,\ldots,k_\epsilon-1}\|\nabla f(x_j)\|)^{1-\alpha}(1+\max_{j=0,\ldots,k_\epsilon-1}\|B_j\|)^{\beta-1}a_{\min}}{\Delta_0}\right)\\ &\leq |\log_{\gamma_2}(\gamma_4)||\mathcal{S}(\epsilon)| + (1-\alpha)\log_{\gamma_2}(\epsilon) + (\beta-1)\log_{\gamma_2}(1+\mu(1+|\mathcal{S}(\epsilon)|^p)) + \log_{\gamma_2}\left(\frac{a_{\min}}{\Delta_0}\right). \quad \square\end{aligned}$$

The bound (17) is minimized in (α, β) for $\alpha = \beta = 1$, (17) as the logarithmic terms cancel out. The bound reduces to

$$|\mathcal{U}(\epsilon)| \leq |\log_{\gamma_2}(\gamma_4)||\mathcal{S}(\epsilon)| + \log_{\gamma_2}\left(\frac{a_{\min}}{\Delta_0}\right). \quad (18)$$

Thus, the scaling $\|\nabla f(x_k)\|/(1+\|B_k\|)$ of the trust-region radius improves the worst-case complexity bound. We now compare our most favorable bound with that of [1, Theorem 2.3.7].

Corollary 1. *Let $\alpha = \beta = 1$ in (2). Under the assumptions of Theorem 2, if $0 \leq p < 1$,*

$$\begin{aligned}k_\epsilon &\leq \left(|\log_{\gamma_2}(\gamma_4)|+1\right)\left(\left[(1-p)\,(1+2\mu)\,\kappa_1\,\epsilon^{-2}+1\right]^{1/(1-p)} - 1\right) + \log_{\gamma_2}\left(\frac{a_{\min}}{\Delta_0}\right) \quad (19)\\ &= O\left(\epsilon^{-2/(1-p)}\right)\end{aligned}$$

If $p = 1$,

$$\begin{aligned}k_\epsilon &\leq \left(|\log_{\gamma_2}(\gamma_4)|+1\right)\left(\exp\left[(1+2\mu)\kappa_1\,\epsilon^{-2}\right] - 1\right) + \log_{\gamma_2}\left(\frac{a_{\min}}{\Delta_0}\right). \quad (20)\\ &= O\left(\exp\left[(1+2\mu)\kappa_1\,\epsilon^{-2}\right]\right)\end{aligned}$$

When $p = 0$, i.e., the model Hessians are uniformly bounded, (19) reduces to

$$k_\epsilon \leq \left(|\log_{\gamma_2}(\gamma_4)| + 1\right) \frac{4(L + 2\mu L)}{\gamma_1 \eta_1 (1 - \eta_2)} (f(x_0) - f_{\text{low}})\, \epsilon^{-2} + \log_{\gamma_2}\left(\frac{\gamma_1(1 - \eta_2)}{2L\Delta_0}\right), \tag{21}$$

where, for simplicity, we set $\kappa_{\text{mdc}} = \frac{1}{2}$ and consider that $L \geq 1$.

In the classical trust-region method [2], [1, Theorem 2.3.7] reduces to

$$k_\epsilon \leq \left(|\log_{\gamma_2}(\gamma_4)| + 1\right) \frac{4(L + 2\mu)}{\gamma_1 \eta_1 (1 - \eta_2)} (f(x_0) - f_{\text{low}})\, \epsilon^{-2} + \log_{\gamma_2}(\epsilon) + \log_{\gamma_2}\left(\frac{\gamma_1(1 - \eta_2)}{2(L + 2\mu)\, \Delta_0}\right), \tag{22}$$

where we assumed that $\|\nabla f(x_0)\| \leq \Delta_0$ to simplify. It is now apparent that, although the $\log_{\gamma_2}(\epsilon)$ of (22) does not appear in (21), the constant in front of the leading term ϵ^{-2} is worse in (21). Nevertheless, the added generality of our analysis allows us to treat the case $p > 0$.

Note that it is allowed to set $\gamma_4 = 1$, and in this case, (18) implies that the number of unsuccessful iterations is uniformly bounded independently of ϵ. Note also that in our setting, neither the classical radius $\alpha = \beta = 0$ nor the setting $\alpha = 1$, $\beta = 0$ enjoy the most favorable complexity bound. For a better analysis on the impact of selecting α and β, a numerical comparison of different choices of such parameters will be conducted in Section 5.

3.1 Sharpness of the complexity bound

Let $\epsilon \in (0, 1)$ and $c > 0$. Our goal is to construct smooth $f : \mathbb{R} \to \mathbb{R}$ that satisfies Problem Assumption 2.1 and Model Assumption 3.1 and for which Algorithm 1 requires exactly

$$k_\epsilon = \begin{cases} \lfloor \epsilon^{-2/(1-p)} \rfloor, & \text{if } 0 \leq p < 1 \\ \left\lfloor \exp(c\epsilon^{-2}) \right\rfloor, & \text{if } p = 1 \end{cases}$$

function and gradient evaluations to produce x_{k_ϵ} with $|f'(x_{k_\epsilon})| \leq \epsilon$. The construction follows the guidelines of Cartis et al. [1] and proceeds as Leconte and Orban [9]. We begin by recalling a key result on Hermite interpolation.

Proposition 1 (1, Theorem *A*.9.2; 9, Proposition 6)**.** *Let $k_\epsilon \in \mathbb{N}$. Consider real sequences $\{f_k\}$, $\{g_k\}$, and $\{x_k\}$ for $k \in \{0, \ldots, k_\epsilon\}$. For $k = 0, \ldots, k_\epsilon$, let $s_k := x_{k+1} - x_k$, and assume that*

$$|f_{k+1} - (f_k + g_k s_k)| \leq \kappa_f s_k^2,$$
$$|g_{k+1} - g_k| \leq \kappa_f |s_k|,$$

for $k = 0, \ldots, k_\epsilon - 1$, where $\kappa_f \geq 0$. Then, there exists continuously differentiable $f : \mathbb{R} \to \mathbb{R}$ such that

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

for $k = 0, \ldots, k_\epsilon$. Furthermore, if

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

for $k = 0, \ldots, k_\epsilon$, then both $|f|$ and $|f'|$ are bounded by a constant depending only on κ_f.

