Newton-MR: Inexact Newton Method with minimum residual sub-problem solver

Fred Roosta [a,b,c,*], Yang Liu [a], Peng Xu [d], Michael W. Mahoney [e]

[a] *School of Mathematics and Physics, University of Queensland, Australia*
[b] *The Centre for Information Resilience (CIRES), Australia*
[c] *International Computer Science Institute, Berkeley, USA*
[d] *Institute for Computational and Mathematical Engineering, Stanford University, USA*
[e] *International Computer Science Institute and Department of Statistics, University of California at Berkeley, USA*

A R T I C L E   I N F O

A B S T R A C T

We consider a variant of inexact Newton Method [20,40], called Newton-MR, in which the least-squares sub-problems are solved approximately using Minimum Residual method [79]. By construction, Newton-MR can be readily applied for unconstrained optimization of a class of non-convex problems known as invex, which subsumes convexity as a sub-class. For invex optimization, instead of the classical Lipschitz continuity assumptions on gradient and Hessian, Newton-MR's global convergence can be guaranteed under a weaker notion of joint regularity of Hessian and gradient. We also obtain Newton-MR's problem-independent local convergence to the set of minima. We show that fast local/global convergence can be guaranteed under a novel inexactness condition, which, to our knowledge, is much weaker than the prior related works. Numerical results demonstrate the

* Corresponding author.

E-mail addresses: fred.roosta@uq.edu.au (F. Roosta), yang.liu2@uq.edu.au (Y. Liu), pengxu@stanford.edu (P. Xu), mmahoney@stat.berkeley.edu (M.W. Mahoney).

performance of Newton-MR as compared with several other
Newton-type alternatives on a few machine learning problems.
© 2022 The Author(s). Published by Elsevier Ltd on behalf
of Association of European Operational Research Societies
(EURO). This is an open access article under the CC BY
license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

For a twice differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, consider the unconstrained optimization problem

$$\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}). \tag{1}$$

The canonical example of second-order methods is arguably the classical Newton's method whose iterations are typically written as

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k, \quad \text{with} \quad \mathbf{p}_k = - \left[\mathbf{H}(\mathbf{x}_k) \right]^{-1} \mathbf{g}(\mathbf{x}_k), \tag{2}$$

where $\mathbf{x}_k$, $\mathbf{g}(\mathbf{x}_k) = \nabla f(\mathbf{x}_k)$, $\mathbf{H}(\mathbf{x}_k) = \nabla^2 f(\mathbf{x}_k)$, and α_k are respectively the current iterate, the gradient, the Hessian matrix, and the step-size that is often chosen using an Armijo-type line-search to enforce sufficient decrease in f [76]. When f is smooth and strongly convex, it is well known that the local and global convergence rates of the classical Newton's method are, respectively, quadratic and linear [18,73,76]. For such problems, the Hessian matrix is uniformly positive definite. As a result, if forming the Hessian matrix explicitly and/or solving the linear system in (2) exactly is prohibitive, the update direction $\mathbf{p}_k$ is obtained approximately using conjugate gradient (CG), resulting in the celebrated Newton-CG [76].

Newton-CG is elegantly simple in that its iterations merely involve solving linear systems followed by a certain type of line-search. Despite its simplicity, it has been shown to enjoy various desirable theoretical and algorithmic properties including insensitivity to problem ill-conditioning [81,94] as well as robustness to hyper-parameter tuning [15, 60]. However, in the absence of strong-convexity, Newton's method lacks any favorable convergence analysis. Indeed, recall that to obtain the update direction, $\mathbf{p}_k$, Newton-CG aims at (approximately) solving quadratic sub-problems of the form ([76])

$$\min_{\mathbf{p} \in \mathbb{R}^d} \langle \mathbf{g}(\mathbf{x}_k), \mathbf{p} \rangle + \frac{1}{2} \langle \mathbf{p}, \mathbf{H}(\mathbf{x}_k) \mathbf{p} \rangle . \tag{3}$$

However, if the Hessian is indefinite (as in non-convex settings) or if its positive semi-definite but $\mathbf{g}(\mathbf{x}_k) \notin \mathrm{Range}\,(\mathbf{H}(\mathbf{x}_k))$ (as in weakly convex problems), (3) is simply unbounded below.

In this light, many Newton-type variants have been proposed which aim at extending Newton-CG beyond strongly convex problems to more general settings, e.g., trust-region [32] and cubic regularization [25,26,75]. However, many of these extensions involve sub-problems that are themselves non-trivial to solve, e.g., the sub-problems of trust-region and cubic regularization methods are themselves non-linear and (potentially) non-convex [92,93]. This has prompted many authors to design effective methods for approximately solving such sub-problems, e.g., [16,24,25,34,52,61,85]. These methods can be highly effective in many settings. For example, the trust-region method coupled with CG-Steihaug [85] has shown great promise in several machine learning applications [93]. However, solving such non-trivial sub-problems can be a major challenge in many other settings. For instance, despite the obvious advantages of the CG-Steihaug method, e.g., one matrix-vector product per iteration and the ability to naturally detect negative curvature directions, it is not guaranteed to solve the trust-region sub-problem to an arbitrary accuracy. This can have negative consequences for the convergence speed of the trust-region method. Indeed, if either of the negative curvature or the boundary is encountered too early, the CG-Steihaug method terminates and the resulting step is only slightly, if at all, better than the Cauchy direction [32]. If this occurs too often, the trust-region method's convergence can slow down to be comparable to that of the simple gradient descent method.

