Les Cahiers du GERAD ISSN: 0711–2440

A proximal modified quasi-Newton method for nonsmooth regularized optimization

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

G–2024–64

Septembre 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

A proximal modified quasi-Newton method for nonsmooth regularized optimization

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

Septembre 2024
Les Cahiers du GERAD
G–2024–64

Abstract : We develop R2N, a modified quasi-Newton method for minimizing the sum of a $\mathcal{C}^1$ function f and a lower semi-continuous prox-bounded h. Both f and h may be nonconvex. At each iteration, our method computes a step by minimizing the sum of a quadratic model of f, a model of h, and an adaptive quadratic regularization term. A step may be computed by way of a variant of the proximal-gradient method. An advantage of R2N over competing trust-region methods is that proximal operators do not involve an extra trust-region indicator. We also develop the variant R2DH, in which the model Hessian is diagonal, which allows us to compute a step without relying on a subproblem solver when h is separable. R2DH can be used as standalone solver, but also as subproblem solver inside R2N. We describe non-monotone variants of both R2N and R2DH. Global convergence of a first-order stationarity measure to zero holds without relying on local Lipschitz continuity of ∇f, while allowing model Hessians to grow unbounded, an assumption particularly relevant to quasi-Newton models. Under Lipschitz-continuity of ∇f, we establish a tight worst-case evaluation complexity bound of $O(1/\epsilon^{2/(1-p)})$ to bring said measure below $\epsilon > 0$, where $0 \leq p < 1$ controls the growth of model Hessians. Specifically, the latter must not diverge faster than $|\mathcal{S}_k|^p$, where $\mathcal{S}_k$ is the set of successful iterations up to iteration k. When $p = 1$, we establish the tight exponential complexity bound $O(\exp(c\epsilon^{-2}))$ where $c > 0$ is a constant. We describe our Julia implementation and report numerical experience on a classic basis-pursuit problem, an image denoising problem, a minimum-rank matrix completion problem, and a nonlinear support vector machine. In particular, the minimum-rank problem cannot be solved directly at this time by a trust-region approach as corresponding proximal operators are not known analytically.

Keywords : Nonsmooth optimization, nonconvex optimization, regularized optimization, composite optimization, modified quasi-Newton method, proximal quasi-Newton method, proximal gradient method

Résumé: Nous développons R2N, une méthode quasi-Newton modifiée pour minimiser la somme d'une fonction $\mathcal{C}^1$ f et d'une fonction h semi-continue inférieurement et prox-bornée. Les fonctions f et h peuvent toutes deux être non convexes. À chaque itération, notre méthode calcule un pas en minimisant la somme d'un modèle quadratique de f, d'un modèle de h, et d'un terme de régularisation quadratique adaptatif. Un pas peut être calculé au moyen d'une variante de la méthode du gradient proximal. Un avantage de R2N par rapport aux méthodes de région de confiance concurrentes est que les opérateurs proximaux n'impliquent pas d'indicateur de région de confiance supplémentaire. Nous développons également la variante R2DH, dans laquelle la hessienne du modèle est diagonale, ce qui permet de calculer un pas sans avoir recours à un solveur de sous-problème lorsque h est séparable. R2DH peut être utilisé comme solveur autonome, mais aussi comme solveur pour le sous-problème de R2N. Nous décrivons également des variantes non monotones de R2N et R2DH. La convergence globale d'une mesure de stationnarité de premier ordre vers zéro est garantie sans supposer que ∇f est (localement) Lipschitz continue, et sans imposer aux hessiennes du modèle d'être uniformément bornée, une hypothèse particulièrement pertinente pour les modèles quasi-Newton. Sous l'hypothèse que ∇f est Lipschitz continue, nous établissons une borne de complexité d'évaluation atteinte dans le pire des cas de $O(1/\epsilon^{2/(1-p)})$ pour amener ladite mesure sous $\epsilon > 0$, où $0 \leq p < 1$ contrôle la croissance des hessiennes du modèle. Plus précisément, ces dernières ne doivent pas diverger plus rapidement que $|\mathcal{S}_k|^p$, où $\mathcal{S}_k$ est l'ensemble des itérations réussies jusqu'à l'itération k. Lorsque $p = 1$, nous établissons la borne de complexité exponentielle atteinte $O(\exp(c\epsilon^{-2}))$, où $c > 0$ est une constante. Nous décrivons notre implémentation en Julia et rapportons des expériences numériques sur un problème classique de poursuite de base, un problème de débruitage d'image, un problème de complétion matricielle de rang minimal, et une machine à vecteurs de support non linéaire. En particulier, le problème de rang minimal ne peut pas être résolu directement à ce jour par une approche de région de confiance car les opérateurs proximaux correspondants ne sont pas connus analytiquement.

Acknowledgements: The research of Y. Diouane and D. Orban is partially supported by an NSERC Discovery Grant.

1 Introduction

We consider problems of the form

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

where $f : \mathbb{R}^n \to \mathbb{R}$ is $\mathcal{C}^1$ on $\mathbb{R}^n$, and $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is lower semi-continuous (lsc). Both f and h may be nonconvex. Our motivation is to develop a modified Newton variant of the trust-region algorithms of Aravkin et al. [3] and Leconte and Orban [21] because the proximal operators used in the subproblems should be easier to derive as they do not need to account for a trust-region indicator. We introduce method R2N, at each iteration of which the sum of a quadratic model of f, a model of h, and an adaptive quadratic regularization term, is approximately minimized. Both models may be nonconvex. The Hessian of the quadratic model of f may be that of f if it exists, or an approximation such as those derived from quasi-Newton updates. We establish global convergence of R2N under the assumption that the models of h are prox-bounded and approximate $h(x+s)$ as $o(\|s\|)$—an assumption that covers composite terms with Hölder Jacobian, see Model Assumption 5.1 for details. No assumption on local Lipschitz continuity of ∇f is required, nor is boundedness of the model Hessians, provided they do not diverge too fast. Specifically, if B_k is the model Hessian at iteration k, we require that the series with general term $1/(1 + \max_{0\leq j\leq k} \|B_j\|)$ diverge—an assumption similar to that used in trust-region methods [13, §8.4]. Our assumptions are significantly weaker than assumptions commonly found in the analysis of competing methods, and, consequently, the applicability of R2N is significantly more general—see the related research section below for details.

R2N specializes to method R2DH when B_k is diagonal, as did the solver of Leconte and Orban [21]. For a number of choices of separable h that are relevant in applications, steps can be computed explicitly without resort to an iterative subproblem solver. R2DH can be used as standalone solver or as subproblem solver inside R2N.

We also develop complexity results inspired from those of Leconte and Orban [22] and Diouane et al. [16], that account for potentially unbounded model Hessians. Specifically, we require that either $\|B_k\| = O(|\mathcal{S}_k|^p)$ for some $0 \leq p \leq 1$, where $\mathcal{S}_k$ is the set of successful iterations up to iteration k. When $0 \leq p < 1$, we establish a tight $O(\epsilon^{-2/(1-p)})$ complexity, and when $p = 1$, we establish a tight exponential complexity, i.e., a bound in $O(\exp(c\epsilon^{-2}))$ where $c > 0$ is a constant. Though the latter bound is tight, it is not known if it is attained for a quasi-Newton update.

We provide efficient implementations of R2N and R2DH. The latter can use one of several diagonal quasi-Newton updates. Both have non-monotone variants that preserve their convergence and complexity properties. Our open-source Julia implementations are available from [5]. In Section 8, we illustrate the performance of R2N and R2DH on challenging problems, including minimum-rank problems for which the trust-region methods of [3, 4] are impractical.

Contributions and related research

The proximal-gradient method [17, 26] is the prototypical first-order method for (1). Vast amounts of literature consider variants but restrict f and/or h to be convex, impose that f have (locally) Lipschitz-continuous gradient, or that h be Lipschitz continuous. For instance, [24] develop a proximal Newton method that requires f and h convex, a positive semi-definite Hessian, and solve the subproblem via the proximal-gradient method. Cartis et al. [10] require h to be globally Lipschitz continuous. Kanzow and Lechner [19] develop an approach closely related to ours, but for convex h. Others dispense with convexity but require coercivity of $f + h$ [25]. We are aware of few references that allow both f and h to be nonconvex. Bolte et al. [8] design an alternating method for problems with partitioned sets of variables. Stella et al. [31] propose a line search limited-memory BFGS method named PANOC. Themelis et al. [32] devise ZeroFPR, a non-monotone line search proximal quasi-Newton method based on the forward-backward envelope. Boţ et al. [9], study a proximal method with momentum. The last three converge if $f + h$ satisfies the Kurdyka-Łojasiewicz (KŁ) assumption. Kanzow and Mehlitz [20]

merely assume that f is $\mathcal{C}^1$, but that h is bounded below by an affine function, which is stronger than our prox-boundedness assumption.

Our work follows the scheme laid out by Aravkin et al. [3]; a trust-region framework applicable to nonconvex f and/or h, and that does not require coercivity or KŁ assumptions. They also describe a method named R2 that amounts to a proximal-gradient method with adaptive step size, and that may be viewed as R2N where B_k is set to zero at each iteration, effectively reducing to a first-order method. Aravkin et al. [4] specialize their trust-region method to problems where f has a least-squares structure, and develop a Levenberg-Marquardt variant named LM that may also be viewed as a special case of R2N for least-squares f. If J_k is the least-squares residual's Jacobian at x_k, their model of f uses $B_k = J_k^T J_k$. Leconte and Orban [21] devise variants of the trust-region method of [3] for separable h in which the model Hessian is a diagonal quasi-Newton approximation. They also devise non-monotone schemes that are shown to significantly improve performance in certain cases. All of [3, 4, 21] assume uniformly bounded second-order information in the model of f.

Leconte and Orban [22] revisit the trust-region method of [3] but allow for unbounded model Hessians. They establish global convergence and a worst-case complexity bound of $O(\epsilon^{-2/(1-p)})$ provided $\|B_k\| = O(|\mathcal{S}_k|^p)$ with $0 \leq p < 1$. To the best of our knowledge, they were the first to use that assumption and to obtain a complexity bound in the presence of unbounded model Hessians. Unfortunately, their analysis does not generalize to $p = 1$.

Potentially unbounded model Hessians are a relevant assumption in several contexts, including quasi-Newton methods. Conn et al. [13, §8.4.1.2] show that the SR1 approximation satisfies $\|B_k\| = O(|\mathcal{S}_k|)$, and a similar bound for BFGS when f is convex. Powell [28] establishes a similar bound for his PSB update. Even though it is not currently known whether those bounds are tight, the case $p = 1$ covers them.

Diouane et al. [16] generalized the results of [22] to $p = 1$ and provided tighter complexity constants when $0 \leq p < 1$ in the context of trust-region methods for smooth optimization, i.e., $h = 0$. Our complexity analysis draws from [16, 22].

Notation

Unless otherwise noted, if x is a vector, $\|x\|$ denotes its Euclidean norm and if A is a matrix, $\|A\|$ denotes its spectral norm. For positive sequences $\{a_k\}$ and $\{b_k\}$, we say that $a_k = o(b_k)$ if and only if $\limsup_k a_k/b_k = 0$. The cardinality of a finite set $\mathcal{A}$ is denoted $|\mathcal{A}|$. We denote $\mathbb{N}_0$ the set of positive integers.

2 Background

We recall relevant concepts of variational analysis, e.g., [29].

The domain of h is $\operatorname{dom} h := \{x \in \mathbb{R}^n \mid h(x) < \infty\}$. Because h is proper, $\operatorname{dom} h \neq \varnothing$. If $P : \mathbb{R}^n \to \mathbb{R}^n$ is a set-valued function, $\operatorname{dom} P = \{x \in \mathbb{R}^n \mid P(x) \neq \varnothing\}$.

Definition 1. (Limiting subdifferential) Consider $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$ and $\bar{x} \in \mathbb{R}^n$ such that $\phi(\bar{x}) < +\infty$. We say that $v \in \mathbb{R}^n$ is a regular subgradient of ϕ at $\bar{x}$ if

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

The set $\hat{\partial}\phi(\bar{x})$ of all regular subgradients of ϕ at $\bar{x}$ is called the Fréchet subdifferential.

The limiting subdifferential of ϕ at $\bar{x}$ is the set $\partial\phi(\bar{x})$ of all $v \in \mathbb{R}^n$ such that there is $\{x_k\} \to \bar{x}$ with $\{\phi(x_k)\} \to \phi(\bar{x})$ and $\{v_k\} \to v$ with $v_k \in \hat{\partial}\phi(x_k)$ for all k.

If $\phi = f + h$ with f continuously differentiable and h lower semi-continuous, then $\partial\phi(x) = \nabla f(x) + \partial h(x)$ [29, Theorem 10.1].

Definition 2. (Proximal Operator) Let $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be proper and lower semi-continuous. The proximal operator of h with step length $\nu > 0$ is

$$\operatorname*{prox}_{\nu h}(x) := \operatorname*{argmin}_y \; h(y) + \tfrac{1}{2}\nu^{-1}\|y - x\|^2.$$

Without further assumptions on h, the proximal operator might be empty, or contain one or more elements.

By [29, Exercise 8.8c], $\bar{x}$ is first-order stationary for (1) if $0 \in \nabla f(\bar{x}) + \partial h(\bar{x})$.

3 Models

For $\sigma \geq 0$, $x \in \mathbb{R}^n$, and $B(x) = B(x)^T \in \mathbb{R}^{n\times n}$, consider the models

$$\varphi(s; x) := f(x) + \nabla f(x)^T s + \tfrac{1}{2} s^T B(x) s \tag{2a}$$

$$\psi(s; x) \approx h(x + s) \tag{2b}$$

$$m(s; x, \sigma) := \varphi(s; x) + \tfrac{1}{2}\sigma\|s\|^2 + \psi(s; x). \tag{2c}$$

Note that (2c) represents a regularized second-order model of the objective of (1), where f and h are modeled separately. More details, on the use of such model to solve (1), will be given in Section 4. By construction, $\varphi(0; x) = f(x)$ and $\nabla\varphi(0; x) = \nabla f(x)$. We make the following assumption on (2b).