We proceed as Leconte and Orban [9], but consider a scenario where Model Assumption 3.1 holds for any $p \in [0, 1]$. For $k \in \{0, \ldots, k_\epsilon\}$, we set

$$\omega_k := \frac{k_\epsilon - k}{k_\epsilon} \quad \text{and} \quad g_k := -\epsilon(1 + \omega_k). \tag{23}$$

By definition, $|g_k| > \epsilon$ for all $k \in \{0, \ldots, k_\epsilon - 1\}$ and $|g_{k_\epsilon}| = \epsilon$.

Let $p \in [0,1]$. Define

$$B_0 := 1 \quad \text{and} \quad B_k := k^p \text{ for all } k = 1, \ldots, k_\epsilon, \tag{24}$$

and

$$x_0 := 0 \quad \text{and} \quad x_{k+1} := x_k + s_k \text{ for all } k = 0, \ldots, k_\epsilon - 1, \tag{25}$$

where

$$s_k := -B_k^{-1} g_k > 0 \text{ for all } k = 0, \ldots, k_\epsilon - 1. \tag{26}$$

Finally, set

$$f_0 := \begin{cases} 8\epsilon^2 + 4c, & \text{if } p = 1 \\ 8\epsilon^2 + \frac{4}{1-p}, & \text{if } 0 \leq p < 1 \end{cases} \quad \text{and} \quad f_{k+1} := f_k + g_k s_k \text{ for all } k = 0, \ldots, k_\epsilon - 1. \tag{27}$$

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

Lemma 6. *The sequence $\{f_k\}$ defined in (27) is decreasing and*

$$f_k \in [0,\ f_0] \quad \textit{for all } k = 0, \ldots, k_\epsilon. \tag{28}$$

Proof. Because $f_{k+1} - f_k = g_k s_k < 0$, $\{f_k\}$ is decreasing.

For $k \in \{0, 1\}$, $f_0 - f_1 = -g_0 s_0 = 4\epsilon^2 > 0$, so that (28) holds. For $k = 2, \ldots, k_\epsilon$,

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

Because $1 + \omega_i \leq 2$,

$$\sum_{i=1}^{k-1} (1+\omega_i)^2 i^{-p} \leq \sum_{i=1}^{k-1} 4 i^{-p} = 4\left(1 + \sum_{i=2}^{k-1} i^{-p}\right) \leq 4\left(1 + \sum_{i=2}^{k-1} \int_{i-1}^{i} t^{-p}\,\mathrm{d}t\right) = 4\left(1 + \int_1^{k-1} t^{-p}\,\mathrm{d}t\right),$$

so that

$$f_0 - f_k \leq 4\epsilon^2 \left(2 + \int_1^{k-1} t^{-p}\,\mathrm{d}t\right).$$

If $p = 1$,

$$f_0 - f_k \leq 4\epsilon^2 \left(2 + \log(k-1)\right) \leq 4\epsilon^2 \left(2 + \log(k)\right) \leq 4\epsilon^2 \left(2 + \log(k_\epsilon)\right) \leq 4\epsilon^2 \left(2 + c\epsilon^{-2}\right).$$

Thus, $f_0 - f_k \leq f_0$, which implies $f_k \geq 0$. Similarly, if $0 \leq p < 1$,

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

Again, $f_0 - f_k \leq f_0$, and $f_k \geq 0$. □

The next theorem establishes slow convergence of Algorithm 1.

Theorem 4. Let $0 < \epsilon < 1$ and $c \geq 0$. Algorithm 1 applied to minimize $f : \mathbb{R}^n \to \mathbb{R}$ satisfying Problem Assumption 2.1 and Model Assumption 3.1 with $0 \leq p \leq 1$ may require as many as

$$k_\epsilon := \begin{cases} \lfloor \epsilon^{-2/(1-p)} \rfloor & \text{if } 0 \leq p < 1 \\ \lfloor \exp(c\epsilon^{-2}) \rfloor & \text{if } p = 1 \end{cases}$$

iterations to produce x_{k_ϵ} such that $\|\nabla f(x_{k_\epsilon})\| \leq \epsilon$.

Proof. We check that the iterates (25), the function values (27), the gradient values (23), and the step sizes (26) satisfy the assumptions of Proposition 1. By construction, for $k = 0, \ldots, k_\epsilon$, $|f_{k+1} - f_k - g_k s_k| = 0$, and

$$|g_{k+1} - g_k| = \epsilon(\omega_k - \omega_{k+1}) = \epsilon k_\epsilon^{-1} \leq \epsilon k^{-1} \leq |g_k| k^{-1} = B_k^{-1}|g_k| = |s_k|.$$

Lemma 6 yields $f_k \in [0,\, f_0]$ for $k = 0, \ldots, k_\epsilon$, and (23) implies

$$|g_k| \leq 2\epsilon \leq 2 \quad \text{and} \quad 0 < s_k \leq |g_k| \leq 2, \quad k = 0, \ldots, k_\epsilon.$$

The bounds above pave the way for Proposition 1, with $\kappa_f := \max(f_0, 2)$.