Recently, [68] proposes a novel Newton-type method, which not only enjoys sub-problems in the form of linear-systems, but as long as these sub-problems are solved exactly, it has provably fast global convergence for weakly convex objectives. To allow for optimization of a larger class of problems beyond convex, while maintaining the simplicity offered by linear sub-problems, one can consider replacing (3) with an equivalently simple ordinary least-squares (OLS)

$$\min_{\mathbf{p} \in \mathbb{R}^d} \ \frac{1}{2} \left\| \mathbf{H}(\mathbf{x}_k)\mathbf{p} + \mathbf{g}(\mathbf{x}_k) \right\|^2. \tag{4}$$

Clearly, regardless of the Hessian matrix, there is always a solution (in fact at times infinitely many solutions) to the above OLS problem. For example, exact minimum norm solution to the above OLS amounts to iterations of the form

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k, \quad \text{with} \quad \mathbf{p}_k = -\left[\mathbf{H}(\mathbf{x}_k) \right]^\dagger \mathbf{g}(\mathbf{x}_k), \tag{5}$$

where $\mathbf{A}^\dagger$ denotes the Moore-Penrose generalized inverse of matrix $\mathbf{A}$, and α_k is an appropriately chosen step-size. Our aim in this paper is to give an in-depth treatment of the above alternative and study the convergence properties of the resulting algorithm, which we call *Newton-MR*,[1] under a variety of settings.

---

[1] The term "MR" refers to the fact that the sub-problems are in the form of **Minimum Residual**, i.e., least squares, as depicted in Algorithms 1 and 2.

The rest of this paper is organized as follows. We end Section 1 by introducing the notation used throughout the paper. In Section 2, we take a look at high-level consequences of using a direction from (an approximation of) (4). Convergence properties of Newton-MR are gathered in Section 3. This is done, first, by extensive discussion on various assumptions underlying the analysis in Section 3.1, followed by a detailed theoretical development to obtain local and global convergence of Newton-MR in Section 3.2. Numerical examples are provided in Section 4. Conclusions and further thoughts are gathered in Section 5.

Notation   In what follows, vectors and matrices are denoted by bold lowercase and bold uppercase letters, e.g., $\mathbf{v}$ and $\mathbf{V}$, respectively. We use regular lowercase and uppercase letters to denote scalar constants, e.g., c or L. The transpose of a real vector $\mathbf{v}$ is denoted by $\mathbf{v}^\mathsf{T}$. For two vectors, $\mathbf{v}, \mathbf{w}$, their inner-product is denoted as $\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^\mathsf{T}\mathbf{w}$. For a vector $\mathbf{v}$, and a matrix $\mathbf{V}$, $\|\mathbf{v}\|$ and $\|\mathbf{V}\|$ denote the vector ℓ_2 norm and the matrix spectral norm, respectively. For any $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$, $\mathbf{y} \in [\mathbf{x}, \mathbf{z}]$ denotes $\mathbf{y} = \mathbf{x} + \tau(\mathbf{z} - \mathbf{x})$ for some $0 \le \tau \le 1$. For a natural number n, we denote $[n] = \{1, 2, \ldots, n\}$. For any finite collection $\mathcal{S}$, its cardinality is denoted by $|\mathcal{S}|$. The subscript, e.g., $\mathbf{x}_k$, denotes iteration counter. For a matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$, Range $(\mathbf{A})$ and Null $(\mathbf{A})$ denote, respectively, its range, i.e., Range $(\mathbf{A}) = \{\mathbf{A}\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^d\}$ and its null-space, i.e., Null $(\mathbf{A}) = \{\mathbf{y} \in \mathbb{R}^d \mid \mathbf{A}\mathbf{y} = 0\}$. The Moore-Penrose generalized inverse of a matrix $\mathbf{A}$ is denoted by $\mathbf{A}^\dagger$. The identity matrix is written as $\mathbf{I}$ and the indicator function of a set $\mathcal{S}$ is denoted as $\mathbb{1}_{\{\mathcal{S}\}}$. When the minimum is attained at more than one point, "Arg min" denotes the set of minimizers, otherwise "arg min" implies a unique minimizer. Finally, we use $\mathbf{g}(\mathbf{x}) \triangleq \nabla f(\mathbf{x})$ and $\mathbf{H}(\mathbf{x}) \triangleq \nabla^2 f(\mathbf{x})$ for the gradient and the Hessian of f at $\mathbf{x}$, respectively, and at times we drop the dependence on $\mathbf{x}$ by simply using $\mathbf{g}$ and $\mathbf{H}$, e.g., $\mathbf{g}_k = \mathbf{g}(\mathbf{x}_k)$ and $\mathbf{H}_k = \mathbf{H}(\mathbf{x}_k)$.

2. Newton-MR

Before providing a detailed look into the convergence implications of using (4), let us first study higher-level consequences of such a choice as they relate to solving (1).

2.1. Minimum residual sub-problem solver

For solving OLS problems, a plethora of iterative solvers exists. Each of these iterative methods can be highly effective once applied to the class of problems for which they have been designed. Among them, the one method which has been specifically designed to effectively obtain the least-norm solution for all systems, compatible or otherwise, involving real symmetric but potentially indefinite/singular matrices (such as those in (4)) is MINRES-QLP [29,30]. In fact, by requiring only marginally more computations/storage, it has been shown that MINRES-QLP is numerically more stable than MINRES [79] on ill-conditioned problems.

Perhaps a natural question at this stage is "*what are the descent properties of the (approximate) solution to* (4) *using MINRES-QLP?*" Recall that the tth iteration of MINRES-QLP yields a search direction, $\mathbf{p}_k^{(t)}$, which satisfies

$$\mathbf{p}_k^{(t)} = \underset{\mathbf{p}\in\mathcal{K}_t(\mathbf{H}_k,\mathbf{g}_k)}{\arg\min}\ r_k(\mathbf{p}) \triangleq \|\mathbf{g}_k + \mathbf{H}_k\mathbf{p}\|, \tag{6}$$

where $\mathcal{K}_t(\mathbf{H}_k,\mathbf{g}_k) = \text{Span}\{\mathbf{g}_k, \mathbf{H}_k\mathbf{g}_k, \ldots, [\mathbf{H}_k]^{t-1}\mathbf{g}_k\}$ denotes the Krylov sub-space of order t, [49]. Since $\mathbf{0} \in \mathcal{K}_t(\mathbf{H}_k,\mathbf{g}_k)$, it follows that for any t in (6), we have $r_k(\mathbf{p}_k^{(t)}) \leq r_k(\mathbf{0})$, which implies

$$\left\langle \mathbf{p}_k^{(t)}, \mathbf{H}_k\mathbf{g}_k \right\rangle \leq -\left\|\mathbf{H}_k\mathbf{p}_k^{(t)}\right\|^2/2 \leq 0.$$