Model Assumption 3.1. For any $x \in \mathbb{R}^n$, $\psi(\cdot; x)$ is proper, lower semi continuous and prox-bounded with threshold $\lambda_x \in \mathbb{R}_+ \cup \{+\infty\}$ [29, Definition 1.23]. In addition, $\psi(0; x) = h(x)$, and $\partial\psi(0; x) = \partial h(x)$.

We make the following additional assumption and say that $\{\psi(\cdot; x)\}$ is *uniformly prox-bounded*.

Model Assumption 3.2. There is $\lambda \in \mathbb{R}_+ \cup \{+\infty\}$ such that $\lambda_x \geq \lambda$ for all $x \in \mathbb{R}^n$.

Model Assumption 3.2 is satisfied if h itself is prox-bounded and we select $\psi(s; x) := h(x + s)$ for all x. Let

$$p(x, \sigma) := \min_s \; m(s; x, \sigma) \leq m(0; x, \sigma) = f(x) + h(x) \tag{3a}$$

$$P(x, \sigma) := \operatorname*{argmin}_s \; m(s; x, \sigma), \tag{3b}$$

be the value function and the set of minimizers of (2c), respectively.

For $x \in \mathbb{R}^n$, $s \in P(x, \sigma) \implies 0 \in \nabla\varphi(s; x) + \sigma s + \partial\psi(s; x)$. Our first result states properties of the domain of p and P as given in (3a) and (3b).

Lemma 1. *Let Model Assumption 3.1 be satisfied and $B(x) = B(x)^T$ for all $x \in \mathbb{R}^n$. Then,* $\operatorname{dom} p = \mathbb{R}^n \times \mathbb{R}$. *In addition, if Model Assumption 3.2 holds,* $\operatorname{dom} P \supseteq \{(x, \sigma) \mid \sigma > \max(\lambda^{-1} - \lambda_{\min}(B(x)), \lambda^{-1})\}$, *where $\lambda_{\min}(B(x))$ is the smallest eigenvalue of $B(x)$.*

Proof. By definition of the domain and Model Assumption 3.1,

$$\begin{aligned}\operatorname{dom} p = \{(x, \sigma) \mid \inf_s m(s; x, \sigma) < +\infty\} &= \{(x, \sigma) \mid \exists s \; m(s; x, \sigma) < +\infty\} \\ &= \{(x, \sigma) \mid \exists s \; \psi(s; x) < +\infty\} = \mathbb{R}^n \times \mathbb{R},\end{aligned}$$

because $\psi(\cdot; x)$ is proper. Moreover,

$$\operatorname{dom} P = \{(x, \sigma) \mid \exists s(x, \sigma) \in \mathbb{R}^n, \; m(s(x, \sigma); x, \sigma) = \inf_s m(s; x, \sigma)\}.$$

Write

$$m(s;x,\sigma) = \varphi(s;x) + \tfrac{1}{2}(\sigma - \lambda^{-1})\|s\|^2 + \psi(s;x) + \tfrac{1}{2}\lambda^{-1}\|s\|^2.$$

By Model Assumption 3.2 and [29, Exercise 1.24(c)], there is $b \in \mathbb{R}$ such that $\psi(s;x) + \frac{1}{2}\lambda^{-1}\|s\|^2 \geq b$ for all $s \in \mathbb{R}^n$. Let $a \in \mathbb{R}$. The above and (2a) imply that the level set $\{s \in \mathbb{R}^n \mid m(s;x,\sigma) \leq a\}$ is contained in

$$\{s \in \mathbb{R}^n \mid \nabla f(x)^T s + \tfrac{1}{2}s^T(B(x) + (\sigma - \lambda^{-1})I)s \leq a - b - f(x)\},$$

which is a bounded set for $\sigma > \lambda^{-1} - \lambda_{\min}(B(x))$, i.e., $m(\cdot;x,\sigma)$ is level-bounded. Thus, [29, Theorem 1.9] implies that $\inf_s m(s;x,\sigma)$ is attained, i.e., that $P(x,\sigma) \neq \varnothing$. □

In Lemma 1, $\operatorname{dom} P = \{(x,\sigma) \mid \sigma > \max(\lambda^{-1} - \lambda_{\min}(B(x)),\ \lambda^{-1})\}$ does not hold in general. Consider for example a situation where $\psi(s;x)$ is bounded below for all $x \in \mathbb{R}^n$, i.e., each $\lambda_x = +\infty$. We can choose $\lambda = +\infty$. Assume also that, for a given $x \in \mathbb{R}^n$, $\varphi(s;x) = 0$ for all s, and $\psi(s;x)$ level-bounded. Then, $\lambda_{\min}(B(x)) = 0$, and for $\sigma = 0 = \lambda^{-1}$, $m(s;x,\sigma) = \psi(s;x)$. Therefore, $P(x,\sigma) \neq \varnothing$.

For a given $s(x,\sigma) \in P(x,\sigma)$, we define

$$\xi(x,\sigma) := f(x) + h(x) - (\varphi(s(x,\sigma);x) + \psi(s(x,\sigma);x)). \tag{4}$$

The next result relates (4) to first-order stationary for (1) and (2c).

Lemma 2. *Let Model Assumption 3.1 be satisfied, and $x \in \mathbb{R}^n$ and $\sigma \geq 0$ be given. Then, $\xi(x,\sigma) = 0 \iff 0 \in P(x,\sigma) \implies x$ is first-order stationary for (1).*

Proof. Firstly, $\xi(x,\sigma) = 0$ occurs if and only if $p(x,\sigma) = f(x) + h(x) = \varphi(0;x) + \psi(0;x)$, which occurs if and only if $0 \in P(x,\sigma)$. Therefore, $0 \in \partial m(0;x,\sigma) = \nabla\varphi(0;x) + \partial\psi(0;x) = \nabla f(x) + \partial h(x)$, and x is first-order stationary for (1). The same logic applies in the opposite direction, and establishes the equivalence. □

The following proposition states some properties of (3a) and (3b).

Proposition 1. *Let Model Assumptions 3.1 and 3.2 be satisfied. Assume also that ∇f is bounded over $\mathbb{R}^n$. Let $\epsilon > 0$. Then,*

1. *at any (x,σ) such that $\sigma \geq \lambda^{-1} - \lambda_{\min}(B(x)) + \epsilon$, p is finite and lsc, and $P(x,\sigma)$ is nonempty and compact;*
2. *if $\{(x_k,\sigma_k)\} \to (\bar{x},\bar{\sigma})$ with $\sigma_k \geq \lambda^{-1} - \lambda_{\min}(B(x_k)) + \epsilon$ for all k in such a way that $\{p(x_k,\sigma_k)\} \to p(\bar{x},\bar{\sigma})$, and for each k, $s_k \in P(x_k,\sigma_k)$, then $\{s_k\}$ is bounded and all its limit points are in $P(\bar{x},\bar{\sigma})$;*
3. *for any $x \in \mathbb{R}^n$, $p(\bar{x},\cdot)$ is continuous at any $\bar{\sigma} \geq \lambda^{-1} - \lambda_{\min}(B(\bar{x})) + \epsilon$ and $\{p(x_k,\sigma_k)\} \to p(\bar{x},\bar{\sigma})$ holds in part 2.*

Proof. The proof consists in establishing that (2c) is level-bounded in s locally uniformly in (x,σ) [29, Definition 1.16] for $\sigma \geq \lambda^{-1} - \lambda_{\min}(B(x)) + \epsilon$ and applying [29, Theorem 1.17]. It is nearly identical to that of [4, Proposition 3.2] and is omitted. □

Even though model (2c) is natural for incorporating second-order information, it is generally difficult to compute an exact minimizer of it. We proceed as Aravkin et al. [3, 4] and consider a simpler first-order model that will allow us to define an implementable stationary measure, to set minimal requirements steps computed in the course of the iterations of the algorithm of Section 4, and to derive convergence properties. This first-order model generalizes the concept of *Cauchy point* ("cp") when solving (1). For fixed $\nu > 0$ and $x \in \mathbb{R}^n$, define

$$\varphi_{\text{cp}}(s;x) := f(x) + \nabla f(x)^T s \tag{5a}$$

$$m_{\text{cp}}(s; x, \nu^{-1}) := \varphi_{\text{cp}}(s; x) + \tfrac{1}{2}\nu^{-1}\|s\|^2 + \psi(s; x) \tag{5b}$$

$$p_{\text{cp}}(x, \nu^{-1}) := \min_s \; m_{\text{cp}}(s; x, \nu^{-1}) \leq m_{\text{cp}}(0; x, \nu^{-1}) = f(x) + h(x) \tag{5c}$$

$$P_{\text{cp}}(x, \nu^{-1}) := \operatorname*{argmin}_s \; m_{\text{cp}}(s; x, \nu^{-1}) \tag{5d}$$

$$\xi_{\text{cp}}(x, \nu^{-1}) := f(x) + h(x) - (\varphi_{\text{cp}}(s_{\text{cp}}(x, \nu^{-1}); x) + \psi(s_{\text{cp}}(x, \nu^{-1}); x)), \tag{5e}$$

where $s_{\text{cp}} \in P_{\text{cp}}(x, \nu^{-1})$. By [8, Lemma 2],

$$\xi_{\text{cp}}(x, \nu^{-1}) \geq \tfrac{1}{2}\nu^{-1}\|s_{\text{cp}}(x, \nu^{-1})\|^2 \geq 0. \tag{6}$$

In the smooth case, i.e., $h = 0$ and $\psi = 0$, $s_{\text{cp}} = -\nu\nabla f(x)$, so that

$$\xi_{\text{cp}}(x, \nu^{-1}) \geq \tfrac{1}{2}\nu^{-1}\|s_{\text{cp}}\|^2 = \tfrac{1}{2}\nu\|\nabla f(x)\|^2,$$

which suggests $\nu^{-1/2}\xi_{\text{cp}}(x, \nu^{-1})^{1/2}$ as a stationarity measure that generalizes the norm of the gradient to the nonsmooth setting.

The next results establish corresponding properties of p_{cp} and P_{cp}. The proofs are similar to those of Lemmas 1 and 2 and proposition 1 and are omitted.

Lemma 3. *Let Model Assumption 3.1 be satisfied. Then,* $\operatorname{dom} p_{\text{cp}} = \mathbb{R}^n \times \mathbb{R}$. *If Model Assumption 3.2 holds,* $\operatorname{dom} P_{\text{cp}} \supseteq \{(x, \nu^{-1}) \mid \nu > \max(\lambda^{-1} - \lambda_{\min}(B(x)),\, \lambda^{-1})\}$.

The next result characterizes first-order stationarity for (1).

Lemma 4. *Let Model Assumption 3.1 be satisfied and* $\nu > 0$. *Then,*
$\xi_{\text{cp}}(x, \nu^{-1}) = 0 \iff 0 \in P_{\text{cp}}(x, \nu^{-1}) \implies x$ *is first-order stationary for* (1).

The following result states properties of (5c) and (5d).

Proposition 2. *Let Model Assumptions 3.1 and 3.2 be satisfied and* $\nabla f(x)$ *be bounded over* $\mathbb{R}^n$. *Let* $\epsilon > 0$. *Then,*

1. *at any* (x, ν^{-1}) *with* $\nu^{-1} \geq \lambda^{-1} + \epsilon$, p_{cp} *is finite and lsc, and* $P_{\text{cp}}(x, \nu^{-1})$ *is nonempty and compact;*
2. *if* $\{(x_k, \nu_k^{-1})\} \to (\bar{x}, \bar{\nu}^{-1})$ *with* $\nu_k^{-1} \geq \lambda^{-1} + \epsilon$ *for all* k *in such a way that* $\{p_{\text{cp}}(x_k, \nu_k^{-1})\} \to p_{\text{cp}}(\bar{x}, \bar{\nu}^{-1})$, *and for each* k, $s_k \in P_{\text{cp}}(x_k, \nu_k^{-1})$, *then* $\{s_k\}$ *is bounded and all its limit points are in* $P_{\text{cp}}(\bar{x}, \bar{\nu}^{-1})$;
3. *for any* $\bar{x} \in \mathbb{R}^n$ *and any* $\bar{\nu}^{-1} \geq \lambda^{-1} + \epsilon$, $p_{\text{cp}}(\bar{x}, \cdot)$ *is continuous at* $\bar{\nu}$ *and* $\{p_{\text{cp}}(x_k, \nu_k^{-1})\} \to p_{\text{cp}}(\bar{x}, \bar{\nu}^{-1})$ *holds in part 2.*

The main idea of the algorithm proposed in Section 4 is that (2c) is approximately minimized at each iteration. In order to establish convergence, the step s thus computed is required to satisfy *Cauchy decrease*, which we define as in [3, 4]:

$$\varphi(0; x) + \psi(0; x) - (\varphi(s; x) + \psi(s; x)) \geq (1 - \theta_1)\xi_{\text{cp}}(x, \nu^{-1}), \tag{7}$$

for a preset value of $\theta_1 \in (0, 1)$. In other words, s must result in a decrease in $\varphi(\cdot; x) + \psi(\cdot; x)$ that is at least a fraction of the decrease of the Cauchy model $\varphi_{\text{cp}}(\cdot; x) + \psi(\cdot; x)$ obtained with the Cauchy step s_{cp} and a well-chosen step length ν.

The following result parallels [21, Proposition 4] and establishes that if a step s reduces (2c) at least as much as s_{cp} does, Cauchy decrease holds. This observation is important because the first step of the proximal-gradient method from $s = 0$ applied to (5b) and to (2c) with step length ν is the same, and that step is precisely s_{cp}. Therefore, a step s may be obtained by continuing the proximal-gradient iterations on (2c) from s_{cp}.