The final step is to check coherence with Algorithm 1, and ensure that only successful iterations occur so that x_{k+1} is always updated and $|s_k| \leq \frac{|g_k|^\alpha}{(1+B_k)^\beta}\Delta_k$. Each iteration $k = 0, \ldots, k_\epsilon - 1$ is successful because

$$\rho_k = \frac{f_k - f_{k+1}}{f_k - f_k - g_k s_k - \frac{1}{2}B_k s_k^2} = \frac{-g_k s_k}{\frac{1}{2}B_k^{-1} g_k^2} = \frac{B_k^{-1} g_k^2}{\frac{1}{2}B_k^{-1} g_k^2} = 2.$$

Thus, $x_{k+1} = x_k + s_k$. There remains to show that each s_k is inside the trust region. Consider first $k = 0$. Set $\Delta_0 := 2^{2-\alpha}$. By construction, $(1 + B_0)^\beta = 2^\beta \leq 2$, because $\beta \leq 1$. Moreover, using the fact that $\alpha \leq 1$,

$$|s_0| = |g_0| = |g_0|^\alpha |g_0|^{1-\alpha} = |g_0|^\alpha (2\epsilon)^{1-\alpha} \leq \frac{2(2^{1-\alpha})|g_0|^\alpha}{(1 + B_0)^\beta} = \frac{|g_0|^\alpha}{(1 + B_0)^\beta}\Delta_0.$$

Similarly, for $k \geq 1$, the fact that $B_k \geq 1$ implies

$$|s_k| = |B_k^{-1} g_k| = \frac{|g_k|}{B_k} \leq \frac{2|g_k|^\alpha |g_k|^{1-\alpha}}{1 + B_k} \leq \frac{2(2)^{1-\alpha}|g_k|^\alpha}{(1 + B_k)^\beta} = \frac{|g_k|^\alpha}{(1 + B_k)^\beta}\Delta_0 \leq \frac{|g_k|^\alpha}{(1 + B_k)^\beta}\Delta_k.$$

Thus, each s_k is inside the trust-region.

We deduce from Proposition 1 that there exists continuously differentiable $f : \mathbb{R} \to \mathbb{R}$ with Lipschitz-continuous gradient such that, for $k = 0, \ldots, k_\epsilon$, $f(x_k) = f_k$ and $f'(x_k) = g_k$. As the range of f is within $[-\kappa_f,\, \kappa_f]$, f is bounded below. This concludes the proof. □

Figure 1 shows plots of f and f' for $p = 0$, 0.5 and 1.

4 Complexity when the bound on $\|B_k\|$ depends only on the current iteration

We now assume that $\{B_k\}$ is bounded by a multiple of k. Among other scenarios, this suggests that in Algorithm 1, one is allowed to update the model Hessian at both successful and unsuccessful iterations.

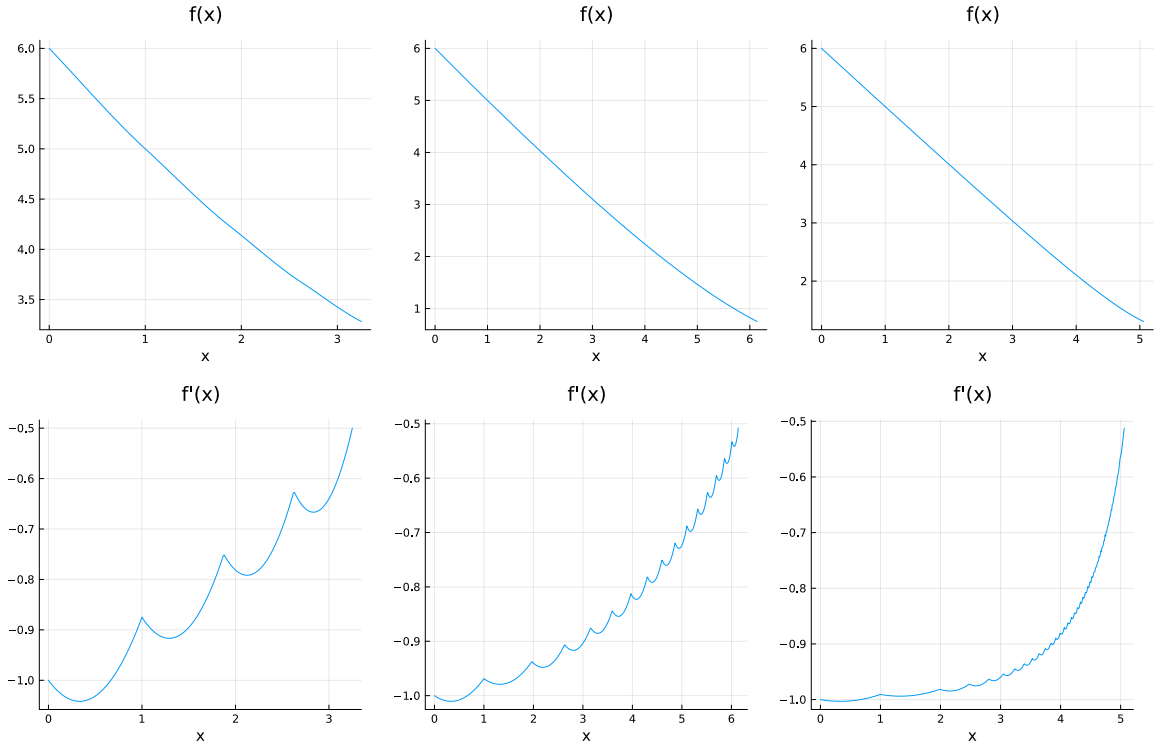
We replace Model Assumption 3.1 with the following assumption.

Model Assumption 4.1. *There exists $\mu > 0$ and $0 \leq p \leq 1$ such that for all $k \geq 0$,*

$$\max_{j=0,\ldots,k} \|B_j\| \leq \mu(1 + k^p).$$

As before, we begin with the case where only a finite number of successful iterations is generated.

Theorem 5. Let Problem Assumption 2.1 and Model Assumption 4.1 be satisfied. If Algorithm 1 only generates finitely many successful iterations, then $x_k = x^*$ for all sufficiently large k where $\|\nabla f(x^*)\| = 0$.



(a) $p = 0$. **(b)** $p = 0.5$. **(c)** $p = 1$ **and** $c = 1$.