In other words, $\mathbf{p}_k^{(t)}$ from MINRES, for all t, is a descent direction for $\|\mathbf{g}(\mathbf{x})\|$ at $\mathbf{x}_k$. Let $\mathbf{p}_k = \mathbf{p}_k^{(t)}$ for some $t \geq 1$. As a consequence of such particular descent property of $\mathbf{p}_k$, the step-size, α_k, can be chosen using Armijo-type line search [76] to reduce the norm of the gradient. Hence, we chose $\alpha_k \leq 1$ as the largest value such that for some $0 < \rho < 1$, we get

$$\|\mathbf{g}_{k+1}\|^2 \leq \|\mathbf{g}_k\|^2 + 2\rho\alpha_k \left\langle \mathbf{p}_k, \mathbf{H}_k\mathbf{g}_k \right\rangle. \tag{7}$$

This can be approximately achieved using any standard back-tracking line-search strategy.

2.2. Invexity

From the above discussion, we always have $\langle \mathbf{p}_k, \mathbf{H}_k\mathbf{g}_k \rangle \leq 0$, which implies that

$$\|\mathbf{g}_{k+1}\| \leq \|\mathbf{g}_k\|. \tag{8}$$

In other words, by appropriate application of Hessian and regardless of non-convexity of the problem, one can always obtain descent directions corresponding to the auxiliary problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{g}(\mathbf{x})\|. \tag{9}$$

In certain applications, e.g., chemical physics [2,66] and deep learning [46–48], instead of minimizing the function, the goal is to find zeros of its gradient field, which can be obtained by solving (9). However, when the goal is solving (1), a natural question is "*what is the class of objective functions for which* (9) *is equivalent to* (1)*?*" Clearly, any global optimum of (1) is also a solution to (9). However, the converse only holds for a special class of non-convex functions, known as *invex* [55,69].

Definition 1 *(Invexity).* Let $\mathcal{X} \subseteq \mathbb{R}^d$ be an open set. A differentiable function $f : \mathcal{X} \to \mathbb{R}$ is said to be invex on $\mathcal{X}$ if there exists a vector-valued function $\boldsymbol{\phi} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^d$ such that

$$f(\mathbf{y}) - f(\mathbf{x}) \geq \langle \boldsymbol{\phi}(\mathbf{y}, \mathbf{x}), \mathbf{g}(\mathbf{x}) \rangle, \qquad \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}. \tag{10}$$

The class of functions satisfying (10) with the same $\boldsymbol{\phi}$ is denoted by $\mathcal{F}_{\boldsymbol{\phi}}$. The class of all invex functions is denoted by $\mathcal{F}$.

Invexity characterizes the class of functions for which the first order optimality condition is also sufficient. It is also easily seen that, by choosing $\boldsymbol{\phi}(\mathbf{y}, \mathbf{x}) = \mathbf{y} - \mathbf{x}$ in Definition 1, differentiable convex functions are subsumed in the class of invex functions; see [14,69] for detailed treatments of invexity. Much like all generalized convex functions [23], invexity has been introduced in order to weaken, as much as possible, the convexity requirements in optimization problems, and extend many results related to convex optimization theory to more general functions.

Invexity has recently garnered attention in the machine learning community. Restricted to a special subclass, often referred to as Polyak-Łojasiewicz functions (cf. Section 3.2.3), numerous works have established the convergence of first-order optimization algorithms, e.g., [7,37,57,59,78,90]. In addition, it has recently been shown that several deep learning models enjoy such invex properties in certain regimes, e.g., [6,64]. For more general non-convex machine learning models, [33] proposes a novel regularization framework, which ensures the invexity of the regularized function and is shown to be advantageous over the classical ℓ_2-regularization in terms of the sensitivity to hyper-parameter tuning and out-of-sample generalization performance.

2.3. Connections to root finding algorithms

Considering (9) in lieu of $\mathbf{g}(\mathbf{x}) = 0$ is, in fact, a special case of a more general framework for solving non-linear system of equations involving a vector valued function $\mathbf{F} : \mathbb{R}^m \to \mathbb{R}^n$. More specifically, as an alternative to solving $\mathbf{F}(\mathbf{x}) = 0$, minimization of $\|\mathbf{F}(\mathbf{x})\|$ has been considered extensively in the literature; e.g., [10,11,28,74,80,97,100]. Consequently, Newton-MR can be regarded as a member of the class of inexact Newton methods with line-search for solving nonlinear systems [35,40,41,70]. In our case, the non-linear system of equations arises from optimality condition $\mathbf{g}(\mathbf{x}) = 0$, i.e., $\mathbf{F}(\mathbf{x}) = \mathbf{g}(\mathbf{x})$, which has a symmetric Jacobian. Hence, through the perspective of the application of MINRES within its iterations, Newton-MR can be viewed as a special case of more general Newton-GMRES algorithms for solving non-linear systems [1,12,13,19,20,58].

In light of these connections, the bare-bones iterations of Newton-MR do not constitute novel elements of our work here. What sets our contributions in this paper apart is studying the plethora of desirable theoretical and algorithmic properties of Newton-MR in the context of (9) and its connection to (1). In particular, we give a thorough treatment of the implications of the assumptions that we make as well as the novel inexactness

conditions that we propose as they relate to convergence of Newton-MR. For example, in most prior works on Newton-GMRES, the Jacobian of the non-linear function is assumed full-rank [12,13,20,91]. This, in our setting here, amounts to assuming that the Hessian matrix is invertible, which can be easily violated in the non-convex settings. In fact, to address rank deficient cases, the ubiquitous approach has been to employ regularization techniques such as Levenberg-Marquardt [8,9,36,45,53,95]. By introducing a novel assumption on the interplay between the gradient and the sub-space spanned by the Hessian matrix, we obviate the need for full-rank assumptions or the Hessian regularization techniques. As another example, to guarantee convergence, we will introduce a novel inexactness condition that constitute a major relaxation from the typical relative residual conditions used, almost ubiquitously, in prior works.