Proposition 3. *Let Model Assumption 3.1 be satisfied. Let $x \in \mathbb{R}^n$, $\theta_1 \in (0, 1)$, $\sigma > 0$ and let s_{cp} be computed with $\nu = \theta_1/(\|B(x)\| + \sigma)$. Assume $s \in \mathbb{R}^n$ is such that $m(s; x, \sigma) \leq m(s_{\text{cp}}; x, \sigma)$. Then, s satisfies (7).*

Proof. Let $x \in \mathbb{R}^n$, $\sigma > 0$, and $s \in \mathbb{R}^n$, such that $m(s; x, \sigma) \leq m(s_{\text{cp}}; x, \sigma)$. Then,

$$\begin{aligned}\varphi(s; x) + \psi(s; x) + \tfrac{1}{2}\sigma\|s\|^2 &\leq \varphi(s_{\text{cp}}; x) + \psi(s_{\text{cp}}; x) + \tfrac{1}{2}\sigma\|s_{\text{cp}}\|^2 \\ &= \varphi_{\text{cp}}(s_{\text{cp}}; x) + \psi(s_{\text{cp}}; x) + \tfrac{1}{2}s_{\text{cp}}^T B(x)s_{\text{cp}} + \tfrac{1}{2}\sigma\|s_{\text{cp}}\|^2.\end{aligned}$$

The Cauchy-Schwarz inequality $s_{\text{cp}}^T B(x)s_{\text{cp}} \leq \|B(x)\|\|s_{\text{cp}}\|^2$, the identity $\varphi(0; x) = \varphi_{\text{cp}}(0; x)$ and (6) yield

$$\begin{aligned}\varphi(0; x) + \psi(0; x) - \varphi(s; x) - \psi(s; x) &\geq \xi_{\text{cp}}(x, \nu^{-1}) - \tfrac{1}{2}(\|B(x)\| + \sigma)\|s_{\text{cp}}\|^2 + \tfrac{1}{2}\sigma\|s\|^2 \\ &\geq \xi_{\text{cp}}(x, \nu^{-1}) - \tfrac{1}{2}(\|B(x)\| + \sigma)\|s_{\text{cp}}\|^2 \\ &\geq \xi_{\text{cp}}(x, \nu^{-1}) - (\|B(x)\| + \sigma)\nu\xi_{\text{cp}}(x, \nu^{-1}) \\ &= (1 - \theta_1)\xi_{\text{cp}}(x, \nu^{-1}).\end{aligned}$$

□

Computing $\|B(x)\|$ in the spectral norm comes at a cost. However, as we now illustrate, an inexact computation is sufficient in order to ensure (7). Assume that we are able to compute $\beta(x) \approx \|B(x)\|$ such that $\beta(x) \geq \mu\|B(x)\|$ for $0 < \mu < 1$, and set $\nu = \theta_1/(\beta(x) + \sigma)$. The proof of Proposition 3 continues to apply unchanged until the very last line, which becomes

$$\begin{aligned}\varphi(0; x) + \psi(0; x) - (\varphi(s; x) + \psi(s; x)) &\geq (1 - (\|B(x)\| + \sigma)\nu)\xi_{\text{cp}}(x, \nu^{-1}) \\ &= \left(1 - \theta_1\frac{\|B(x)\| + \sigma}{\beta(x) + \sigma}\right)\xi_{\text{cp}}(x, \nu^{-1}).\end{aligned}$$

If $\beta(x) \leq \|B(x)\|$, $(\|B(x)\| + \sigma)/(\beta(x) + \sigma) \leq \|B(x)\|/\beta(x) \leq 1/\mu$, so that

$$\left(1 - \theta_1\frac{\|B(x)\| + \sigma}{\beta(x) + \sigma}\right)\xi_{\text{cp}}(x, \nu^{-1}) \geq (1 - \theta_1/\mu)\xi_{\text{cp}}(x, \nu^{-1}).$$

Thus, as long as $\theta_1 < \mu$, (7) is satisfied with θ_1 replaced with θ_1/μ.

If, on the other hand, $\beta(x) \geq \|B(x)\|$, then $(\|B(x)\| + \sigma)/(\beta(x) + \sigma) \leq 1$, and

$$\left(1 - \theta_1\frac{\|B(x)\| + \sigma}{\beta(x) + \sigma}\right)\xi_{\text{cp}}(x, \nu^{-1}) \geq (1 - \theta_1)\xi_{\text{cp}}(x, \nu^{-1}),$$

and (7) holds unchanged.

The above observation also allows us to replace $\|B(x)\|$ in the denominator of ν with, e.g., $\|B(x)\|_1$, $\|B(x)\|_\infty$ or $\|B(x)\|_F$ if $B(x)$ is available as an explicit matrix, or indeed with any other norm of $B(x)$.

4 A modified quasi-Newton method for nonsmooth optimization

We are in position to describe a modified quasi-Newton method to solve (1) named R2N. By contrast with trust-region-based approaches [3, 21], proximal operators are easier to evaluate in the R2N subproblem as they do not include a trust-region indicator.

At iteration k, we choose a step length $\nu_k > 0$ based on the regularization parameter $\sigma_k > 0$ and the norm of the model Hessian $B(x_k)$ at the current iterate $x_k \in \mathbb{R}^n$ as in Proposition 3. We then compute

the Cauchy step $s_{k,\text{cp}}$ as a minimizer of (5b). A step s_k is subsequently computed that satisfies the assumptions of Proposition 3.

The rest of the algorithm is standard. The decrease in $f + h$ 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 σ_k is possibly reduced. If the model turns out to predict poorly the actual decrease, the trial point is rejected and σ_k is increased. Algorithm 1 states the whole procedure.

The interaction between σ_k and the unknown threshold λ_{x_k} works as in [3, Algorithm 6.1] and [4]. If $\sigma_k \leq \lambda_{x_k}^{-1}$, $\psi(s_k; x_k) = -\infty$, and according to the rules of extended arithmetic, which state that $\pm\infty \cdot 0 = 0 \cdot (\pm\infty) = (\pm\infty)/(\pm\infty) := 0$ [29], $\rho_k = 0$. Consequently, s_k will be rejected at Line 10, and σ_{k+1} will be set larger than σ_k at Line 11. After a finite number of such increments, σ_k will surpass $\lambda_{x_k}^{-1}$, resulting in a step with finite $\psi(s_k; x_k)$. In effect, Model Assumption 3.2 is only required to hold at the iterates generated by the algorithm.

Algorithm 1 R2N: A proximal modified Quasi-Newton method.

1: Choose constants $0 < \theta_1 < 1 < \theta_2$, $0 < \eta_1 \leq \eta_2 < 1$ and $0 < \gamma_3 \leq 1 < \gamma_1 \leq \gamma_2$.
2: Choose $x_0 \in \mathbb{R}^n$ where h is finite, $0 < \sigma_{\min} < \sigma_0$.
3: **for** $k = 0, 1, \ldots$ **do**
4: Choose $B_k := B(x_k) \in \mathbb{R}^{n\times n}$ such that $B_k = B_k^T$.
5: Compute $\nu_k := \theta_1/(\|B_k\| + \sigma_k)$.
6: Compute $s_{k,\text{cp}} \in \operatorname{argmin}_s \; m_{\text{cp}}(s; x_k, \nu_k^{-1})$ and $\xi_{\text{cp}}(x_k, \nu_k^{-1})$ as defined in (5e).
7: Compute a step s_k such that $m(s_k; x_k, \sigma_k) \leq m(s_{k,\text{cp}}; x_k, \sigma_k)$.
8: If $\|s_k\| > \theta_2 \, \|s_{k,\text{cp}}\|$, reset $s_k = s_{k,\text{cp}}$.
9: Compute the ratio

$$\rho_k := \frac{f(x_k) + h(x_k) - (f(x_k + s_k) + h(x_k + s_k))}{\varphi(0; x_k) + \psi(0; x_k) - (\varphi(s_k; x_k) + \psi(s_k; x_k))}.$$

10: If $\rho_k \geq \eta_1$, set $x_{k+1} = x_k + s_k$. Otherwise, set $x_{k+1} = x_k$.
11: Update the regularization parameter according to

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

12: **end for**

Importantly, R2N does not require $B_k \succeq 0$, which may be useful in practice in order to capture natural problem curvature. In addition, we allow $\{B_k\}$ to be unbounded. In Section 5, we establish convergence provided it does not diverge too fast, using an assumption similar to that used in trust-region methods [13, §8.4]. In Section 6, we study the effect of using different bounds on $\{\|B_k\|\}$ on worst-case evaluation complexity. The complexity results are obtained by adapting results from [16] (in the context of trust-region methods for smooth optimization) to R2N.

Our main working assumption is the following.

Problem Assumption 4.1. The function f is continuously differentiable over the set $\{x \in \mathbb{R}^n \mid (f + h)(x) \leq (f + h)(x_0)\}$ and h is proper and lower semi-continuous.

Problem Assumption 4.1 is very mild as one does not require boundedness nor Lipschitz continuity of f or ∇f, in contrast with [4, Problem Assumption 4.1] or the assumptions of Kanzow and Lechner [19]. For instance, our analysis includes cases where f is continuously differentiable, but whose gradient is not locally Lipschitz continuous at $x = 0$, e.g., $f(x) = |x|^{\frac{3}{2}}$.

In the next sections, we derive convergence and worst-case complexity analysis for Algorithm 1. We will repeatedly use the notation

$$\mathcal{S} := \{i \in \mathbb{N} \mid \rho_i \geq \eta_1\} \qquad \text{(all successful iterations)} \tag{8a}$$
$$\mathcal{S}_k := \{i \in \mathcal{S} \mid i \leq k\} \qquad \text{(successful iterations until iteration } k) \tag{8b}$$
$$\mathcal{U} := \{i \in \mathbb{N} \mid \rho_i < \eta_1\} \qquad \text{(all unsuccessful iterations)} \tag{8c}$$
$$\mathcal{U}_k := \{i \in \mathbb{N} \mid i \notin \mathcal{S}, \; i \leq k\} \qquad \text{(unsuccessful iterations until iteration } k). \tag{8d}$$

5 Convergence analysis of Algorithm 1

In this section, we investigate the convergence properties of Algorithm 1 under Problem Assumption 4.1. We show that $\liminf_{k\to\infty} \nu_k^{-\frac{1}{2}} \xi_{\text{cp}}(x_k, \nu_k^{-1})^{\frac{1}{2}} = 0$. We stress that the obtained convergence properties of Algorithm 1 are more general than those of [4, 19, 20], and do not require boundedness of the model Hessians nor (local) Lipschitz continuity of ∇f.

We first establish lower bounds on $\xi_{\text{cp}}(x_k, \nu_k^{-1})$ in terms of $\|s_k\|$.

Lemma 5. *For all $k \in \mathbb{N}$,*

$$\xi_{\text{cp}}(x_k, \nu_k^{-1}) \geq \tfrac{1}{2\theta_2^2} \nu_k^{-1} \|s_k\|^2. \tag{9}$$

Additionally, for any $\alpha > 0$,

$$\nu_k^{-\frac{1}{2}} \xi_{\text{cp}}(x_k, \nu_k^{-1})^{\frac{1}{2}} \geq \alpha \quad \Rightarrow \quad \xi_{\text{cp}}(x_k; \nu_k^{-1}) \geq \frac{\alpha}{\theta_2\sqrt{2}} \|s_k\|. \tag{10}$$

Proof. From Algorithm 1, we have $\|s_k\| \leq \theta_2 \|s_{k,\text{cp}}\|$. Hence,

$$\xi_{\text{cp}}(x_k, \nu_k^{-1}) \geq \tfrac{1}{2} \nu_k^{-1} \|s_{k,\text{cp}}\|^2 \geq \tfrac{1}{2\theta_2^2} \nu_k^{-1} \|s_k\|^2.$$

If $\nu_k^{-\frac{1}{2}} \xi_{\text{cp}}(x_k, \nu_k^{-1})^{\frac{1}{2}} \geq \alpha$,

$$\xi_{\text{cp}}(x_k, \nu_k^{-1}) \geq \alpha \nu_k^{\frac{1}{2}} \xi_{\text{cp}}(x_k, \nu_k^{-1})^{\frac{1}{2}} \geq \alpha \nu_k^{\frac{1}{2}} (\tfrac{1}{2\theta_2^2} \nu_k^{-1} \|s_k\|^2)^{\frac{1}{2}} = \frac{\alpha}{\theta_2\sqrt{2}} \|s_k\|. \qquad \square$$

The next lemma shows that the convergence of $\{x_k\}_{k\in\mathbb{N}}$ holds if the objective is bounded below, the algorithm generates infinitely many successful iterations and the stationarity measure $\nu_k^{-1/2} \xi_{\text{cp}}(x_k, \nu_k^{-1})^{1/2}$ is bounded away from zero.

Lemma 6. *Assume that Algorithm 1 generates infinitely many successful iterations and that there is $(f+h)_{low} \in \mathbb{R}$ such that $(f+h)(x_k) \geq (f+h)_{low}$ for all $k \in \mathbb{N}$. Additionally, assume, that there is $\alpha > 0$ such that for all $k \in \mathbb{N}$, $\nu_k^{-1/2} \xi_{\text{cp}}(x_k, \nu_k^{-1})^{1/2} \geq \alpha$. Then, $\{x_k\}_{k\in\mathbb{N}}$ is a Cauchy sequence, and hence converges.*

Proof. For all $k \in \mathcal{S}$, using (10) from Lemma 5, we have

$$\begin{aligned} f(x_k) + h(x_k) - f(x_{k+1}) - h(x_{k+1}) &\geq \eta_1(1-\theta_1)\xi_{\text{cp}}(x_k, \nu_k^{-1}) \\ &\geq \frac{\eta_1(1-\theta_1)\alpha}{\theta_2\sqrt{2}} \|s_k\| = \frac{\eta_1(1-\theta_1)\alpha}{\theta_2\sqrt{2}} \|x_{k+1} - x_k\|. \end{aligned}$$

Summing over all successful iterations from 1 to k, we obtain

$$\begin{aligned} f(x_0) + h(x_0) - (f+h)_{\text{low}} &\geq \sum_{j\in\mathcal{S}_k} f(x_j) + h(x_j) - f(x_{j+1}) - h(x_{j+1}) \\ &\geq \frac{\eta_1(1-\theta_1)\alpha}{\theta_2\sqrt{2}} \sum_{j\in\mathcal{S}_k} \|x_{j+1} - x_j\| \\ &= \frac{\eta_1(1-\theta_1)\alpha}{\theta_2\sqrt{2}} \sum_{j=0}^{k} \|x_{j+1} - x_j\| . \end{aligned}$$

Thus, $\sum_{j\in\mathbb{N}} \|x_{j+1} - x_j\| < +\infty$. Hence, $\{x_k\}_{k\in\mathbb{N}}$ is a Cauchy sequence, and converges. $\square$

The following lemma shows that when $\{x_k\}$ converges and $\{\sigma_k\}$ diverges along common subsequences, the corresponding subsequence of $\{s_k\}$ converges to zero.