Figure 1: Form of f (top) and f' (bottom) over $[x_0, x_{k_\epsilon}]$ using $\epsilon = 0.5$ for three values of p.

Proof. Assume by contradiction that there exists $\nu > 0$ such that $\|\nabla f(x_k)\| \geq \nu > 0$ for all $k \in \mathbb{N}$, and let k_f be the last successful iteration. Necessarily, $x_k = x_{k_f}$ for all $k \geq k_f$, and hence, $\{x_k\} \to x^* := x_{k_f}$. By Lemma 3, for all k,

$$\Delta_k \geq \frac{(\min_{j=0,\ldots,k} \|\nabla f(x_j)\|)^{1-\alpha}}{(1 + \max_{j=0,\ldots,k} \|B_j\|)^{1-\beta}} a_{\min} \geq \frac{\nu^{1-\alpha}}{(1 + \mu(1+k^p))^{1-\beta}} a_{\min}.$$

On unsuccessful iterations, Algorithm 1 reduces Δ_k by a factor at least γ_2. Hence, for all $k \geq k_f$,

$$\Delta_k \leq \gamma_2^{k-k_f} \Delta_{k_f}.$$

We combine the above inequalities, and obtain

$$\frac{\nu^{1-\alpha}}{(1 + \mu(1+k^p))^{1-\beta}} a_{\min} \leq \gamma_2^{k-k_f} \Delta_{k_f},$$

which may be rewritten

$$0 < \frac{\nu^{1-\alpha} \gamma_2^{k_f}}{\Delta_{k_f}} a_{\min} \leq \gamma_2^k (1 + \mu(1+k^p))^{1-\beta}.$$

However, the above is a contradiction as the right-hand side goes to zero as $k \to \infty$. Thus, by contradiction, $\liminf_{k\to\infty} \|\nabla f(x_k)\| = \|\nabla f(x_{k_f})\| = \|\nabla f(x^*)\| = 0$. □

Let $\tau \in \mathbb{N}_0$ and k_0 be the index of the first successful iteration. Define

$$\mathcal{T}_k^\tau := \{j = k_0, \ldots, k \mid j \leq \tau |\mathcal{S}_j|\}, \tag{29a}$$

$$\mathcal{W}_k^\tau := \{j = k_0, \ldots, k \mid j > \tau|\mathcal{S}_j|\}. \tag{29b}$$

Note that the set $\mathcal{T}_k^\tau$ might be empty. The next technical result will allow us to relate the number of successful iterations to the number of iterations.

Lemma 7. *Let $\{r_j\}_{j\in\mathbb{N}}$ be a non-decreasing positive real sequence. For any $k \geq k_0$,*

$$\tau \sum_{j\in\mathcal{S}_k} \frac{1}{r_j} \geq \sum_{j\in\mathcal{T}_k^\tau} \frac{1}{r_j} = \sum_{j=k_0}^{k} \frac{1}{r_j} - \sum_{j\in\mathcal{W}_k^\tau} \frac{1}{r_j},$$

where $\mathcal{T}_k^\tau$ and $\mathcal{W}_k^\tau$ are defined in (29).

Proof. If $\mathcal{T}_k^\tau$ is empty, the result is trivial, using the convention that the sum over an empty set is zero.

Otherwise, let $k \geq k_0$ and $j \in \mathcal{T}_k^\tau$, i.e., $k_0 \leq j \leq k$ and $j \leq \tau|\mathcal{S}_j|$. Define $\mathcal{S}_j^\tau$ as the set of elements of $\mathcal{S}_j$, in ascending order, where each element is repeated τ times:

$$\mathcal{S}_j^\tau := \Big\{ \underbrace{i_1, i_1, \ldots, i_1}_{\tau \text{ times}}, \underbrace{i_2, i_2, \ldots, i_2}_{\tau \text{ times}}, \ldots, \underbrace{i_{|\mathcal{S}_j|}, i_{|\mathcal{S}_j|}, \ldots, i_{|\mathcal{S}_j|}}_{\tau \text{ times}} \;\Big|\; i_l \in \mathcal{S}_j \text{ for } l \in \{1, \ldots, |\mathcal{S}_j|\} \Big\}. \tag{30}$$

By construction, $|\mathcal{S}_j^\tau| = \tau|\mathcal{S}_j| \geq j$, hence $\mathcal{S}_j^\tau$ must contain at least j elements. Let i_j^τ be the j-th element of $\mathcal{S}_j^\tau$. In particular, $i_j^\tau \in \mathcal{S}_j$, and because each element of $\mathcal{S}_j$ is less than j, $i_j^\tau \leq j$. Because $\mathcal{S}_j^\tau \subseteq \mathcal{S}_k^\tau$ also holds, $i_j^\tau \in \mathcal{S}_k^\tau$. We have just showed that

$$\{i_j^\tau \mid j \in \mathcal{T}_k^\tau\} \subseteq \mathcal{S}_k^\tau. \tag{31}$$

As $\{r_j\}$ is non-decreasing and $i_j^\tau \leq j$, $r_{i_j^\tau} \leq r_j$. We sum over $j \in \mathcal{T}_k^\tau$, and obtain

$$\sum_{j\in\mathcal{T}_k^\tau} \frac{1}{r_{i_j^\tau}} \geq \sum_{j\in\mathcal{T}_k^\tau} \frac{1}{r_j}. \tag{32}$$

Positivity of $\{r_j\}_{j\in\mathbb{N}}$ and (31) then yield

$$\sum_{j\in\mathcal{S}_k^\tau} \frac{1}{r_j} \geq \sum_{j\in\mathcal{T}_k^\tau} \frac{1}{r_{i_j^\tau}}. \tag{33}$$

Thus, we combine (29), (30), (32) and (33) to conclude that

$$\tau \sum_{j\in\mathcal{S}_k} \frac{1}{r_j} = \sum_{j\in\mathcal{S}_k^\tau} \frac{1}{r_j} \geq \sum_{j\in\mathcal{T}_k^\tau} \frac{1}{r_{i_j^\tau}} \geq \sum_{j\in\mathcal{T}_k^\tau} \frac{1}{r_j} = \sum_{j=k_0}^{k} \frac{1}{r_j} - \sum_{j\in\mathcal{W}_k^\tau} \frac{1}{r_j}. \qquad \square$$

The following lemma also plays a key role in deriving our worst-case complexity bound.