3. Theoretical analysis

In this section, we study the convergence properties of some variants of Newton-MR (Algorithms 1 and 2). For this, in Section 3.1, we first give the assumptions underlying our analysis. Under these assumptions, in Section 3.2, we provide detailed local/global convergence analysis of these Newton-MR variants.

3.1. Assumptions

For the theoretical analysis of Newton-MR, we make the following blanket assumptions regarding the properties of the objective function f in (1). Some of these assumptions might seem unconventional at first, however, they are simply generalizations of many typical assumptions made in the similar literature. For example, a strongly-convex function with Lipschitz continuous gradient and Hessian satisfies all of the following assumptions in this Section.

Assumption 1 *(Differentiability).* The function f is twice-differentiable.

In particular, all the first partial derivatives are themselves differentiable, but the second partial derivatives are allowed to be discontinuous. Recall that requiring the first partials be differentiable implies the equality of crossed-partials, which amounts to the symmetric Hessian matrix [56, pp. 732-733].

In the literature for the analysis of non-convex Newton-type methods for (1), to obtain *non-asymptotic quantitative* convergence rates, it is typically assumed that the function is sufficiently smooth in that its gradient and Hessian are Lipschitz continuous. Specifically, for some $0 \leq L_\mathbf{g} < \infty$, $0 \leq L_\mathbf{H} < \infty$, and $\forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, it is assumed that

$$\|\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y})\| \leq L_\mathbf{g} \|\mathbf{x} - \mathbf{y}\|, \tag{11a}$$

$$\|\mathbf{H}(\mathbf{x}) - \mathbf{H}(\mathbf{y})\| \leq L_\mathbf{H} \|\mathbf{x} - \mathbf{y}\|. \tag{11b}$$

For Newton-MR, the smoothness is required for the auxiliary function $\|\mathbf{g}(\mathbf{x})\|^2$, which amounts to a more relaxed smoothness condition.

Assumption 2 *(Moral-smoothness).* For any $\mathbf{x}_0 \in \mathbb{R}^d$, there is $0 \leq L(\mathbf{x}_0) < \infty$, such that

$$\|\mathbf{H}(\mathbf{x})\mathbf{g}(\mathbf{x}) - \mathbf{H}(\mathbf{y})\mathbf{g}(\mathbf{y})\| \leq L(\mathbf{x}_0) \|\mathbf{x} - \mathbf{y}\|^\beta, \quad \forall (\mathbf{x}, \mathbf{y}) \in \mathcal{X}_0 \times \mathbb{R}^d, \tag{12}$$

where $\mathcal{X}_0 \triangleq \left\{ \mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{g}(\mathbf{x})\| \leq \|\mathbf{g}(\mathbf{x}_0)\| \right\}$ and $0 < \beta < \infty$.

Note that in Assumption 2, the constant $L(\mathbf{x}_0)$ depends on the choice of $\mathbf{x}_0$. We recall that, for $\beta = 1$, Assumption 2 is similar to the assumption that is often made in the analysis of many non-linear least-squares algorithms for minimizing $\|\mathbf{F}(\mathbf{x})\|^2$ for some non-linear mapping $\mathbf{F}(\mathbf{x})$ in which $\nabla \|\mathbf{F}(\mathbf{x})\|^2$ is assumed Lipschitz continuous (on a level-set). We refer to Assumption 2 as "moral-smoothness" since by (12), it is only the action of Hessian on the gradient that is required to be Lipschitz continuous, and each gradient and/or Hessian individually can be highly irregular, e.g., gradient can be very non-smooth and Hessian can even be discontinuous.

Example 1 *(A morally-smooth function with discontinuous Hessian).* Consider a quadratically smoothed variant of hinge-loss function,

$$f(\mathbf{x}) = \frac{1}{2} \max \left\{ 0, b \langle \mathbf{a}, \mathbf{x} \rangle \right\}^2,$$

for a given $(\mathbf{a}, b) \in \mathbb{R}^d \times \mathbb{R}$. It is easy to see that

$$\mathbf{g}(\mathbf{x}) = b^2 \langle \mathbf{a}, \mathbf{x} \rangle \, \mathbf{a} \, \mathbb{1}_{\{b\langle \mathbf{a}, \mathbf{x} \rangle > 0\}}, \quad \forall \mathbf{x} \in \mathbb{R}^d,$$

$$\mathbf{H}(\mathbf{x}) = b^2 \mathbf{a}\mathbf{a}^\mathsf{T} \, \mathbb{1}_{\{b\langle \mathbf{a}, \mathbf{x} \rangle > 0\}}, \quad \forall \mathbf{x} \notin \mathcal{N},$$

where $\mathcal{N} \triangleq \left\{ \mathbf{x} \in \mathbb{R}^d \mid b \langle \mathbf{a}, \mathbf{x} \rangle = 0 \right\}$. Clearly, Hessian is discontinuous on $\mathcal{N}$. Since $\mathcal{N}$ is a set of measure zero, i.e., $\mu(\mathcal{N}) = 0$ with respect to the Lebesgue measure μ, we can arbitrarily define $\mathbf{H}(\mathbf{x}) \triangleq \mathbf{0}$ for $\mathbf{x} \in \mathcal{N}$ (in fact, any other definition would work as well). It follows that for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have