Lemma 7. *Let Problem Assumption 4.1 and Model Assumptions 3.1 and 3.2 be satisfied and assume that there is an index set $\mathcal{K} \subseteq \mathbb{N}$ such that $\lim_{k\in\mathcal{K}} \sigma_k = +\infty$ and $\{x_k\}_{k\in\mathcal{K}}$ is bounded. Then, $\lim_{k\in\mathcal{K}} s_k = \lim_{k\in\mathcal{K}} s_{k,\text{cp}} = 0$.*

Proof. By contradiction, assume that there is an index set $\mathcal{K}' \subseteq \mathcal{K}$ and $\alpha > 0$ such that $\|s_{k,\text{cp}}\| \geq \alpha$ for all $k \in \mathcal{K}'$. By definition of $s_{k,\text{cp}}$ and Model Assumption 3.1, $f(x_k) + h(x_k) = m_{\text{cp}}(0; x_k, \nu_k^{-1}) \geq m_{\text{cp}}(s_{k,\text{cp}}; x_k, \nu_k^{-1})$. Hence,

$$\begin{aligned} f(x_k) + h(x_k) &\geq \varphi_{\text{cp}}(s_{k,\text{cp}}; x_k) + \psi(s_{k,\text{cp}}; x_k) + \tfrac{1}{2}\nu_k^{-1}\|s_{k,\text{cp}}\|^2 \\ &= f(x_k) + \nabla f(x_k)^T s_{k,\text{cp}} + \tfrac{1}{2\theta_1}(\|B_k\| + \sigma_k)\|s_{k,\text{cp}}\|^2 + \psi(s_{k,\text{cp}}; x_k) \\ &\geq f(x_k) + \nabla f(x_k)^T s_{k,\text{cp}} + \tfrac{1}{2\theta_1}\sigma_k\|s_{k,\text{cp}}\|^2 + \psi(s_{k,\text{cp}}; x_k) \\ &\geq f(x_k) - \|\nabla f(x_k)\|\|s_{k,\text{cp}}\| + \tfrac{1}{2}(\frac{\sigma_k}{\theta_1} - \lambda^{-1})\|s_{k,\text{cp}}\|^2 \\ &\quad + \psi(s_{k,\text{cp}}; x_k) + \tfrac{1}{2}\lambda^{-1}\|s_{k,\text{cp}}\|^2. \end{aligned}$$

By Model Assumption 3.2 and [29, Exercise 1.24(*c*)], there is $b_h \in \mathbb{R}$ such that $\psi(s; x) + \frac{1}{2}\lambda^{-1}\|s\|^2 \geq b_h$ for all s and x. Hence, for all sufficiently large $k \in \mathcal{K}'$, $\sigma_k > \lambda^{-1}$ and

$$\begin{aligned} f(x_k) + h(x_k) &\geq f(x_k) - \|\nabla f(x_k)\|\|s_{k,\text{cp}}\| + \tfrac{1}{2}(\frac{\sigma_k}{\theta_1} - \lambda^{-1})\|s_{k,\text{cp}}\|^2 + b_h \\ &\geq f(x_k) - \|\nabla f(x_k)\|\|s_{k,\text{cp}}\| + \tfrac{1}{2}\alpha(\frac{\sigma_k}{\theta_1} - \lambda^{-1})\|s_{k,\text{cp}}\| + b_h \\ &= f(x_k) + \left(\tfrac{1}{2}\alpha(\frac{\sigma_k}{\theta_1} - \lambda^{-1}) - \|\nabla f(x_k)\|\right)\|s_{k,\text{cp}}\| + b_h. \end{aligned} \tag{11}$$

Since $\{x_k\}_{k\in\mathcal{K}'}$ is bounded, so are $\{f(x_k)\}_{k\in\mathcal{K}'}$ and $\{\nabla f(x_k)\}_{k\in\mathcal{K}'}$ by Problem Assumption 4.1. Let $b_f := \min_{k\in\mathcal{K}'} f(x_k) > -\infty$ and $b_{f'} = \max_{k\in\mathcal{K}'} \|\nabla f(x_k)\| < \infty$. Because $\{f(x_k) + h(x_k)\}$ is nonincreasing, (11) yields

$$f(x_0) + h(x_0) \geq f(x_k) + h(x_k) \geq b_f + \left(\tfrac{1}{2}\alpha(\frac{\sigma_k}{\theta_1} - \lambda^{-1}) - b_{f'}\right)\|s_{k,\text{cp}}\| + b_h. \tag{12}$$

As $\lim_{k\in\mathcal{K}'} \sigma_k = +\infty$, for k sufficiently large, $\frac{1}{2}\alpha(\frac{\sigma_k}{\theta_1} - \lambda^{-1}) > b_{f'}$. Thus, for all sufficiently large $k \in \mathcal{K}'$, (12) combines with $\|s_{k,\text{cp}}\| \geq \alpha$ to give

$$f(x_0) + h(x_0) \geq b_f + \left(\tfrac{1}{2}\alpha(\frac{\sigma_k}{\theta_1} - \lambda^{-1}) - b_{f'}\right)\alpha + b_h,$$

which is a contradiction because the right-hand side diverges. Thus, $\lim_{k\in\mathcal{K}} \|s_{k,\text{cp}}\| = 0$. Finally, since $\|s_k\| \leq \theta_2\|s_{k,\text{cp}}\|$, we get also $\lim_{k\in\mathcal{K}} \|s_k\| = 0$. □

For the remainder of this section, we need the following assumption.

Model Assumption 5.1. For all $k \in \mathbb{N}$, the model function $\psi(\cdot, x_k)$ satisfies

$$|h(x_k + s_k) - \psi(s_k; x_k)| = o(\|s_k\|) \quad \text{as} \quad s_k \to 0. \tag{13}$$

Model Assumption 5.1 is trivially satisfied if, at each iteration k, we set $\psi(s; x_k) = h(x_k + s)$, which is what Kanzow and Lechner [19] do. However, the assumption also holds when $h(x) = g(c(x))$, where $c : \mathbb{R}^n \to \mathbb{R}^m$ has Lipschitz-continuous or α_h-Hölder-continuous Jacobian, $g : \mathbb{R}^m \to \mathbb{R}$ is L-Lipschitz continuous, and we choose $\psi(s; x_k) := g(c(x_k) + \nabla c(x_k)^T s)$. Indeed, there exists $M > 0$ such that $|h(x_k + s) - \psi(s; x_k)| \leq L\|c(x_k + s) - c(x_k) - \nabla c(x_k)^T s\| \leq LM\|s\|^{1+\alpha_h} = o(\|s\|)$.

Theorem 1. *Let Problem Assumption 4.1 and Model Assumptions 3.1, 3.2 and 5.1 be satisfied. Assume that there is an index set $\mathcal{K} \subseteq \mathbb{N}$ so that (i) there is $\alpha > 0$ such that $\nu_k^{-1/2}\xi_{\text{cp}}(x_k, \nu_k^{-1})^{1/2} \geq \alpha$ for all $k \in \mathcal{K}$, (ii) $\{\sigma_k(1 + \|B_k\|)^{-1}\}_{k\in\mathcal{K}}$ is unbounded and (iii) $\{x_k\}_{k\in\mathcal{K}}$ is bounded. Then, there is an index set $\mathcal{K}' \subseteq \mathcal{K}$ such that for all $k \in \mathcal{K}'$ sufficiently large, k is a very successful iteration.*

Proof. By Assumption (ii), there is an index set $\mathcal{K}' \subset \mathcal{K}$ such that $\lim_{k\in\mathcal{K}'} \sigma_k(1 + \|B_k\|)^{-1} = \infty$. Since $\sigma_k \geq \sigma_k(1 + \|B_k\|)^{-1}$, we also have $\lim_{k\in\mathcal{K}'} \sigma_k = \infty$. Lemma 7 then implies $\lim_{k\in\mathcal{K}'} \|s_k\| = 0$. For all $k \in \mathcal{K}'$, Model Assumption 5.1 combines with (7) and a Taylor expansion of f about x_k to give

$$\begin{aligned}
|\rho_k - 1| &= \left| \frac{(f+h)(x_k + s_k) - (\varphi(s_k; x_k) + \psi(s_k; x_k))}{\varphi(0; x_k) + \psi(0; x_k) - (\varphi(s_k; x_k) + \psi(s_k; x_k))} \right| \\
&= \left| \frac{(f+h)(x_k + s_k) - (f(x_k) + \nabla f(x_k)^T s_k + \frac{1}{2} s_k^T B_k s_k + \psi(s_k; x_k))}{\varphi(0; x_k) + \psi(0; x_k) - (\varphi(s_k; x_k) + \psi(s_k; x_k))} \right| \\
&\leq \frac{|f(x_k + s_k) - f(x_k) - \nabla f(x_k)^T s_k|}{(1-\theta_1)\xi_{\text{cp}}(x_k, \nu_k^{-1})} + \frac{\|B_k\|\|s_k\|^2}{2(1-\theta_1)\xi_{\text{cp}}(x_k, \nu_k^{-1})} \\
&\quad + \frac{|h(x_k + s_k) - \psi(s_k; x_k)|}{(1-\theta_1)\xi_{\text{cp}}(x_k, \nu_k^{-1})} \\
&= \frac{o(\|s_k\|)}{(1-\theta_1)\xi_{\text{cp}}(x_k, \nu_k^{-1})} + \frac{\|B_k\|\|s_k\|^2}{2(1-\theta_1)\xi_{\text{cp}}(x_k, \nu_k^{-1})} + \frac{o(\|s_k\|)}{(1-\theta_1)\xi_{\text{cp}}(x_k, \nu_k^{-1})} \\
&\leq \frac{o(\|s_k\|)}{\xi_{\text{cp}}(x_k, \nu_k^{-1})} + \frac{(1 + \|B_k\|)\|s_k\|^2}{2(1-\theta_1)\xi_{\text{cp}}(x_k, \nu_k^{-1})}.
\end{aligned} \tag{14}$$

By Assumption (i), Lemma 5 implies $\xi_{\text{cp}}(x_k; \nu_k^{-1}) \geq \frac{\alpha}{\theta_2\sqrt{2}}\|s_k\|$ for all $k \in \mathcal{K}'$, which we apply to the first term in the right-hand side of (14). Similarly, (9) implies

$$\xi_{\text{cp}}(x_k, \nu_k^{-1}) \geq \tfrac{1}{2\theta_2^2}\nu_k^{-1}\|s_k\|^2 = \tfrac{1}{2\theta_1\theta_2^2}(\|B_k\| + \sigma_k)\|s_k\|^2 \geq \tfrac{1}{2\theta_1\theta_2^2}\sigma_k\|s_k\|^2,$$

which we apply to the second term in the right-hand side of (14). Hence, (14) simplifies to

$$|\rho_k - 1| \leq \frac{o(\|s_k\|)}{\frac{\alpha}{\theta_2\sqrt{2}}\|s_k\|} + \frac{(1 + \|B_k\|)\|s_k\|^2}{\frac{(1-\theta_1)}{\theta_1\theta_2^2}\sigma_k\|s_k\|^2} = \frac{o(\|s_k\|)}{\|s_k\|} + \frac{\theta_1\theta_2^2}{(1-\theta_1)\sigma_k(1 + \|B_k\|)^{-1}}. \tag{15}$$

By Assumption (ii), the right-hand side of (15) converges to zero. Thus, for all sufficiently large $k \in \mathcal{K}'$, $|\rho_k - 1| \leq 1 - \eta_2$, which implies that $\rho_k \geq \eta_2$. □

Theorem 1 shares similarities with [4, Theorem 4.1] but uses weaker assumptions. In particular, compared to [4, Theorem 4.1], we do not use the Lipschitz continuity of ∇f nor do we require model Hessians to be uniformly bounded.

Lemma 8. *Let Problem Assumption 4.1 and Model Assumptions 3.1, 3.2 and 5.1 be satisfied. Assume that $\{x_k\}_{k\in\mathbb{N}}$ is bounded and that there is $\alpha > 0$ such that for all $k \in \mathbb{N}$, $\nu_k^{-1/2}\xi_{\text{cp}}(x_k, \nu_k^{-1})^{1/2} \geq \alpha$. Then, $\{\sigma_k(1 + \|B_k\|)^{-1}\}_{k\in\mathbb{N}}$ is bounded.*

Proof. Assume, by contradiction, that $\{\sigma_k(1 + \|B_k\|)^{-1}\}_{k\in\mathbb{N}}$ is unbounded. Since $\{\sigma_k\}_{k\in\mathbb{N}}$ increases only on unsuccessful iterations and $\{(1 + \|B_k\|)^{-1}\}_{k\in\mathbb{N}}$ is bounded, $\{\sigma_k(1 + \|B_k\|)^{-1}\}_{k\in\mathcal{U}}$ must be unbounded, where $\mathcal{U}$ is defined in (8). Hence, using Theorem 1, we deduce that there is an index set $\mathcal{U}' \subseteq \mathcal{U}$ such that for all $k \in \mathcal{U}'$ sufficiently large, k is a very successful iteration, i.e., $k \in \mathcal{S}$, which is absurd. □

Consider the following assumption

Model Assumption 5.2. The sequence $\{B_k\}_{k\in\mathbb{N}}$ satisfies:

$$\sum_{k\in\mathbb{N}} \frac{1}{r_k} = +\infty, \qquad r_k := \max_{0\le j\le k} \|B_j\| + 1.$$

The next theorem examines the case where Algorithm 1 generates only a finite number of successful iterations.