Lemma 8. *Let Problem Assumption 2.1 and Model Assumption 4.1 be satisfied. Assume that $\tau \in \mathbb{N}_0$ is chosen so that $\gamma_4\gamma_2^{\tau-1} < 1$. Let $\epsilon > 0$ and k_ϵ be the first iteration such that $\|\nabla f(x_{k_\epsilon})\| \leq \epsilon$. Then,*

$$\sum_{k\in\mathcal{W}_{k_\epsilon-1}^\tau} \frac{1}{(1+\mu(1+k^p))} \leq \frac{\Delta_0\epsilon^{\alpha-1}}{a_{\min}}\xi_\beta, \qquad \xi_\beta := \sum_{k\in\mathbb{N}} \frac{\left(\gamma_4\gamma_2^{\tau-1}\right)^{k/\tau}}{(1+\mu(1+k^p))^\beta} < \infty, \tag{34}$$

where $a_{\min}$ is as in Lemma 3 and $\mathcal{W}_{k_\epsilon-1}^\tau$ is defined in (29b).

Proof. Let $k \in \mathcal{W}^\tau_{k_\epsilon-1}$ and $r_k := 1+\mu(1+k^p)$. Lemma 3, the update mechanism of Δ_k in Algorithm 1 and Model Assumption 3.1 together yield

$$\frac{a_{\min}\epsilon^{1-\alpha}}{r_k} = \frac{a_{\min}\epsilon^{1-\alpha}}{r_k^{1-\beta}\, r_k^{\beta}} \leq \frac{a_{\min}\min_{j=0,\ldots,k}\|\nabla f(x_j)\|^{1-\alpha}}{\left(1+\max_{j=0,\ldots,k}\|B_j\|\right)^{1-\beta} r_k^{\beta}} \leq \frac{\Delta_k}{r_k^{\beta}} \leq \frac{\gamma_4^{|S_k|}\gamma_2^{|U_k|}}{r_k^{\beta}}\Delta_0 = \frac{\gamma_4^{|S_k|}\gamma_2^{k-|S_k|}}{r_k^{\beta}}\Delta_0.$$

Because $k \in \mathcal{W}^\tau_{k_\epsilon-1}$, $k > \tau|\mathcal{S}_k|$, which, together with the fact that $\gamma_4 > 1$ and $0 < \gamma_2 < 1$ leads to

$$\frac{\gamma_4^{|S_k|}\gamma_2^{k-|S_k|}}{r_k^{\beta}}\Delta_0 < \frac{\gamma_4^{k/\tau}\gamma_2^{k-k/\tau}}{r_k^{\beta}}\Delta_0 = \frac{\left(\gamma_4\gamma_2^{\tau-1}\right)^{k/\tau}}{r_k^{\beta}}\Delta_0.$$

We sum over $k \in \mathcal{W}^\tau_{k_\epsilon-1}$, and obtain

$$a_{\min}\epsilon^{1-\alpha} \sum_{k\in\mathcal{W}^\tau_{k_\epsilon-1}} \frac{1}{r_k} \leq \Delta_0 \sum_{k\in\mathbb{N}} \frac{\left(\gamma_4\gamma_2^{\tau-1}\right)^{k/\tau}}{r_k^{\beta}}, \tag{35}$$

which is (34). The series on the right-hand side of (35) is convergent. Indeed, let u_k denote its general term. Because $\{r_k\}$ is non-decreasing,

$$\lim_{k\to+\infty}\frac{u_{k+1}}{u_k} = \lim_{k\to+\infty}\left(\frac{\left(\gamma_4\gamma_2^{\tau-1}\right)^{(k+1)/\tau}}{r_{k+1}^{\beta}}\right)\left(\frac{r_k^{\beta}}{\left(\gamma_4\gamma_2^{\tau-1}\right)^{k/\tau}}\right) \leq (\gamma_4\gamma_2^{\tau-1})^{1/\tau} < 1. \qquad \square$$

Note that it is possible to choose τ as required by Lemma 8, as it suffices to pick $\tau < \log_{\gamma_2}(\gamma_4^{-1})+1$, where the right-hand side is larger than 1.

We are ready to state our main result on the evaluation complexity of Algorithm 1 under Model Assumption 4.1.

Theorem 6. Let Problem Assumption 2.1 and Model Assumption 4.1 be satisfied. Assume that Algorithm 1 generates infinitely many successful iterations. Let k_0 be the index of the first successful iteration, μ and p be as in Model Assumption 3.1, and τ and ξ_β be as in Lemma 8. Let $\epsilon > 0$, and k_ϵ be the first iteration such that $\|\nabla f(x_{k_\epsilon})\| \leq \epsilon$. Define

$$\kappa_2 := \frac{\tau(f(x_0)-f_{\text{low}})}{\eta_1\kappa_{\text{mdc}}a_{\min}} > 0, \qquad \kappa_3 := \frac{\Delta_0\xi_\beta}{a_{\min}} > 0.$$

If $0 \leq p < 1$,

$$\begin{aligned} k_\epsilon &\leq \left[(1-p)\frac{(1+\mu(1+(1+k_0)^p))}{(1+k_0)^p}\left(\kappa_2\,\epsilon^{-2}+\kappa_3\,\epsilon^{\alpha-1}\right)+(k_0+1)^{1-p}\right]^{1/(1-p)} - 1 \\ &= O\left(\epsilon^{-2/(1-p)}+\epsilon^{(1-\alpha)/(1-p)}\right). \end{aligned} \tag{36}$$