$$\begin{aligned}
\|\mathbf{H}(\mathbf{x})\mathbf{g}(\mathbf{x}) - \mathbf{H}(\mathbf{x})\mathbf{g}(\mathbf{x})\| &= b^4 \|\mathbf{a}\|^2 \left\| \langle \mathbf{a}, \mathbf{x} \rangle \mathbf{a} \, \mathbb{1}_{\{b\langle \mathbf{a}, \mathbf{x} \rangle > 0\}} - \langle \mathbf{a}, \mathbf{y} \rangle \mathbf{a} \, \mathbb{1}_{\{b\langle \mathbf{a}, \mathbf{y} \rangle > 0\}} \right\| \\
&\leq b^4 \|\mathbf{a}\|^3 \left| \langle \mathbf{a}, \mathbf{x} \rangle \, \mathbb{1}_{\{b\langle \mathbf{a}, \mathbf{x} \rangle > 0\}} - \langle \mathbf{a}, \mathbf{y} \rangle \, \mathbb{1}_{\{b\langle \mathbf{a}, \mathbf{y} \rangle > 0\}} \right| \\
&\leq b^4 \|\mathbf{a}\|^4 \|\mathbf{x} - \mathbf{y}\|.
\end{aligned}$$

The last inequality follows by considering four cases with $\mathbb{1}_{\{b\langle \mathbf{a}, \mathbf{x} \rangle > 0\}} = \mathbb{1}_{\{b\langle \mathbf{a}, \mathbf{y} \rangle > 0\}}$ and $\mathbb{1}_{\{b\langle \mathbf{a}, \mathbf{x} \rangle > 0\}} \neq \mathbb{1}_{\{b\langle \mathbf{a}, \mathbf{y} \rangle > 0\}}$. Indeed, for the former two cases, the last inequality follows immediately. For the latter two cases, suppose without loss of generality that $\mathbb{1}_{\{b\langle \mathbf{a}, \mathbf{x} \rangle > 0\}} = 1$ and $\mathbb{1}_{\{b\langle \mathbf{a}, \mathbf{y} \rangle > 0\}} = 0$. In this case, we use the fact that by $b \langle \mathbf{a}, \mathbf{y} \rangle \leq 0$, we have $b \langle \mathbf{a}, \mathbf{x} \rangle \leq b \langle \mathbf{a}, \mathbf{x} \rangle - b \langle \mathbf{a}, \mathbf{y} \rangle$.

It is easy to show that (12) is implied by (11), and hence, it constitutes a relaxation on the typical smoothness requirements for gradient and Hessian, commonly found in the literature.

Lemma 1 *(Moral-smoothness* (12) *is less strict than smoothness* (11)*). Suppose* (11) *holds. Then, for any* $\mathbf{x}_0$*, we have* (12) *with* $\beta = 1$ *and* $L(\mathbf{x}_0) = L_{\mathbf{g}}^2 + L_{\mathbf{H}} \|\mathbf{g}(\mathbf{x}_0)\|$.

Proof. Suppose (11) is satisfied and fix any $\mathbf{x}_0$ with the corresponding $\mathcal{X}_0$ as defined in Assumption 2. Since for any two matrices $\mathbf{A}, \mathbf{B}$ and two vectors $\mathbf{x}, \mathbf{y}$, we have $\|\mathbf{A}\mathbf{x} - \mathbf{B}\mathbf{y}\| \le \|\mathbf{A}\|\|\mathbf{x} - \mathbf{y}\| + \|\mathbf{y}\|\|\mathbf{A} - \mathbf{B}\|$, it implies that $\forall \mathbf{x} \in \mathcal{X}_0, \ \forall \mathbf{y} \in \mathbb{R}^d$

$$\|\mathbf{H}(\mathbf{y})\mathbf{g}(\mathbf{y}) - \mathbf{H}(\mathbf{x})\mathbf{g}(\mathbf{x})\| \le \|\mathbf{H}(\mathbf{y})\| \|\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y})\| + \|\mathbf{g}(\mathbf{x})\| \|\mathbf{H}(\mathbf{x}) - \mathbf{H}(\mathbf{y})\|$$
$$\le L_{\mathbf{g}}^2 \|\mathbf{y} - \mathbf{x}\| + L_{\mathbf{H}} \|\mathbf{g}(\mathbf{x})\| \|\mathbf{y} - \mathbf{x}\|$$
$$\le \left(L_{\mathbf{g}}^2 + L_{\mathbf{H}} \|\mathbf{g}(\mathbf{x}_0)\| \right) \|\mathbf{y} - \mathbf{x}\|,$$

where the last inequality follows since $\mathbf{x} \in \mathcal{X}_0$. □

By Lemma 1, any smooth function satisfying (11) would also satisfy Assumption 2 with $\beta = 1$. Below, we bring examples which show that the converse does not necessarily hold.

Example 2 *(Smoothness* (11) *is strictly stronger than moral-smoothness* (12)*).* In this example, for three regimes of $0 < \beta < 1, \beta = 1$, and $\beta > 1$, we give concrete counter-examples of functions on $\mathbb{R}$, i.e., $d = 1$, which satisfy (12) but not (11).

1. For $0 < \beta < 1$, consider a solution to the following ODE

$$f''(x)f'(x) = x^\beta, \ x > x_0 \triangleq \left((1 + \beta)a \right)^{\frac{1}{(1+\beta)}}, \quad f(x_0) = b, \quad f'(x_0) = 0,$$

where $b \in \mathbb{R}, a \in \mathbb{R}^+$. Such a solution can be of the form

$$f(x) = b + \sqrt{\frac{2}{1 + \beta}} \int_{x_0}^{x} \sqrt{t^{\beta+1} - (1 + \beta)a} \, dt, \tag{13a}$$

and satisfies (12) with $0 < \beta < 1$, and $L = 1$. It is easily verified that $f''(x)$ and $f'''(x)$ are both unbounded; hence f violates (11). Note that, on (x_0, ∞), (13a) is convex, and hence by definition, it is invex.