Theorem 2. *Let Problem Assumption 4.1 and Model Assumptions 3.1, 3.2, 5.1 and 5.2 be satisfied. If Algorithm 1 generates finitely many successful iterations, then there is $x^* \in \mathbb{R}^n$ such that $x_k = x^*$ for all sufficiently large k, and $\liminf_{k\to\infty} \nu_k^{-\frac{1}{2}} \xi_{\text{cp}}(x_k, \nu_k^{-1})^{\frac{1}{2}} = 0$.*

Proof. Assume, by contradiction, that there is $\alpha > 0$ such that $\nu_k^{-\frac{1}{2}} \xi_{\text{cp}}(x_k, \nu_k^{-1})^{\frac{1}{2}} \ge \alpha$ for all $k \in \mathbb{N}$ and let k_f be the last successful iteration. Hence, $x_k = x_{k_f}$ for all $k \ge k_f$ and $\lim_k x_k = x_{k_f} = x^*$. Using Lemma 8, $\{\sigma_k(1 + \|B_k\|)^{-1}\}_{k\in\mathbb{N}}$ is bounded by a constant $b_\sigma > 0$. This, implies, that for all $k > k_f$,

$$\frac{1}{r_k} = \frac{1}{1 + \max_{0\le j\le k} \|B_j\|} \le \frac{1}{1 + \|B_k\|} \le \frac{b_\sigma}{\sigma_k}.$$

Thus, Model Assumption 5.2 implies

$$\sum_{k=k_f+1}^{\infty} \frac{1}{\sigma_k} = +\infty. \tag{16}$$

On the other hand, all $k > k_f$, k is an unsuccessful iteration. The mechanism of Algorithm 1 then ensures $\frac{\sigma_k}{\sigma_{k+1}} \le \frac{1}{\gamma_1} < 1$ for all $k > k_f$. But this implies that $\sum_{k=k_f+1}^{\infty} \frac{1}{\sigma_k}$ converges, which contradicts (16). □

To the best of our knowledge, Theorem 2 is the first convergence result of a regularized or trust-region method that does not rely on the boundedness of the regularization parameter or trust-region radius in the case of a finite number of successful iterations. In the absence of such boundedness, Theorem 2 does not provide information on the stationarity of x^*. Note that it is not because the algorithms performs a finite number of successful iterations that $s_k = 0$ for all k sufficiently large. For this reason, there is no guarantee that x^* will be first-order stationary for (1) after a finite number of iterations.

Now we consider the case where the number of successful iterations is infinite. Let $\tau \in \mathbb{N}_0$ and k_0 be the index of the first successful iteration, and define, as in [16],

$$\mathcal{T}_k^\tau = \{j = k_0, \dots, k \mid j \le \tau|\mathcal{S}_j|\}, \tag{17a}$$
$$\mathcal{W}_k^\tau = \{j = k_0, \dots, k \mid j > \tau|\mathcal{S}_j|\}. \tag{17b}$$

The next lemma provides a series comparison result that will be used in the proof of the main theorem.

Lemma 9 (16, Lemma 7). *Let $\{r_j\}_{j\in\mathbb{N}}$ be a non-decreasing positive real sequence. For any $k \ge k_0$,*

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

where $\mathcal{T}_k^\tau$ and $\mathcal{W}_k^\tau$ are defined in (17).

The following lemma plays a key role in deriving a convergence result in the case where the number of successful iterations is infinite.

Lemma 10. *Let Problem Assumption 4.1 and Model Assumptions 3.1, 3.2, 5.1 and 5.2 be satisfied. Assume that (i) $\tau \in \mathbb{N}_0$ is chosen so that $\gamma_3\gamma_1^{\tau-1} > 1$, (ii) there is $\alpha > 0$ such that $\nu_k^{-1/2}\xi_{\text{cp}}(x_k, \nu_k^{-1})^{1/2} \geq \alpha$ for all $k \geq k_0$, and (iii) $\{x_k\}_{k\geq k_0}$ is bounded. Then, $\left\{\sum_{j\in\mathcal{W}_k^\tau} \frac{1}{r_j}\right\}_{k\geq k_0}$ is bounded, where r_j is as in Model Assumption 5.2.*

Proof. For any $j \geq 0$, Lemma 8 and the update mechanism of Algorithm 1 imply that there is $b_\sigma > 0$ such that

$$\frac{1}{r_j} \leq \frac{1}{1+\|B_j\|} \leq \frac{b_\sigma}{\sigma_j} \leq \frac{b_\sigma}{\gamma_3^{|S_j|}\gamma_1^{|U_j|}\sigma_0} = \frac{b_\sigma}{\gamma_3^{|S_j|}\gamma_1^{j-|S_j|}\sigma_0}.$$

Consider now $k \geq k_0$ and $j \in \mathcal{W}_k^\tau$. Then, $j > \tau|\mathcal{S}_j|$, which, together with the fact that $\gamma_1 > 1$ and $0 < \gamma_3 \leq 1$ leads to

$$\frac{1}{r_j} \leq \frac{b_\sigma}{\gamma_3^{|S_j|}\gamma_1^{j-|S_j|}\sigma_0} < \frac{b_\sigma}{\gamma_3^{j/\tau}\gamma_1^{j-j/\tau}\sigma_0} = \frac{b_\sigma}{(\gamma_3\gamma_1^{\tau-1})^{j/\tau}\sigma_0}.$$

We sum the above inequalities over $j \in \mathcal{W}_k^\tau$ and use the fact that $\gamma_3\gamma_1^{\tau-1} > 1$ to obtain

$$\sum_{j\in\mathcal{W}_k^\tau} \frac{1}{r_j} < \frac{b_\sigma}{\sigma_0} \sum_{j\in\mathcal{W}_k^\tau} \frac{1}{(\gamma_3\gamma_1^{\tau-1})^{j/\tau}} \leq \frac{b_\sigma}{\sigma_0} \sum_{j\in\mathbb{N}} \frac{1}{(\gamma_3\gamma_1^{\tau-1})^{j/\tau}} < \infty. \qquad \square$$

We state now our main convergence result.

Theorem 3. *Let Problem Assumption 4.1 and Model Assumptions 3.1, 3.2, 5.1 and 5.2 be satisfied. Assume that Algorithm 1 generates infinitely many successful iterations and that there is $(f+h)_{low} \in \mathbb{R}$ such that $(f+h)(x_k) \geq (f+h)_{\text{low}}$ for all $k \in \mathbb{N}$. Then, $\liminf_{k\to+\infty} \nu_k^{-1/2}\xi_{\text{cp}}(x_k, \nu_k^{-1})^{1/2} = 0$.*

Proof. By contradiction, assume that there is $\alpha > 0$ such that for all $k \in \mathbb{N}$, $\nu_k^{-1/2}\xi_{\text{cp}}(x_k, \nu_k^{-1})^{1/2} \geq \alpha$. Lemma 6 shows that $\{x_k\}_{k\in\mathbb{N}}$ is convergent, hence bounded. Lemma 8 then implies that $\{\sigma_k(1+\|B_k\|)^{-1}\}_{k\in\mathbb{N}}$ is bounded, say by $b_\sigma > 0$. Equivalently, $\sigma_k \leq b_\sigma(1+\|B_k\|)$. For any $j \in \mathcal{S}$, $\rho_j \geq \eta_1$, and

$$\begin{aligned}
f(x_j) + h(x_j) - f(x_{j+1}) - h(x_{j+1}) &\geq \eta_1(1-\theta_1)\xi_{\text{cp}}(x_j, \nu_j^{-1}) \\
&\geq \eta_1(1-\theta_1)\nu_j\alpha^2 \\
&= \frac{\eta_1(1-\theta_1)\theta_1\alpha^2}{\sigma_j + \|B_j\|} \\
&\geq \frac{\eta_1(1-\theta_1)\theta_1\alpha^2}{(b_\sigma(1+\|B_j\|) + \|B_j\|)} \\
&\geq \frac{\eta_1(1-\theta_1)\theta_1\alpha^2}{(1+b_\sigma)(1+\|B_j\|)} \\
&\geq \frac{\eta_1(1-\theta_1)\theta_1\alpha^2}{1+b_\sigma}\ \frac{1}{r_j},
\end{aligned}$$

where r_j is defined in Model Assumption 5.2. Let $k \geq k_0$. We sum the above inequalities over all $j \in \mathcal{S}_k$, and obtain

$$f(x_0) + h(x_0) - f(x_{k+1}) - h(x_{k+1}) \geq \frac{\eta_1(1-\theta_1)\theta_1\alpha^2}{1+b_\sigma} \sum_{j\in\mathcal{S}_k} \frac{1}{r_j}.$$

Since $f+h$ is bounded below, it follows that $\sum_{k\in\mathcal{S}} \frac{1}{r_k} < \infty$. Let $\tau \in \mathbb{N}_0$ be chosen so that $\gamma_3\gamma_1^{\tau-1} > 1$. By Lemma 10, $\sum_{j\in\mathcal{W}_k^\tau} \frac{1}{r_j}$ is uniformly bounded for all $k \geq k_0$. However, Lemma 9 yields that for all $k \geq k_0$,

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

which implies that $\sum_{k=k_0}^{\infty} \frac{1}{r_k}$ converges, and contradicts Model Assumption 5.2. □

Note that the assumptions involved in Theorem 3 are weaker compared to existing methods in the literature—see Section 1.

6 Complexity analysis of Algorithm 1

In this section, we study the evaluation complexity of Algorithm 1 in the case where the model Hessians are allowed to be unbounded. We replace Model Assumption 5.1 with the following.

Model Assumption 6.1. There is $\kappa_{\mathrm{m}} > 0$ such that for all $k \in \mathbb{N}$,

$$|(f+h)(x_k+s_k) - (\varphi+\psi)(s_k;x_k)| \leq \kappa_{\mathrm{m}}(1+\|B_k\|)\|s_k\|^2. \tag{18}$$

Note that if ∇f is Lipschitz continuous and $\psi(;x)$ satisfies Model Assumption 5.1 then Model Assumption 6.1 is satisfied as discussed by Leconte and Orban [22].

The next lemma will allow us to show that $\{\sigma_k(1+\|B_k\|^{-1})\}$ is bounded

Lemma 11. *Let Problem Assumption 4.1 and Model Assumptions 3.1, 3.2 and 6.1 be satisfied. Define*

$$b_{\mathrm{succ}} := \frac{2\kappa_{\mathrm{m}}}{1-\eta_2} > 0.$$

If x_k is not first-order stationary and $\sigma_k(1+\max_{0\leq j\leq k}\|B_j\|)^{-1} \geq b_{\mathrm{succ}}$, iteration k is very successful and $\sigma_{k+1} < \sigma_k$.

Proof. Because x_k is not first-order stationary, $s_k \neq 0$. By definition of s_k, $m(0;x_k,\sigma_k) \geq m(s_k;x_k,\sigma_k)$. Hence,

$$\varphi_k(0;x_k) + \psi_k(0;x_k) \geq \varphi_k(s_k;x_k) + \psi_k(s_k;x_k) + \tfrac{1}{2}\sigma_k\|s_k\|^2. \tag{19}$$

Thus, Model Assumption 6.1 yields

$$\begin{aligned}
|\rho_k - 1| &= \left|\frac{(f+h)(x_k+s_k) - (\varphi_k(s_k;x_k)+\psi_k(s_k;x_k))}{\varphi_k(0;x_k)+\psi_k(0;x_k) - (\varphi_k(s_k;x_k)+\psi_k(s_k;x_k))}\right| \\
&\leq \frac{\kappa_{\mathrm{m}}(1+\|B_k\|)\|s_k\|^2}{\frac{1}{2}\sigma_k\|s_k\|^2} \\
&\leq \frac{2\kappa_{\mathrm{m}}}{\sigma_k(1+\max_{0\leq j\leq k}\|B_j\|)^{-1}} \leq \frac{2\kappa_{\mathrm{m}}}{b_{\mathrm{succ}}} = 1-\eta_2.
\end{aligned}$$

Thus, we obtain $\rho_k \geq \eta_2$, meaning that the iteration k is very successful. □

The next theorem shows that $\{\sigma_k(1+\|B_k\|^{-1})\}$ is bounded.

Theorem 4. *Let Problem Assumption 4.1 and Model Assumptions 3.1, 3.2 and 6.1 be satisfied. For all $k \in \mathbb{N}$, if x_k is not stationary,*

$$\sigma_k(1+\max_{0\leq j\leq k}\|B_j\|)^{-1} \leq b_{\max} := \min\{\sigma_0(1+\|B_0\|)^{-1}, \gamma_2 b_{\mathrm{succ}}\} > 0.$$

Proof. Set $b_k := \sigma_k(1 + \max_{0\leq j\leq k} \|B_j\|)^{-1}$ for all k. We proceed by induction. For $k = 0$, $\sigma_0(1 + \|B_0\|)^{-1} \leq b_{\max}$ by definition. Assume that $b_k \leq b_{\max}$ for $k \geq 0$.

Assume first that $b_k < b_{\text{succ}}$. Because $\{(1 + \max_{0\leq j\leq k} \|B_j\|)^{-1}\}$ is non-increasing, the update of σ_k in Algorithm 1 ensures that

$$b_{k+1} = (1 + \max_{0\leq j\leq k+1} \|B_j\|)^{-1}\sigma_{k+1} \leq (1 + \max_{0\leq j\leq k} \|B_j\|)^{-1}\gamma_2\sigma_k = \gamma_2 b_k < \gamma_2 b_{\text{succ}} \leq b_{\max}.$$

Now, assume conversely that $b_k \geq b_{\text{succ}}$. Lemma 11 implies that iteration k is very successful, and $\sigma_{k+1} < \sigma_k$. Thus,

$$b_{k+1} = (1 + \max_{0\leq j\leq k+1} \|B_j\|)^{-1}\sigma_{k+1} < (1 + \max_{0\leq j\leq k} \|B_j\|)^{-1}\sigma_k = b_k \leq b_{\max}. \qquad \square$$

Additionally, instead of Model Assumption 5.2, we assume that model Hessians grow at most linearly with $|\mathcal{S}_k|$, which covers multiple quasi-Newton approximations—see Section 1.