If $p = 1$,

$$k_\epsilon \leq (k_0+1)\exp\left[\frac{(1+\mu(2+k_0))}{1+k_0}\left(\kappa_2\,\epsilon^{-2}+\kappa_3\,\epsilon^{\alpha-1}\right)\right] - 1. \tag{37}$$

Proof. Let $k \in \mathcal{S}(\epsilon)$. Lemma 4 and Model Assumption 4.1 imply

$$f(x_k) - f(x_k+s_k) \geq \eta_1(m_k(0)-m_k(s_k))$$

$$\geq \eta_1 \kappa_{\text{mdc}} a_{\min} \frac{\min_{j=0,\ldots,k} \|\nabla f(x_j)\|^2}{1 + \max_{j=0,\ldots,k} \|B_j\|}$$
$$\geq \eta_1 \kappa_{\text{mdc}} a_{\min} \epsilon^2 \frac{1}{1 + \mu(1 + k^p)}.$$

We sum the above inequality over all $k \in \mathcal{S}(\epsilon)$, use a telescoping argument, and obtain

$$f(x_0) - f_{\text{low}} \geq \eta_1 \kappa_{\text{mdc}} a_{\min} \epsilon^2 \sum_{k \in \mathcal{S}(\epsilon)} \frac{1}{1 + \mu(1 + k^p)}.$$

Let $r_k \coloneqq 1 + \mu(1 + k^p)$. The sequence $\{r_k\}$ is positive and non-decreasing, so Lemmas 5, 7 and 8 yield

$$\begin{aligned} f(x_0) - f_{\text{low}} &\geq \frac{\eta_1 \kappa_{\text{mdc}} a_{\min} \epsilon^2}{\tau} \left[\sum_{k=k_0}^{k_\epsilon - 1} \frac{1}{r_k} - \sum_{k \in \mathcal{W}^\tau_{k_\epsilon - 1}} \frac{1}{r_k} \right] \\ &\geq \frac{\eta_1 \kappa_{\text{mdc}} a_{\min} \epsilon^2}{\tau} \left[\sum_{k=k_0}^{k_\epsilon - 1} \frac{1}{r_k} - \frac{\Delta_0 \xi_\beta \epsilon^{\alpha-1}}{a_{\min}} \right] \\ &\geq \frac{\eta_1 \kappa_{\text{mdc}} a_{\min} \epsilon^2}{\tau} \left[\frac{(1 + k_0)^p}{r_{k_0}} \int_{k_0+1}^{k_\epsilon + 1} \frac{1}{t^p} \, \mathrm{d}t - \frac{\Delta_0 \xi_\beta \epsilon^{\alpha-1}}{a_{\min}} \right], \end{aligned} \tag{38}$$

where we applied Lemma 5 applied with $k_1 = k_0$ and $k_2 = k_\epsilon - 1$. We distinguish two cases.

- If $0 \leq p < 1$, (38) becomes

$$f(x_0) - f_{\text{low}} \geq \frac{\eta_1 \kappa_{\text{mdc}} a_{\min} \epsilon^2}{\tau} \left[\frac{(1 + k_0)^p}{(1 + \mu(1 + (1 + k_0)^p))} \frac{(k_\epsilon + 1)^{1-p} - (k_0 + 1)^{1-p}}{1 - p} - \frac{\Delta_0 \xi_\beta \epsilon^{\alpha-1}}{a_{\min}} \right],$$

  which is (36).
- If $p = 1$, (38) becomes

$$f(x_0) - f_{\text{low}} \geq \frac{\eta_1 \kappa_{\text{mdc}} a_{\min} \epsilon^2}{\tau} \left[\frac{(1 + k_0)}{(1 + \mu(1 + k_0))} \log\left(\frac{k_\epsilon + 1}{k_0 + 1} \right) - \frac{\Delta_0 \xi_\beta \epsilon^{\alpha-1}}{a_{\min}} \right],$$

  which gives (37). □

The complexity bounds of Theorem 6 have the same nature as those of Theorems 2 and 3; a polynomial bound for $0 \leq p < 1$ and an exponential bound for $p = 1$. In Theorem 6, we bound the total number of iterations to reach an ϵ-first order point; we do no have estimates on the number of successful iterations as we did under Model Assumption 3.1.

In addition, Theorem 6 suggests that when relax Model Assumption 3.1 to Model Assumption 4.1, the complexity bound deteriorates.

Finally, under Model Assumption 4.1, the value of α appears in the complexity bound. In fact, if $\alpha < -1$, the leading term in the complexity bound becomes $O(\epsilon^{(\alpha-1)/(1-p)})$ when $0 < p < 1$, which is worse than $O(\epsilon^{-2/(1-p)})$. Similarly, when $p = 1$, the leading term deteriorates to $O(\exp(\frac{(1+\mu(2+k_0))}{1+k_0} \kappa_3 \epsilon^{1-\alpha}))$. The example of Section 3.1 satisfies Model Assumption 4.1 because all iterations are successful by construction. Neither Lemma 6 nor Theorem 4 depend on the value of α. Thus, whenever $-1 \leq \alpha \leq 1$, the same example illustrates that the bounds of Theorem 6 are sharp. However, when $\alpha < -1$, either a different example is required, or the bound is not sharp. Until we answer this last question, we state the following corollary of Theorem 6, which applies, to most known variants, including the standard trust-region radius update ($\alpha = \beta = 0$) and choice of radius ($\alpha = 1$, $\beta = 0$) made by Curtis et al. [3].

Corollary 2. *Let $-1 \leq \alpha \leq 1$ in (2). Under the assumptions of Theorem 6.*

- *if $0 \leq p < 1$, (36) continues to hold, but is $O\left(\epsilon^{-2/(1-p)}\right)$;*
- *if $p = 1$, (37) continues to hold, but is $O\left(\exp(\kappa_4 \epsilon^{-2})\right)$ with $\kappa_4 := \frac{(1+\mu(2+k_0))}{1+k_0}\kappa_2$,*

and those bounds are sharp up to a constant factor.