2. For $\beta = 1$, the condition (12) implies that $\left\| \left[\nabla^2 f(\mathbf{x}) \right]^2 + \left\langle \nabla^3 f(\mathbf{x}), \mathbf{g}(\mathbf{x}) \right\rangle \right\| \le L$. In one variable, for $b \in \mathbb{R}, a \in \mathbb{R}^{++}$, the solution to the following ODE

$$f''(x)f'(x) = x, \quad x > x_0 \triangleq \sqrt{2a}, \quad f(x_0) = b - a\log(\sqrt{2a}), \quad f'(x_0) = 0,$$

can be written as

$$f(x) = \frac{1}{2}x\sqrt{x^2 - 2a} - a\log\left(\sqrt{x^2 - 2a} + x\right) + b, \tag{13b}$$

which satisfies (12) with $\beta = L = 1$. It can be easily verified that

$$f''(x) = \frac{x}{\sqrt{x^2 - 2a}}, \quad \text{and} \quad f'''(x) = -\frac{2a}{(x^2 - 2a)^{\frac{3}{2}}},$$

which are unbounded, and so, f does not satisfy (11a) or (11b). It is clear that, on (x_0, ∞), (13b) is convex, and hence it is invex.

Similarly, for $b \in \mathbb{R}, a \in \mathbb{R}^{++}$, one can consider

$$f''(x)f'(x) = -x, \quad x \in (-\sqrt{2a}, \sqrt{2a}), \quad f(0) = b, \quad f'(0) = -\sqrt{2a}.$$

The solution to this ODE is of the form

$$f(x) = b - \frac{x}{2}\sqrt{2a - x^2} - a\arctan\left(\frac{x}{\sqrt{2a - x^2}}\right), \tag{13c}$$

which satisfies (12) with $\beta = L = 1$. However, since

$$f''(x) = \frac{x}{\sqrt{2a - x^2}}, \quad \text{and} \quad f'''(x) = \frac{2a}{(2a - x^2)^{\frac{3}{2}}},$$

are both unbounded, such a function does not satisfy either of (11a) or (11b).

3. For $\beta > 1$, the condition (12) implies that $\left[\nabla^2 f(\mathbf{x})\right]^2 + \left\langle \nabla^3 f(\mathbf{x}), \mathbf{g}(\mathbf{x}) \right\rangle = 0$. In one variable, the solution to the following ODE

$$(f''(x))^2 + f'''(x)f'(x) = 0, \quad x > c/2, \quad f(c/2) = a, \quad f'(c/2) = 0, \quad f''(c) = 3b/\sqrt{c},$$

can be written as

$$f(x) = a + b(2x - c)^{3/2}. \tag{13d}$$

This implies that there are non-trivial function which can satisfy (12) for $\beta > 1$. In addition, since

$$f''(x) = \frac{3b}{(2x - c)^{1/2}}, \quad f'''(x) = \frac{-3b}{(2x - c)^{3/2}},$$

are both unbounded, it implies that such f does not satisfy either assumptions in (11). It is clear that, on $(c/2, \infty)$, (13d) is convex, and hence it is invex.

Assumption 2 allows for Hessian to grow unboundedly, however it implies a certain restriction on the growth rate of the action of Hessian on the gradient, i.e., $\mathbf{H}(\mathbf{x})\mathbf{g}(\mathbf{x})$.

Lemma 2 *(Growth rate of $\mathbf{H}(\mathbf{x})\mathbf{g}(\mathbf{x})$). Under Assumption 2, for any $\mathbf{x}_0 \in \mathbb{R}^d$, we have*

$$L(\mathbf{x}_0) \geq \left(\frac{2\beta}{\beta+1}\right)^\beta \frac{\|\mathbf{H}(\mathbf{x})\mathbf{g}(\mathbf{x})\|^{\beta+1}}{\|\mathbf{g}(\mathbf{x})\|^{2\beta}}, \quad \forall \mathbf{x} \in \mathcal{X}_0,$$

where $L(\mathbf{x}_0), \beta$ and $\mathcal{X}_0$ are as defined in Assumption 2.

Proof. If $L(\mathbf{x}_0) = 0$, then the result follows immediately. Suppose $L(\mathbf{x}_0) > 0$. Under Assumption 2, Lemma 5 with $h(\mathbf{x}) = \|\mathbf{g}(\mathbf{x})\|^2 / 2$, gives

$$\|\mathbf{g}(\mathbf{x}+\mathbf{p})\|^2 \leq \|\mathbf{g}(\mathbf{x})\|^2 + 2\langle\mathbf{H}(\mathbf{x})\mathbf{g}(\mathbf{x}), \mathbf{p}\rangle + \frac{2L(\mathbf{x}_0)}{(\beta+1)} \|\mathbf{p}\|^{\beta+1}, \quad \forall \mathbf{x} \in \mathcal{X}_0, \quad \forall \mathbf{p} \in \mathbb{R}^d.$$

Let $m(\mathbf{p}) \triangleq \|\mathbf{g}(\mathbf{x})\|^2 + 2\langle\mathbf{H}(\mathbf{x})\mathbf{g}(\mathbf{x}), \mathbf{p}\rangle + 2L(\mathbf{x}_0)/(\beta+1)\|\mathbf{p}\|^{\beta+1}$. Consider $\mathbf{p}^\star \neq \mathbf{0}$ such that

$$\nabla m(\mathbf{p}^\star) = 2\mathbf{H}(\mathbf{x})\mathbf{g}(\mathbf{x}) + 2L(\mathbf{x}_0)\|\mathbf{p}^\star\|^{\beta-1}\mathbf{p}^\star = 0.$$