Model Assumption 6.2. There are $\mu > 0$ and $0 \leq p \leq 1$ such that, for all $k \in \mathcal{S}$,

$$\max_{0\leq j\leq k} \|B_j\| \leq \mu(1 + |\mathcal{S}_k|^p). \tag{20}$$

Because $|\mathcal{S}_k|$ is non-decreasing with k, (20) is equivalent to $\|B_k\| \leq \mu(1 + |\mathcal{S}_k|^p)$ for all $k \in \mathbb{N}$. The following theorem considers the case with a finite number of successful iterations. The proof follows [4, Theorem 4.2] and is recalled here for completeness.

Theorem 5. *Let Problem Assumption 4.1 and Model Assumptions 3.1, 3.2, 6.1 and 6.2 be satisfied. If Algorithm 1 generates finitely many successful iterations, then $x_k = x^*$ for all sufficiently large k where x^* is a stationary point.*

Proof. Assume by contradiction that x^* is not a stationary point. Because the number of successful iterations is finite, according to Model Assumption 6.2, there is $k_f \in \mathbb{N}$ such that $\|B_k\| \leq \mu(1 + |\mathcal{S}_{k_f}|^p)$ for all $k \geq k_f$, where k_f is the index of the last successful iteration. The mechanism of Algorithm 1 ensures that σ_k increases on unsuccessful iterations. Hence, there must exist an unsuccessful iteration $k > k_f$ such that $\sigma_k \geq b_{\text{succ}}(1 + \mu(1 + |\mathcal{S}_{k_f}|^p)) \geq b_{\text{succ}}(1 + \|B_k\|)$, with b_{succ} defined in Lemma 11. Because x^* is not stationary, we can apply Lemma 11, which shows that k is very successful, and contradicts our assumption. $\square$

We know from Theorem 3 that $\liminf_{k\to+\infty} \nu_k^{-1/2}\xi_{\text{cp}}(x_k, \nu_k^{-1})^{1/2} = 0$ when Algorithm 1 generates infinitely many successful iterations. Let $\epsilon > 0$ and k_ϵ be the first iteration of Algorithm 1 such that $\nu_k^{-1/2}\xi_{\text{cp}}(x_k, \nu_k^{-1})^{1/2} \leq \epsilon$. Define

$$\mathcal{S}(\epsilon) := \mathcal{S}_{k_\epsilon - 1} = \{k \in \mathcal{S} \mid k < k_\epsilon\}, \tag{21a}$$
$$\mathcal{U}(\epsilon) := \mathcal{U}_{k_\epsilon - 1} = \{k \in \mathbb{N} \mid k \notin \mathcal{S} \text{ and } k < k_\epsilon\}. \tag{21b}$$

The next theorems bound k_ϵ. The proofs are similar to [16, Theorem 2].

Theorem 6. *Let Problem Assumption 4.1 and Model Assumptions 3.1, 3.2, 6.1 and 6.2 be satisfied. Assume that Algorithm 1 generates infinitely many successful iterations and that there is $(f + h)_{\text{low}} \in \mathbb{R}$ such that $(f + h)(x_k) \geq (f + h)_{\text{low}}$ for all $k \in \mathbb{N}$. If $0 \leq p < 1$,*

$$|\mathcal{S}(\epsilon)| \leq ((1 - p)\kappa_1\epsilon^{-2} + 1)^{1/(1-p)} - 1 = O(\epsilon^{-2/(1-p)}), \tag{22}$$

where

$$\kappa_1 = \frac{((f + h)(x_0) - (f + h)_{\text{low}})(b_{\max} + 2\mu(1 + b_{\max}))}{\eta_1\theta_1(1 - \theta_1)},$$

and $b_{\max}$ *is as in Theorem 4. If* $p = 1$,

$$|\mathcal{S}(\epsilon)| \leq \exp(\kappa_1 \epsilon^{-2}) - 1. \tag{23}$$

Proof. Let $k \in \mathcal{S}(\epsilon)$, then $\nu_k^{-1/2} \xi_{\text{cp}}(x_k, \nu_k^{-1})^{1/2} \geq \epsilon$ and

$$(f+h)(x_k) - (f+h)(x_k + s_k) \geq \eta_1(1-\theta_1)\xi_{\text{cp}}(x_k; \nu_k^{-1}) \geq \eta_1(1-\theta_1)\nu_k\epsilon^2. \tag{24}$$

Theorem 4 implies

$$\begin{aligned}\nu_k = \frac{\theta_1}{\|B_k\| + \sigma_k} &\geq \frac{\theta_1}{\max_{0\leq j\leq k}\|B_j\| + b_{\max}(1 + \max_{0\leq j\leq k}\|B_j\|)} \\ &= \frac{\theta_1}{b_{\max} + (1+b_{\max})\max_{0\leq j\leq k}\|B_j\|}.\end{aligned}$$

Model Assumption 6.2 then implies

$$\nu_k \geq \frac{\theta_1}{b_{\max} + \mu(1+b_{\max})(1+|\mathcal{S}_k|^p)} = \frac{\theta_1}{|\mathcal{S}_k|^p}\zeta\left(|\mathcal{S}_k|^p\right), \tag{25}$$

where $\zeta : \mathbb{R}_+ \to \mathbb{R}$, $\zeta(x) := x/(b_{\max} + \mu(1+b_{\max})(x+1))$.

Because ζ is non-decreasing and $|\mathcal{S}_k| \geq 1$ (as we have infinitely many successful iterations), $\zeta(|\mathcal{S}_k|^p) \geq \zeta(1) = (1 + 2\mu(1+b_{\max}))^{-1}$. Thus, (25) becomes

$$\nu_k \geq \frac{\theta_1}{b_{\max} + 2\mu(1+b_{\max})}\frac{1}{|\mathcal{S}_k|^p},$$

which combines with (24) to yield

$$(f+h)(x_k) - (f+h)(x_k+s_k) \geq \frac{\eta_1\theta_1(1-\theta_1)\epsilon^2}{b_{\max} + 2\mu(1+b_{\max})}\frac{1}{|\mathcal{S}_k|^p} := C\frac{1}{|\mathcal{S}_k|^p}. \tag{26}$$

We sum over all $k \in \mathcal{S}(\epsilon)$, and obtain

$$(f+h)(x_0) - (f+h)_{\text{low}} \geq C\sum_{k\in\mathcal{S}(\epsilon)}\frac{1}{|\mathcal{S}_k|^p} = C\sum_{k=0}^{|\mathcal{S}(\epsilon)|-1}\frac{1}{|\mathcal{S}_{\phi(k)}|^p},$$

where ϕ is an increasing map from $\{0, \ldots, |\mathcal{S}(\epsilon)|-1\}$ to $\mathcal{S}(\epsilon)$. Thus, by definition of ϕ and $\mathcal{S}_{\phi(k)}$, $|\mathcal{S}_{\phi(k+1)}| = |\mathcal{S}_{\phi(k)}| + 1$ and $|\mathcal{S}_{\phi(0)}| = 1$. In other words, $|\mathcal{S}_{\phi(k)}| = k+1$, and

$$(f+h)(x_0) - (f+h)_{\text{low}} \geq C\sum_{k=0}^{|\mathcal{S}(\epsilon)|-1}\frac{1}{(k+1)^p} = C\sum_{k=1}^{|\mathcal{S}(\epsilon)|}\frac{1}{k^p}.$$

Because $\int_k^{k+1}\frac{1}{t^p}\mathrm{d}t \leq \int_k^{k+1}\frac{1}{k^p}\mathrm{d}t = \frac{1}{k^p}$,

$$(f+h)(x_0) - (f+h)_{\text{low}} \geq C\sum_{k=1}^{|\mathcal{S}(\epsilon)|}\int_k^{k+1}\frac{1}{t^p}\mathrm{d}t = C\int_1^{|\mathcal{S}(\epsilon)|+1}\frac{1}{t^p}\mathrm{d}t. \tag{27}$$

There are two cases to consider:

- if $0 \leq p < 1$, $(f+h)(x_0) - (f+h)_{\text{low}} \geq C\frac{(|\mathcal{S}(\epsilon)|+1)^{1-p}-1}{1-p}$, which is (22);
- if $p = 1$, $(f+h)(x_0) - (f+h)_{\text{low}} \geq C\log(|\mathcal{S}(\epsilon)|+1)$, which is (23). □

Finally, we derive a bound on the cardinality of $\mathcal{U}(\epsilon)$.

Theorem 7. *Let Problem Assumption 4.1 and Model Assumptions 3.1, 3.2 and 6.1 hold. Assume that Algorithm 1 generates infinitely many successful iterations. Then*

$$|\mathcal{U}(\epsilon)| \leq |\log_{\gamma_1}(\gamma_3)||\mathcal{S}(\epsilon)| + \log_{\gamma_1}(1 + \mu(1 + |\mathcal{S}(\epsilon)|^p)) + \frac{\log(b_{\max}/\sigma_0)}{\log(\gamma_1)}, \tag{28}$$

where μ and p are defined in Model Assumption 5.2, $b_{\max}$ as in Theorem 4, and $|\mathcal{S}(\epsilon)|$ is as in Theorem 6.

Proof. The mechanism of Algorithm 1 guarantees that for all $k \in \mathbb{N}$, $|\mathcal{U}_k| \leq |\log_{\gamma_1}(\gamma_3)|\,|\mathcal{S}_k| + \log_{\gamma_1}(\sigma_k/\sigma_0)$. Hence, Theorem 4 yields

$$\begin{aligned} |\mathcal{U}(\epsilon)| &\leq |\log_{\gamma_1}(\gamma_3)||\mathcal{S}(\epsilon)| + \log_{\gamma_1}\left(\frac{b_{\max}(1 + \max_{0\leq j\leq k_\epsilon - 1}\|B_j\|)}{\sigma_0}\right) \\ &\leq |\log_{\gamma_1}(\gamma_3)||\mathcal{S}(\epsilon)| + \log_{\gamma_1}(1 + \mu(1 + |\mathcal{S}(\epsilon)|^p)) + \log_{\gamma_1}\left(\frac{b_{\max}}{\sigma_0}\right). \end{aligned}$$

□

The complexity bound in Theorem 6 is of the same order as that of [4, Lemma 4.3] for trust-region methods when $p = 0$ in Model Assumption 6.2, which corresponds to bounded model Hessians. Unlike [3, Lemma 3.6], the constant θ_2, as defined in the switch on Algorithm 1 of Algorithm 1, does not appear in our complexity bound. Thus, large values of θ_2 in Algorithm 1 will not worsen the complexity bound. In the general case where $p > 0$, our bound is better than that in [22, Theorem 4.2], as their step computation rule makes the bound dependent on θ_2. As p approaches 1, the bound in [22, Theorem 4.2] goes to infinity, whereas ours, though exponential, remains finite, as in [16]. Finally, the same example as in [16, §3.1] shows that our complexity bounds are also tight.

7 Algorithmic refinements

We describe a special case of Algorithm 1 and an extension for which the convergence theory continues to hold, that we exploit in the numerical experiments of Section 8, and that prove to be efficient in practice. As both refinements have already been studied by Leconte and Orban [21] in the context of their trust-region method, we keep our description to a minimum.

7.1 Special case: diagonal model Hessians

If we select B_k to be diagonal in Algorithm 1, a specialized implementation emerges whenever h is separable and $\psi(\cdot; x_k)$ is chosen to be separable at each iteration. For a number of choices of separable h that are of interest in applications, the step s_k may be computed analytically without requiring an iterative subproblem solver. We refer to this implementation as R2DH, where "DH" stands to *diagonal Hessians*. This section is modeled after [21, Section 4], to which we refer the reader for further information.

Diagonal quasi-Newton methods originate from [14, 18, 27]. In order for a variational problem to possess a solution that defines a diagonal update, the classic secant equation is replaced with the *weak* secant equation $s_k^T B_{k+1} s_k = s_k^T y_k$, where $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$. A handful of diagonal updates have been proposed in the literature. The most efficient is probably the *spectral* update $B_{k+1} = \tau_{k+1} I$, where $\tau_{k+1} := s_k^T y_k / s_k^T s_k$ is defined as in the spectral gradient method [7]. Because B_k is a multiple of the identity, h and $\psi(\cdot; x_k)$ need not be separable as the computation of s_k boils down to the evaluation of a proximal operator with step length $1/\sqrt{\tau_k + \sigma_k}$—see Definition 2. Zhu et al. [34] derive an update akin to the well-known PSB formula that may be indefinite. We refer to it below as *PSB*. Andrei [1] derives an update based on a different variational problem that may also be indefinite. We refer to it below as *Andrei*. Additionally, we include a new diagonal variant inspired from the BFGS formula

using a diagonal update. The main idea comes from applying [11, Lemma 5.1] to the last term of the BFGS update, i.e., $y_k y_k^T / s_k^T y_k$, to obtain the diagonal update

$$D_{k+1} = \frac{\sum_{i=1}^n |(y_k)i|}{s_k^T y_k} \operatorname{diag}(|y_k|).$$

This update remains positive as long as $s_k^T y_k > 0$. We refer to this variant below as *DBFGS*. Although DBFGS does not always satisfy the secant equation, our numerical results demonstrate its competitiveness against other state-of-the-art diagonal-based methods. Note that the three updates (i.e., PSB, Andrei and DBFGS) generate B_k that is not a multiple of the identity, and hence h and $\psi(\cdot; x_k)$ should be separable.