5 Numerical illustration

In order to better understand the connection between worst-case complexity bounds and performance in practice, we implemented Algorithm 1 in the Julia programming language using the *JuliaSmoothOptimizers* infrastructure [10]. We altered the implementation of the *trunk* solver [11] to implement the trust-region radius management of Algorithm 1. By default, *trunk* is a standard trust-region method for (1) in which steps are computed using the truncated conjugate-gradient method [17]. No other algorithmic changes were performed. We extracted all unconstrained problems from the *OptimizationProblems* collection [12] with at least two variables, which resulted in 150 problems with dimension ranging from 2 to 100. Larger instances are available by changing the default number of variables in many problems, but our intention here is to obtain preliminary results that may confirm or infirm causality between worst-case complexity and performance. Typical complexity studies in nonconvex deterministic optimization are, unfortunately, almost never accompanied by numerical experiments.

Figure 2 reports our results in the form of Dolan and Moré [5] performance profiles in terms of total number of evaluations of f (left), ∇f (center), and CPU time (right). At each iteration, B_k is the exact Hessian computed via automatic differentiation. The variants tested are $\alpha = \beta = 0$, named `trunk_0_0`, $\alpha = 0$ and $\beta = 1$, named `trunk_0_1`, $\alpha = 1$ and $\beta = 0$, named `trunk_1_0`, and $\alpha = \beta = 1$, named `trunk_1_1`.

Not all variants are equally efficient or robust. The standard implementation $\alpha = \beta = 0$ is both the fastest and most robust with failures on 7 problems. The method using the radius proposed by Curtis et al. [3], $\alpha = 1$ and $\beta = 0$, is a close second in terms of efficiency, but fails on 8 problems. The two other variants are visibly less efficient. The variant `trunk_1_1` ranks third in terms of number of evaluations, but last in terms of CPU time (though the gap with the next method is small), and fails on 7 problems. Finally, `trunk_0_1` ranks last in terms of number of evaluations, and third in terms of CPU time, and fails on 11 problems. In terms of CPU time, the two variants that use $\beta = 0$ are faster because they do not need to compute $\|B_k\|$ at each iteration.

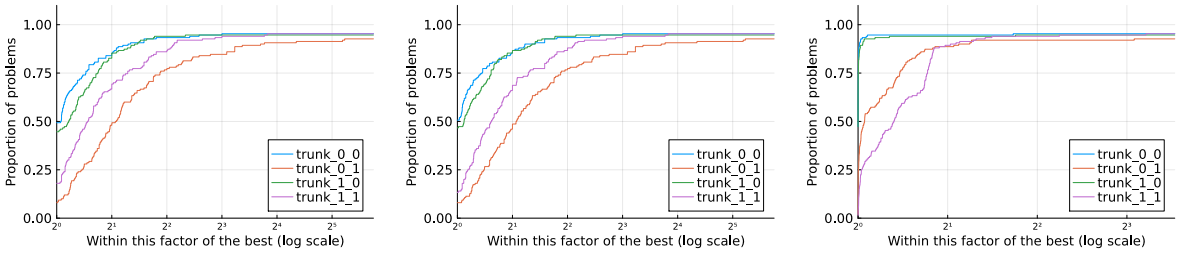


Figure 2: Performance profiles of trunk with exact Hessian in terms of number of objective evaluations (left), gradient evaluations (center), and CPU time (right). Method `trunk_a_b` refers to the variant $\alpha = a$ and $\beta = b$.

Because our analysis allows for Hessian approximations and it has long been suspected that quasi-Newton approximations may diverge as fast as linearly with $|\mathcal{S}_k|$, we ran the same experiment as above but with model Hessians defined as limited-memory BFGS or SR1 approximations with memory 5. The results are reported in Figures 3 and 4. The variants are color-coded as in Figure 2, and the trend is the same.

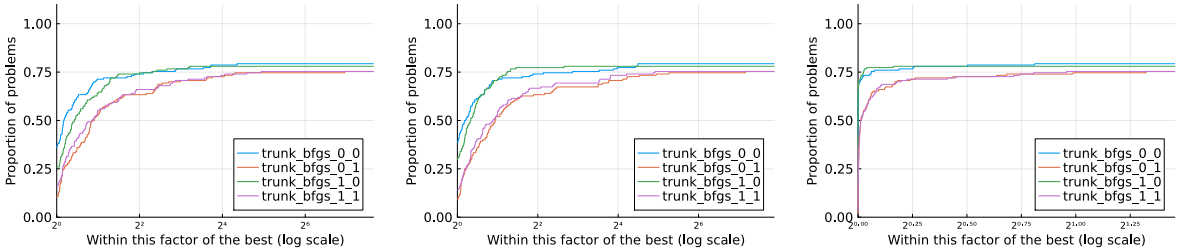


Figure 3: Performance profiles of trunk with LBFGS Hessian approximation in terms of number of objective evaluations (left), gradient evaluations (center), and CPU time (right). Method `trunk_bfgs_a_b` refers to the variant $\alpha = a$ and $\beta = b$.

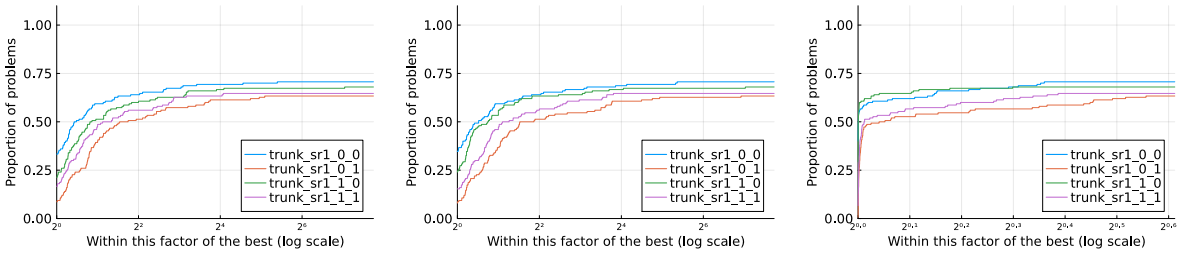


Figure 4: Performance profiles of trunk with LSR1 Hessian approximation in terms of number of objective evaluations (left), gradient evaluations (center), and CPU time (right). Method `trunk_sr1_a_b` refers to the variant $\alpha = a$ and $\beta = b$.