It implies that $\|\mathbf{p}^\star\|^{\beta-1}\mathbf{p}^\star = -\mathbf{H}(\mathbf{x})\mathbf{g}(\mathbf{x})/L(\mathbf{x}_0)$. As a result, for such $\mathbf{p}^\star$, we must have $\|\mathbf{p}^\star\|^\beta = \|\mathbf{H}(\mathbf{x})\mathbf{g}(\mathbf{x})\|/L(\mathbf{x}_0)$, and $\langle\mathbf{H}(\mathbf{x})\mathbf{g}(\mathbf{x}), \mathbf{p}^\star\rangle = -\|\mathbf{H}(\mathbf{x})\mathbf{g}(\mathbf{x})\|\|\mathbf{p}^\star\|$, where the equality follows since $\mathbf{p}^\star$ is a scalar multiple of $\mathbf{H}(\mathbf{x})\mathbf{g}(\mathbf{x})$. Since $m(\mathbf{p})$ is convex, it follows that

$$\min_{\mathbf{p}} m(\mathbf{p}) = \|\mathbf{g}(\mathbf{x})\|^2 - 2\|\mathbf{H}(\mathbf{x})\mathbf{g}(\mathbf{x})\|\|\mathbf{p}^\star\| + \frac{2L(\mathbf{x}_0)}{(\beta+1)}\|\mathbf{p}^\star\|^{\beta+1}$$

$$= \|\mathbf{g}(\mathbf{x})\|^2 - \frac{2\|\mathbf{H}(\mathbf{x})\mathbf{g}(\mathbf{x})\|^{\frac{\beta+1}{\beta}}}{L(\mathbf{x}_0)^{\frac{1}{\beta}}} + \frac{2\|\mathbf{H}(\mathbf{x})\mathbf{g}(\mathbf{x})\|^{\frac{\beta+1}{\beta}}}{(\beta+1)L(\mathbf{x}_0)^{\frac{1}{\beta}}}$$

$$= \|\mathbf{g}(\mathbf{x})\|^2 - \left(\frac{\beta}{\beta+1}\right)\left(\frac{2}{L(\mathbf{x}_0)^{\frac{1}{\beta}}}\right)\|\mathbf{H}(\mathbf{x})\mathbf{g}(\mathbf{x})\|^{\frac{\beta+1}{\beta}}.$$

The result follows since we have $m(\mathbf{p}) \geq \|\mathbf{g}(\mathbf{x}+\mathbf{p})\|^2 \geq 0, \forall \mathbf{p}$. $\square$

We also require the following regularity on the pseudo-inverse of the Hessian matrix.

Assumption 3 *(Pseudo-inverse regularity). For any $\mathbf{x}_0 \in \mathbb{R}^d$, there is a $\gamma(\mathbf{x}_0) > 0$, s.t.*

$$\|[\mathbf{H}(\mathbf{x})]^\dagger\| \leq 1/\gamma(\mathbf{x}_0), \quad \forall \mathbf{x} \in \mathcal{X}_0, \tag{14}$$

where $\mathcal{X}_0$ is as in Assumption 2.

It turns out that (14) is equivalent to the regularity of the Hessian matrix on its *range space*.

Lemma 3. (14) *is equivalent to*

$$\|\mathbf{H}(\mathbf{x})\mathbf{p}\| \geq \gamma(\mathbf{x}_0) \|\mathbf{p}\|, \quad \forall \mathbf{x} \in \mathcal{X}_0, \quad and \quad \forall \mathbf{p} \in Range\left(\mathbf{H}(\mathbf{x})\right). \tag{15}$$

The proof of Lemma 3 follows simply by considering the eigenvalue decomposition of $\mathbf{H}$. Assumption 3 is trivially satisfied for all strongly convex functions. However, it is also satisfied for any function whose (possibly rank-deficient) Hessian is uniformly positive definite in the sup-space spanned by its range.

Example 3 *(Functions satisfying Assumption 3).* A simple example of a non-strongly convex function satisfying Assumption 3 is the undetermined least squares $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 / 2$ with full row rank matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$, $n \leq d$. This problem is clearly only weakly convex since the Hessian matrix, $\mathbf{A}^\mathsf{T}\mathbf{A} \in \mathbb{R}^{d \times d}$, is rank-deficient. However, it is easy to see that (14) holds with $\gamma(\mathbf{x}_0) = \gamma = \sigma_n^2(\mathbf{A})$, where σ_n is the smallest non-zero singular value of $\mathbf{A}$.

Finally, we make the following structural assumption about f with regards to the gradient and its projection onto the range space of Hessian.

Assumption 4 *(Gradient-Hessian null-space property).* For any $\mathbf{x} \in \mathbb{R}^d$, let $\mathbf{U}_\mathbf{x}$ and $\mathbf{U}_\mathbf{x}^\perp$ denote arbitrary orthogonal bases for $\text{Range}(\mathbf{H}(\mathbf{x}))$ and its orthogonal complement, respectively. A function is said to satisfy the Gradient-Hessian Null-Space property, if for any $\mathbf{x}_0 \in \mathbb{R}^d$, there is a $0 < \nu(\mathbf{x}_0) \leq 1$, such that

$$\nu(\mathbf{x}_0) \left\| \left(\mathbf{U}_\mathbf{x}^\perp\right)^\mathsf{T} \mathbf{g}(\mathbf{x}) \right\|^2 \leq \left(1 - \nu(\mathbf{x}_0)\right) \|\mathbf{U}_\mathbf{x}^\mathsf{T}\mathbf{g}(\mathbf{x})\|^2, \quad \forall \mathbf{x} \in \mathcal{X}_0, \tag{16}$$

where $\mathcal{X}_0$ is as in Assumption 2.

Lemma 4 gives some simple, yet useful, consequences of Assumption 4.