R2DH may act as standalone solver for (1) or as subproblem solver to compute s_k in Algorithm 1. Our results in Section 8 illustrate that, in both use cases, R2DH typically outperforms R2 [2, 3, Algorithm 6.1].

7.2 Non-monotone variants

Inspired by the success of the non-monotone spectral gradient method [7], Leconte and Orban [21, Section 6] explain how to modify an algorithm similar to Algorithm 1 to incorporate a non-monotone strategy.

Let $q \in \mathbb{N}$ be a given *memory parameter*. Define $q_k = 1$ if $q = 0$ and $q_k := \min(k, q)$ if $q > 0$, Define also $\mathcal{S}_{q_k}^+$ the set of the q_k *most recent* successful iterations. By convention, we set $\mathcal{S}_0^+ = \{0\}$. An iteration k now considers the objective value at each iteration in $\mathcal{S}_{q_k}^+$. Define

$$(f+h)_{\max,k} := \max\{(f+h)(x_j) \mid j \in \mathcal{S}_{q_k}^+\}. \tag{29}$$

Algorithm 1 corresponds to $q = 0$. The non-monotone strategy consists in enforcing decrease with respect to $(f+h)_{\max,k}$ instead of $(f+h)(x_k)$. In other words, we redefine

$$\rho_k := \frac{(f+h)_{\max,k} - (f+h)(x_k + s_k)}{(f+h)_{\max,k} - (\varphi+\psi)(s_k; x_k)}.$$

As in [21, Section 6], the new expression of ρ_k does not interfere with convergence properties or complexity bounds, except that it changes the constants in the latter.

8 Numerical experiments

Our implementation of Algorithm 1 and all solvers used in the experiments are available in the RegularizedOptimization Julia module [5]. By default, R2N uses an L-BFGS approximation with memory 5, as implemented in the LinearOperators Julia module [23], and uses parameters $\theta_1 = (1+\varepsilon_M^{1/5})^{-1} \approx 0.999$, $\theta_2 = 1/\varepsilon_M \approx 10^{15}$, $\eta_1 = \varepsilon_M^{1/4} \approx 10^{-4}$, $\eta_2 = 0.9$, and $\sigma_0 = \varepsilon_M^{1/3} \approx 10^{-6}$, where ε_M is the machine epsilon. The reason for defining values based on ε_M is that our code may be run in various floating-point arithmetics. Here, however, all tests are run in double precision. If iteration k of Algorithm 1 is very successful, $\sigma_{k+1} = \sigma_k/3$; if iteration k is unsuccessful, $\sigma_{k+1} = 3\sigma_k$. Otherwise, $\sigma_{k+1} = \sigma_k$.

R2N stops as soon as

$$\nu_k^{-1/2} \xi_{cp}(x_k, \nu_k)^{1/2} < \epsilon_a + \epsilon_r \nu_0^{-1/2} \xi_{cp}(x_0, \nu_0)^{1/2}, \tag{30}$$

where $\epsilon_a = \epsilon_r = \varepsilon_M^{3/10} \approx 10^{-5}$ are an absolute and relative tolerance, respectively, or it exceeds the budget of $1,000$ iterations or $3,600$ seconds of CPU time. To solve the subproblem in Line 7 of

Algorithm 1, we use either R2 [3, Algorithm 6.1], or one of several R2DH variants (Spec, PSB, Andrei, or DBFGS) as described in Section 7, as well as the non-monotone spectral R2DH (R2DH-Spec-NM) with memory 5. R2 initializes $\nu_0 = 1.0$. R2N and R2DH initialize ν_0 according to Line 5 of Algorithm 1. The subproblem solvers terminate as soon as

$$\hat{\nu}_k^{-1/2}\hat{\xi}_{cp}\left(x_k + s, \hat{\nu}_k\right)^{1/2} \leq \begin{cases} 10^{-3} & \text{if } k = 0, \\ \min\left(\left(\nu_k^{-1}\xi_{cp}\right)^{3/2}, 10^{-3}\left(\nu_k^{-1}\xi_{cp}\right)^{1/2}\right) & \text{if } k > 0, \end{cases}$$

where $\xi_{cp} = \xi_{cp}\left(x_k, \nu_k\right)$, $\hat{\nu}_k$ and $\hat{\xi}_{cp}$ are the step size and first-order stationarity measure related to the subproblem solver. Note that R2 and all the R2DH variants can also be used to solve (1) directly. All quasi-Newton approximations are initialized to the identity. In all experiments, we use $\psi(s; x) := h(x + s)$.

Our objective is to minimize the number of objective and gradient evaluations, as they are generally expensive to compute, while assuming that the proximal operators of common regularizers such that ℓ_0 and ℓ_1 norms are comparatively cheap to evaluate. We include also other test problems with the nuclear norm and the rank regularizers.

In our figures, we set $(f + h)^*$ to the best value found by all the solvers. We plot $\Delta(f + h)(x_k) = (f + h)(x_k) - (f + h)^*$ against the iterations to illustrate progress towards that best value. We also report the following solver statistics in tables: the final value of f at convergence; the final h/λ, where λ is a weight on the regularizer h; the final stationarity measure $\sqrt{\xi/\nu}$; the number of evaluations of the smooth objective ($\#f$); the number of evaluations of the gradient ($\#\nabla f$); the number of proximal operator evaluations (#prox); and the elapsed time t in seconds.

8.1 Basis pursuit denoise (BPDN)

The first set of experiments focuses on the basis pursuit denoise problem as described in [3], which is common in statistical and compressed sensing applications. The goal is to recover a sparse signal $x_{\text{true}} \in \mathbb{R}^n$ from noisy observed data $b \in \mathbb{R}^m$. This problem can be formulated as

$$\underset{x}{\text{minimize}} \;\; \tfrac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_0, \tag{31}$$

where A is $m \times n$ and randomly generated with orthonormal rows. We set $m = 2{,}000$, $n = 5{,}120$, and $b := Ax_{\text{true}} + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, 0.01)$. The true signal x_{true} is a vector of zeros, except for 100 of its components. We set $\lambda = 0.1\|A^T b\|_\infty$. All algorithms start from the same randomly generated, hence non-sparse, x_0. For this problem, we compare R2 with R2DH variants (Spec, PSB, Andrei, and DBFGS).

Figure 1 shows that all solvers reach similar accuracy, except for the PSB and Andrei variants. R2DH-Spec-NM displays the best performance, followed by the R2DH-Spec and closely by R2DH-DBFGS variants, although it requires more evaluations to achieve stationarity. Table 1 shows that all R2DH variants surpass R2 in all measures, except for R2DH-PSB and R2DH-Andrei, which require more evaluations and seem to converge to a non stationary point. R2 and all other R2DH variants identify a similarly-sparse solution. R2DH-Andrei requires significantly more evaluations and time than other R2DH variants and hits the iteration limit before (30) is triggered. Note that R2DH-DBFGS requires fewer evaluations than R2 both in gradient and function evaluations but struggles to compete with R2DH-Spec-NM and R2DH-Spec. R2DH-Spec-NM is more efficient than R2DH-Spec, it avoids the unsuccessful iterations that R2DH-Spec falls into. Given the strong performance of R2DH-Spec-NM, we set it as the default R2N subsolver in the following experiments.

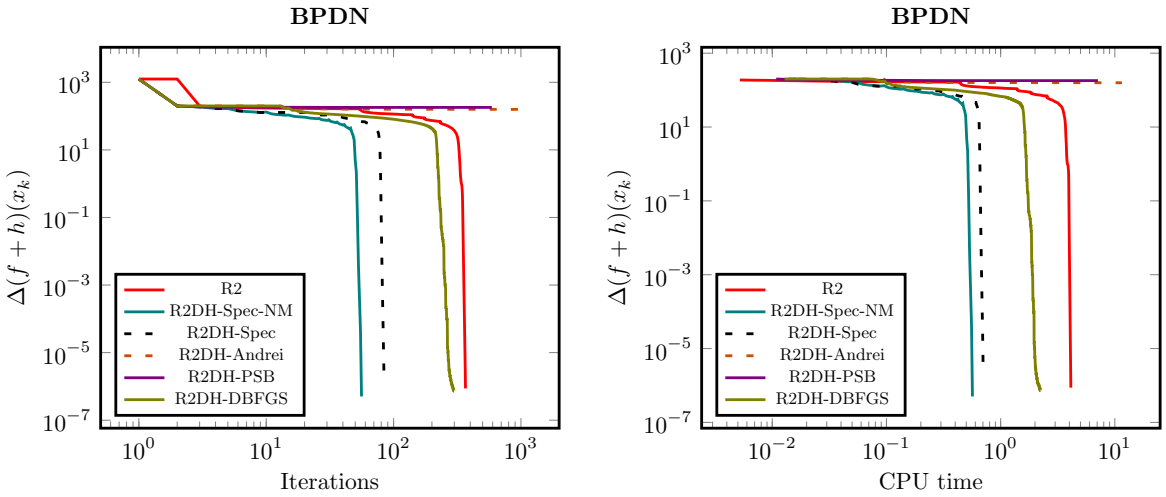

Figure 1: BPDN objective vs. iterations (left) and CPU time (right).

Table 1: Comparison of different solvers on the BPDN problem.

Solver	f	h/λ	$\Delta(f+h)$	$\sqrt{\xi/\nu}$	$\#f$	$\#\nabla f$	$\#prox$	$t(s)$
R2	9.22e−02	100	8.74e−07	8.0e−04	366	362	366	7.37
R2DH-Spec-NM	9.22e−02	100	5.05e−07	5.9e−04	57	56	56	0.89
R2DH-Spec	9.22e−02	100	0.00e+00	3.6e−04	86	57	85	0.97
R2DH-Andrei	3.42e+00	2926	1.58e+02	8.6e−01	1001	988	1991	20.61
R2DH-PSB	3.69e−06	3400	1.81e+02	6.7e−04	592	591	1194	11.85
R2DH-DBFGS	9.22e−02	100	7.26e−07	6.4e−04	300	153	299	5.53

8.2 Matrix completion

We address the matrix completion problem from [33] with rank nuclear norm regularizers to recover a low-rank matrix from noisy observations. The problem is formulated as

$$\underset{X}{\text{minimize}} \ \tfrac{1}{2}\|P_\Omega(X-M)\|_F^2 + \lambda h(X), \tag{32}$$

where $X \in \mathbb{R}^{n\times n}$ and $n = 120$. Here, $\lambda = 10^{-1}$ is a weight, and $h(X)$ is either $\text{rank}(X)$ or $\|X\|_*$, M is formed by applying a standard two-component Gaussian mixture model (GMM) to a low-rank matrix X_r. Specifically, M is computed as:

$$M = (1-c)(X_r + \mathcal{N}(0,\sigma_A^2)) + c(X_r + \mathcal{N}(0,\sigma_B^2)),$$

where $\mathcal{N}(0,\sigma_A^2)$ represents the noise component with variance σ_A^2, and $\mathcal{N}(0,\sigma_B^2)$ represents the influence of outliers with a larger variance σ_B^2. The parameter c controls the relative proportion of noise and outliers in the observed matrix M. Finally, P_Ω is a linear operator that extracts entries $(i,j) \in \Omega$ and sets unobserved entries to zero.

For all solvers, we select a random initial matrix and set the rank of X_r to 40. Given that the smooth part of (32) is a linear least-squares residual, we apply the Levenberg-Marquardt (LM) algorithm from Aravkin et al. [4, Algorithm 4.1], which is a specific instance of R2N with $B_k = J_k^T J_k$, where J_k is the Jacobian of the least-squares residual at iteration k. Notably, R2DH can serve as a subproblem solver within LM—this combination is referred to as LM-R2DH, in contrast to the default LM-R2. We compare the performance of R2, R2DH, LM-R2, and LM-R2DH in Figure 2 and Tables 2 and 3. In Tables 2 and 3, the column $\#\nabla f$ is replaced by the number of Jacobian or adjoint products $\#J$.

Figure 2 shows that R2DH stands out in terms of final objective value for the rank regularizer, while for the nuclear norm regularizer, all solvers achieve similar accuracy. For the rank regularizer,

Table 2 shows that, while R2DH requires more objective evaluations than either LM variant, it performs significantly fewer Jacobian-vector products and provides the solution with the best objective value, and, in particular, the lowest-rank solution. Variants of LM behave almost identically according to Figure 2 and require the fewest objective evaluations, although they demand many Jacobian-vector products, as seen in Tables 2 and 3. LM-R2DH requires fewer gradient evaluations and proximal operator evaluations than LM-R2.

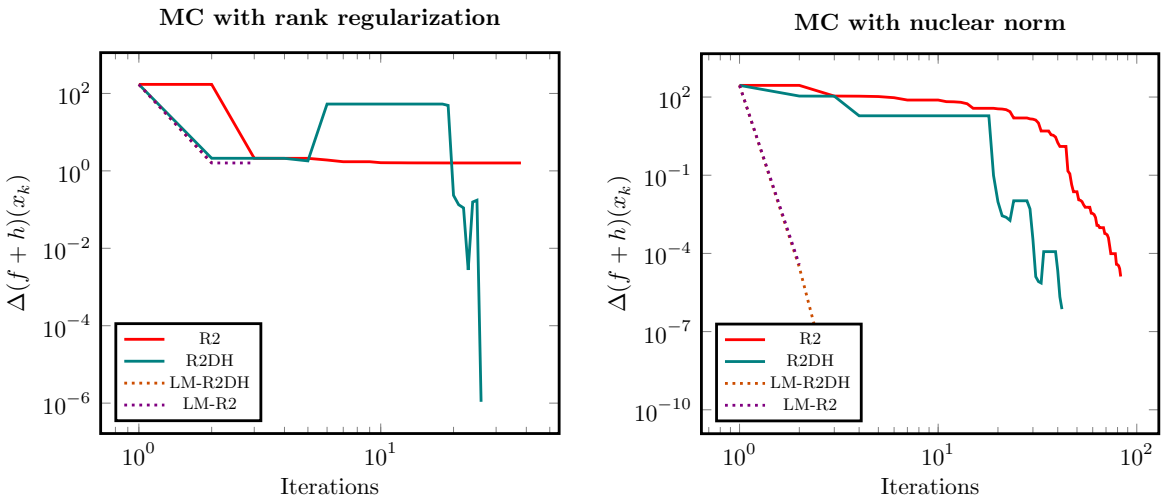

Figure 2: Objectives vs. iterations for MC with rank (left) and nuclear norm (right) regularizers.