Figures 2 to 4 suggest no clear relationship between worst-case complexity and performance in practice, at least on this preliminary, and admittedly restricted, test set. To the authors, this situation is reminiscent of comparisons between the standard trust-region implementation `trunk_0_0` and adaptive cubic regularization (ARC) methods, which became popular due to their favorable $O(\epsilon^{-3/2})$ worst-case evaluation complexity. ARC methods are also covered by the analysis of Cartis et al. [1], and those authors are responsible for much of their development. In the experiments conducted by Dussault et al. [6], however, the performance of ARC methods is not sufficiently compelling to become a method of choice in practice. Future developments may, of course, have us revise that opinion.

6 Conclusion

We extended the complexity analysis of trust-region methods to handle potentially unbounded model Hessians in unconstrained optimization. Unlike traditional complexity analyses that assume uniformly bounded model Hessians, our study covers practical cases, including quasi-Newton updates such as PSB, BFGS, and SR1. We analyzed two regimes of the model Hessian growth: linear in the number of successful iterations and in the total number of iterations. When the model Hessians grow as $O(k^p)$, where k is the iteration counter and $0 \leq p < 1$, we derived a sharp $O(\epsilon^{-2/(1-p)})$ worst-case evaluation complexity bound to reach an ϵ-stationary point. Additionally, for the case where $p = 1$, we established a new $O(\exp(c\epsilon^{-2}))$ worst-case evaluation complexity bound, for some constant $c > 0$. We derived similar sharp bounds when the model Hessians grow linearly with the number of successful iterations. Among others, our results confirmed the profound intuition of Powell [16] on complexity for multiple quasi-Newton approximations.

Ongoing work will extend the complexity analysis to cover the adaptive regularization with cubics (ARC) framework [1], and investigate whether quasi-Newton approximation may indeed grow linearly with the number of successful iterations.

References

[1] 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, Philadelphia, USA, 2022. DOI: 10.1137/1.9781611976991.

[2] A. R. Conn, N. I. Gould, and Ph. L. Toint. Trust-Region Methods. Number 1 in MPS-SIAM Series on Optimization. SIAM, Philadelphia, USA, 2000. DOI: 10.1137/1.9780898719857.

[3] F. E. Curtis, Z. Lubberts, and D. P. Robinson. Concise complexity analyses for trust region methods. Optim. Lett., 12:1713–1724, 2018. DOI: doi.org/10.1007/s11590-018-1286-2.

[4] J. E. Dennis, Jr. and J. J. Moré. Quasi-Newton methods, motivation and theory. SIAM Rev., 19(1):46–89, 1977. DOI: 10.1137/1019005.

[5] E. D. Dolan and J. J. Moré. Benchmarking optimization software with performance profiles. Math. Program., Series B, 29(2):201–213, 2002. DOI: 10.1007/s101070100.

[6] J. Dussault, T. Migot, and D. Orban. Scalable adaptive cubic regularization methods. Math. Program., 2023. DOI: 10.1007/s10107-023-02007-6. Online.

[7] R. Fletcher. An algorithm for solving linearly constrained optimization problems. Math. Program., (2): 133–165, 1972. DOI: 10.1007/BF01584540.

[8] 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, 07 2014. DOI: 10.1007/s10107-014-0794-9.

[9] G. Leconte and D. Orban. Complexity of trust-region methods with unbounded Hessian approximations for smooth and nonsmooth optimization. Cahier du GERAD G–2023–65, GERAD, Montréal, QC, Canada, 2024.

[10] T. Migot, D. Orban, A. Soares Siqueira, and contributors. The JuliaSmoothOptimizers Ecosystem for Numerical Linear Algebra and Optimization in Julia, July 2024. URL `https://jso.dev`.

[11] T. Migot, D. Orban, A. Soares Siqueira, and contributors. JSOSolvers.jl: JuliaSmoothOptimizers optimization solvers, July 2024. URL `https://github.com/JuliaSmoothOptimizers/JSOSolvers.jl`.

[12] T. Migot, D. Orban, A. Soares Siqueira, and contributors. OptimizationProblems.jl: A collection of optimization problems in Julia, July 2024. URL `https://github.com/JuliaSmoothOptimizers/OptimizationProblems.jl`.

[13] 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. DOI: 10.1016/B978-0-12-597050-1.50006-3.

[14] 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. DOI: 10.1016/B978-0-12-468650-2.50005-5.

[15] M. J. D. Powell. On the global convergence of trust region algorithms for unconstrained minimization. Math. Program., (29):297–303, 1984. DOI: 10.1007/BF02591998.

[16] M. J. D. Powell. On the convergence of a wide range of trust region methods for unconstrained optimization. IMA J. Numer. Anal., 30(1):289–301, 2010. DOI: 10.1093/imanum/drp021.

[17] T. Steihaug. The conjugate gradient method and trust regions in large scale optimization. SIAM J. Numer. Anal., 20(3):626–637, 1983. DOI: 10.1137/0720042.

[18] Ph. Toint. Global Convergence of a a of Trust-Region Methods for Nonconvex Minimization in Hilbert Space. IMA J. Numer. Anal., 8(2):231–252, 1988. DOI: 10.1093/imanum/8.2.231.

[19] Ph. Toint. Nonlinear stepsize control, trust regions and regularizations for unconstrained optimization. Optim. Method Softw., 28:1–14, 01 2011. DOI: 10.1080/10556788.2011.610458.