Table 2: Comparison of different solvers for matrix completion problem with rank regularizer.

Solver	f	h/λ	$\Delta(f+h)$	$\sqrt{\xi/\nu}$	$\#f$	$\#J$	$\#prox$	$t(s)$
R2	2.78e−07	111	1.60e+00	2.3e−04	37	31	37	0.20
R2DH	1.34e−11	95	0.00e+00	3.7e−06	28	15	27	0.17
LM-R2DH	2.16e−08	111	1.60e+00	1.7e−04	2	115	48	0.29
LM-R2	1.67e−12	111	1.60e+00	6.4e−07	3	183	61	0.33

Table 3: Comparison of different solvers for matrix completion problem with nuclear norm regularizer.

Solver	f	h/λ	$\Delta(f+h)$	$\sqrt{\xi/\nu}$	$\#f$	$\#J$	$\#prox$	$t(s)$
R2	1.00e−02	7.5e+00	1.30e−05	3.6e−04	82	52	82	0.43
R2DH	1.00e−02	7.5e+00	7.28e−07	2.3e−04	43	19	42	0.21
LM-R2DH	1.00e−02	7.5e+00	1.98e−10	3.1e−06	3	246	108	0.63
LM-R2	1.00e−02	7.5e+00	0.00e+00	2.0e−06	3	340	144	0.86

8.3 General regularized problems

In this section, we illustrate the performance of R2N on two test problems. The first problem addresses an image recognition task using a support vector machine (SVM) similar to those in [3]. The objective is to use this nonlinear SVM to classify digits from the MNIST dataset, specifically distinguishing between "1" and "7", while excluding all other digits. A sparse support is imposed using an ℓ_0 regularizer. The optimization problem is given by

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \ \tfrac{1}{2}\|\mathbf{1} - \tanh(b \odot \langle A, x\rangle)\|^2 + \lambda\|x\|_0,$$

where $\lambda = 10^{-1}$ and $A \in \mathbb{R}^{m\times n}$, with $n = 784$ representing the vectorized size of each image. The dataset includes $m = 13{,}007$ images for training and $m = 2{,}163$ images for testing. Here, $\odot$ denotes the elementwise product between vectors, and $\mathbf{1} = (1, \ldots, 1)$.

The second problem is from [12, 30] and arises in image denoising and deblurring applications. The related optimization problem is given by

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \sum_{i=1}^{n} \log\left((Ax-b)_i^2+1\right) + \lambda\|x\|_1,$$

where $\lambda = 10^{-4}$ and $A \in \mathbb{R}^{n\times n}$ with $n = 256^2$ is a Gaussian blur operator. The term b denotes the blurred image with added Gaussian noise. In our test, b is the blurred version of the cameraman image x^* with added Gaussian noise, i.e., $b = Ax^* + \text{noise}$. The smooth part related to the two optimization problems is neither quadratic nor linear least squares, but a general non-convex problem.

We compare the performance of four methods: R2, R2DH, R2N-R2 (R2N with R2 as a subsolver), and R2N-R2DH (R2N with R2DH as a subsolver).

As shown in Tables 4 and 5, for both problems, the R2N variants outperform R2 and R2DH in terms of objective and gradient evaluations, though they require more proximal operator evaluations. Both R2N-R2DH and R2N-R2 have comparable performance and reach very good solutions compared to the other methods.

Note that for the non-linear SVM problem, as indicated in Figure 3 and Table 4, although R2DH reduces the objective function the most, it requires a higher number of evaluations of f and ∇f than both R2N-R2DH and R2N-R2.

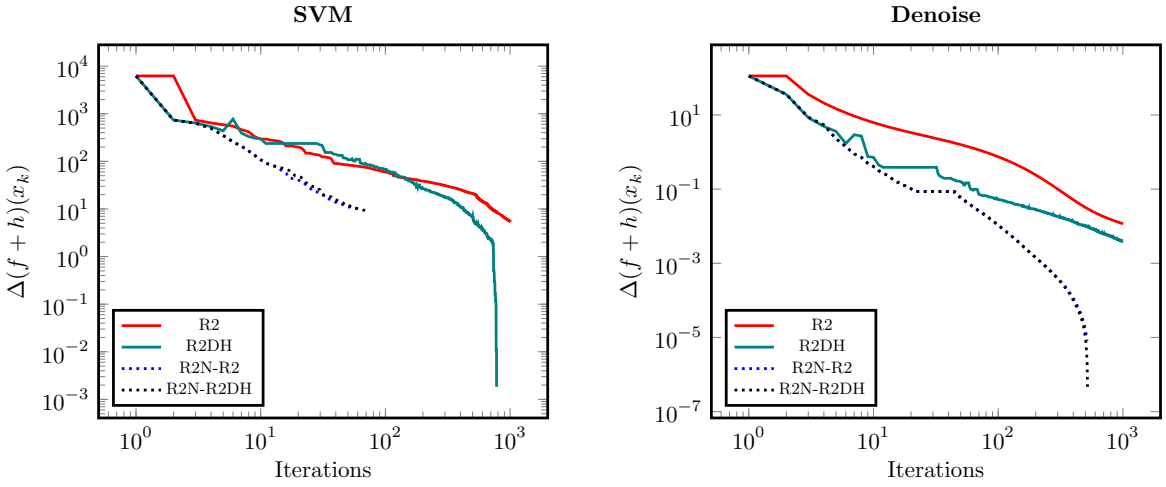

Figure 3: Plots of the objective vs. iterations related to SVM (left) and denoise (right).

Table 4: Comparison of different solvers on the nonlinear SVM problem.

Solver	f	h/λ	$\Delta(f+h)$	$\sqrt{\xi/\nu}$	$\#f$	$\#\nabla f$	$\#prox$	$t(s)$
R2	1.94e+01	175	5.33e+00	1.6e−02	1002	775	1002	9.47
R2DH	1.59e+01	157	0.00e+00	1.8e−03	781	441	780	4.97
R2N-R2	1.85e+01	233	1.02e+01	2.3e−03	59	59	10404	1.21
R2N-R2DH	1.67e+01	237	8.84e+00	2.5e−03	71	71	11284	1.29

For the denoising problem, the R2N variants outperform the R2 and R2DH variants. R2N-R2, where R2 is used as a subsolver, is the best among all the tested methods, requiring the fewest objective and gradient evaluations, followed closely by R2N-R2DH, which provides a more accurate solution, as detailed in Table 5.

Table 5: Comparison of different solvers on the denoising problem.

Solver	f	h/λ	$\Delta(f+h)$	$\sqrt{\xi/\nu}$	$\#f$	$\#\nabla f$	$\#prox$	$t(s)$
R2	6.05e−02	3.7e+03	1.16e−02	2.7e−02	1002	1002	1002	12.02
R2DH	5.92e−02	3.6e+03	3.92e−03	8.4e−03	1001	578	1000	8.85
R2N-R2	5.88e−02	3.6e+03	8.13e−06	3.4e−04	515	494	97093	316.44
R2N-R2DH	5.88e−02	3.6e+03	0.00e+00	3.2e−04	521	499	97411	314.27

9 Discussion

We proposed method R2N, a modified quasi-Newton method for nonsmooth regularized problems. R2N generalizes both R2 [3] and LM [4] and enjoys convergence properties without assuming Lipschitz continuity of ∇f or boundedness of the model Hessians. Inspired by [16], who work on trust-region methods for smooth optimization, we propose a complexity analysis of R2N to handle potentially unbounded model Hessians. 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 by bounding the model Hessian growth with a power of the number of successful iterations—a reasonable bound as, in practice, it is uncommon to update quasi-Newton approximations on unsuccessful iterations. Nevertheless, Diouane et al. [16] show that similar complexity bounds continue to hold when the model Hessians are bounded by a power of the number of iterations, and not just the number of successful iterations. Because their analysis uses similar arguments, their complexity bounds continue to hold for R2N.

Numerical illustrations show the strong potential of our implementation of R2N and some of its variants, both as a main solver and as a subproblem solver. In particular, diagonal variants are competitive with, and often outperform, R2 when used as a subsolver inside R2N. One of the main advantages of R2N in practice is that proximal operators are easier to compute than in TR [3]. We illustrated that advantage by solving rank and nuclear norm-regularized problems. One way to further enhance the performance of R2N is to use a more efficient subproblem solver, e.g., such as that in [6], or, in certain cases, by solving the subproblem exactly as in [15].

R2N convergence analysis arguments can be used to update and strengthen the existing convergence analysis of methods R2, TR, TRDH [21], LM and LMTR [4]. In follow-up research, we aim to identify the nature of limit points under our assumptions.

References

[1] N. Andrei. A diagonal quasi-Newton updating method for unconstrained optimization. Numer. Algor., 81: 575–590, 2019.

[2] A. Aravkin, R. Baraldi, G. Leconte, and D. Orban. Corrigendum: A proximal quasi-Newton trust-region method for nonsmooth regularized optimization. Cahier G–2021–12–SM, GERAD, Montréal, QC, Canada, 2024.

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

[4] A. Y. Aravkin, R. Baraldi, and D. Orban. A Levenberg–Marquardt method for nonsmooth regularized least squares. SIAM J. Sci. Comput., 46(4):A2557–A2581, 2024.

[5] R. Baraldi, G. Leconte, and D. Orban. RegularizedOptimization.jl: Algorithms for regularized optimization, 2024.

[6] S. Becker and J. Fadili. A quasi-Newton proximal splitting method. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, Advances in Neural Information Processing Systems, volume 25. Curran Associates, Inc., 2012.

[7] E. G. Birgin, J. M. Martínez, and M. Raydan. Spectral projected gradient methods: Review and perspectives. J. Stat. Softw., 60(3):1–21, 2014.

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

[9] R. I. Boţ, E. R. Csetnek, and S. László. An inertial forward–backward algorithm for the minimization of the sum of two nonconvex functions. EURO J. Comput. Optim., (4):3–25, 2016.

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

[11] E. Chouzenoux, J.-C. Pesquet, and A. Repetti. Variable metric forward–backward algorithm for minimizing the sum of a differentiable function and a convex function. J. Optimiz. Theory App., 162(1):107–132, 2014.

[12] E. Chouzenoux, S. Martin, and J.-C. Pesquet. A local MM subspace method for solving constrained variational problems in image recovery. J. Math. Imaging Vis., 65(2):253–276, 2023.

[13] A. R. Conn, N. I. Gould, and Ph. L. Toint. Trust-region methods. Number 1 in MOS-SIAM Series on Optimization. SIAM, Philadelphia, USA, 2000.

[14] J. E. Dennis, Jr. and H. Wolkowicz. Sizing and least-change secant methods. SIAM J. Numer. Anal., 30 (5):1291–1314, 1993.

[15] Y. Diouane, M. Gollier, and D. Orban. A nonsmooth exact penalty method for equality-constrained optimization: complexity and implementation. Cahier G–2024–65, GERAD, Montréal, Canada, 2024.

[16] Y. Diouane, M. L. Habiboullah, and D. Orban. Complexity of trust-region methods in the presence of unbounded Hessian approximations. Cahier G–2024–43, GERAD, Montréal, Canada, 2024.

[17] M. Fukushima and H. Mine. A generalized proximal point algorithm for certain non-convex minimization problems. Int. J. Syst. Sci., 12(8):989–1000, 1981.

[18] J.-C. Gilbert and C. Lemaréchal. Some numerical experiments with variable-storage quasi-Newton algorithms. Math. Program., 45:407–435, 1989.

[19] C. Kanzow and T. Lechner. Globalized inexact proximal Newton-type methods for nonconvex composite functions. Comput. Optim. Appl., 78(2):377–410, 2021.

[20] C. Kanzow and P. Mehlitz. Convergence properties of monotone and nonmonotone proximal gradient methods revisited. J. Optimiz. Theory App., 195(2):624–646, 2022.

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

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

[23] G. Leconte, D. Orban, A. Soares Siqueira, and contributors. LinearOperators.jl: Linear Operators for Julia, 2023.

[24] J. D. Lee, Y. Sun, and M. A. Saunders. Proximal Newton-type methods for minimizing composite functions. SIAM J. Optim., 24(3):1420–1443, 2014.

[25] H. Li and Z. Lin. Accelerated proximal gradient methods for nonconvex programming. In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1, NIPS'15, pages 379–387, Cambridge, MA, USA, 2015. MIT Press.

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

[27] J. L. Nazareth. If quasi-Newton then why not quasi-Cauchy? SIAG/OPT Views-and-News, 6:11–14, 1995.

[28] 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.

[29] R. Rockafellar and R. J.-B. Wets. Variational Analysis. Springer Verlag, Heidelberg, Berlin, New York, 1998.

[30] L. Stella, A. Themelis, and P. Patrinos. Forward–backward quasi-Newton methods for nonsmooth optimization problems. Comput. Optim. Appl., 67(3):443–487, 2017.

[31] L. Stella, A. Themelis, P. Sopasakis, and P. Patrinos. A simple and efficient algorithm for nonlinear model predictive control. In 2017 IEEE 56th Annual Conference on Decision and Control (CDC), pages 1939–1944, 2017.

[32] A. Themelis, L. Stella, and P. Patrinos. Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone line seach algorithms. SIAM J. Optim., 28(3):2274–2303, 2018.

[33] Q. Yu and X. Zhang. A smoothing proximal gradient algorithm for matrix rank minimization problem. Comput. Optim. Appl., pages 1–20, 2022.

[34] M. Zhu, J. L. Nazareth, and H. Wolkowicz. The quasi-Cauchy relation and diagonal updating. SIAM J. Optim., 9(4):1192–1204, 1999.