Lemma 4. *Under Assumption 4, we have*

$$\nu(\mathbf{x}_0) \|\mathbf{g}(\mathbf{x})\|^2 \leq \|\mathbf{U}_\mathbf{x}^\mathsf{T}\mathbf{g}(\mathbf{x})\|^2, \qquad \forall \mathbf{x} \in \mathcal{X}_0, \tag{17a}$$

$$\left(1 - \nu(\mathbf{x}_0)\right) \|\mathbf{g}(\mathbf{x})\|^2 \geq \left\| \left(\mathbf{U}_\mathbf{x}^\perp\right)^\mathsf{T} \mathbf{g}(\mathbf{x}) \right\|^2, \qquad \forall \mathbf{x} \in \mathcal{X}_0. \tag{17b}$$

Proof. For a given $\mathbf{x}_0 \in \mathbb{R}^d$, take any $\mathbf{x} \in \mathcal{X}_0$. For (17a), we have

$$\|\mathbf{g}(\mathbf{x})\|^2 = \left\| \left(\mathbf{U}_\mathbf{x}\mathbf{U}_\mathbf{x}^\mathsf{T} + \mathbf{U}_\mathbf{x}^\perp \left(\mathbf{U}_\mathbf{x}^\perp\right)^\mathsf{T}\right) \mathbf{g}(\mathbf{x}) \right\|^2 = \|\mathbf{U}_\mathbf{x}^\mathsf{T}\mathbf{g}(\mathbf{x})\|^2 + \left\| \left(\mathbf{U}_\mathbf{x}^\perp\right)^\mathsf{T} \mathbf{g}(\mathbf{x}) \right\|^2$$

$$\leq \frac{1 - \nu(\mathbf{x}_0)}{\nu(\mathbf{x}_0)} \left\| \mathbf{U}_{\mathbf{x}}^\mathsf{T} \mathbf{g}(\mathbf{x}) \right\|^2 + \left\| \mathbf{U}_{\mathbf{x}}^\mathsf{T} \mathbf{g}(\mathbf{x}) \right\|^2 = \frac{1}{\nu(\mathbf{x}_0)} \left\| \mathbf{U}_{\mathbf{x}}^\mathsf{T} \mathbf{g}(\mathbf{x}) \right\|^2.$$

Also, (17b) is obtained similarly. □

Assumption 4 ensures that, as long as the gradient is non-zero, its angle with the sub-space spanned by the Hessian matrix is uniformly bounded away from zero. In other words, the gradient will never become arbitrarily orthogonal to the range space of Hessian.

Example 4 *(Over-parameterized ERM using linear predictor models)*. Consider an empirical risk minimization (ERM) problem involving linear predictor models [84],

$$f(\mathbf{x}) = \sum_{i=1}^{n} f_i(\mathbf{a}_i^\mathsf{T} \mathbf{x}), \tag{18}$$

where $\mathbf{a}_i \in \mathbb{R}^d, i = 1, \ldots n$ are given data points and each $f_i : \mathbb{R} \to \mathbb{R}$ is some nonlinear misfit or loss corresponding to the ith data point. Consider the over-parameterized settings ($n \leq d$), which has recently garnered significant attention within the machine learning community, e.g., [64,71,86,88]. Assume that the data points are linearly independent, i.e., Range $(\{\mathbf{a}_i\}_{i=1}^n) = \mathbb{R}^n$. Further, suppose each loss function f_i is such that if $f_i''(t) = 0$ then we must also have that $f_i'(t) = 0$. Many loss functions satisfy this assumption, e.g., the convex function t^{2p} for any $p \geq 1$, so that $f_i(\mathbf{a}_i^\mathsf{T} \mathbf{x}) = \langle \mathbf{a}_i, \mathbf{x} \rangle^{2p}$. Another example is when each f_i is such that $f_i''(t) > 0, \forall t$, e.g., logistic function $\log(1 + e^{-t})$ or any strongly convex. Note that even if each f_i is strongly convex, since we have $n \leq d$, the overall objective f in (18) is still only weakly convex.

Define

$$\mathbf{A} = \begin{pmatrix} \mathbf{a}_1^\mathsf{T} \\ \vdots \\ \mathbf{a}_n^\mathsf{T} \end{pmatrix} \in \mathbb{R}^{n \times d}, \quad \mathbf{D} = \begin{pmatrix} f''(\mathbf{a}_1^\mathsf{T} \mathbf{x}) & & & \\ & f''(\mathbf{a}_2^\mathsf{T} \mathbf{x}) & & \\ & & \ddots & \\ & & & f''(\mathbf{a}_n^\mathsf{T} \mathbf{x}) \end{pmatrix} \in \mathbb{R}^{n \times n}.$$

It is easy to see that

$$\mathbf{g}(\mathbf{x}) = \mathbf{A}^\mathsf{T} \begin{pmatrix} f_i'(\mathbf{a}_1^\mathsf{T} \mathbf{x}) \\ \vdots \\ f_i'(\mathbf{a}_n^\mathsf{T} \mathbf{x}) \end{pmatrix}, \quad \mathbf{H}(\mathbf{x}) = \mathbf{A}^\mathsf{T} \mathbf{D} \mathbf{A}.$$

For the gradient to be in the range of the Hessian, we must have for some $\mathbf{v} \in \mathbb{R}^d$

$$\mathbf{A}^\mathsf{T} \mathbf{D} \mathbf{A} \mathbf{v} = \mathbf{A}^\mathsf{T} [f'(\mathbf{a}_1^\mathsf{T} \mathbf{x}), \ldots, f'(\mathbf{a}_n^\mathsf{T} \mathbf{x})]^\mathsf{T}. \tag{19}$$

Under the assumption on f_i, it can be seen that $\mathbf{v} = \mathbf{A}^\dagger \mathbf{D}^\dagger [f'(\mathbf{a}_1^\mathsf{T}\mathbf{x}), \ldots, f'(\mathbf{a}_n^\mathsf{T}\mathbf{x})]^\mathsf{T}$, satisfies (19), where $\mathbf{A}^\dagger = \mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T})^{-1} \in \mathbb{R}^{d \times n}$. Hence, Assumption 4 holds with $\nu(\mathbf{x}_0) = \nu = 1$.

3.2. Convergence analysis

In this section, we discuss the convergence properties of Newton-MR. For this, in Section 3.2.1, we first consider the slightly simpler, yet less practical, case where the sub-problems (20) are solved exactly (Algorithm 1). Clearly, this is too stringent, and as a result, in Section 3.2.2, it is subsequently relaxed to allow inexact solutions (Algorithm 2). Convergence under a generalized variant of Polyak-Łojasiewicz inequality will be treated in Section 3.2.3.

The following simple Lemma relating to (12) is frequently used in our theoretical analysis. The proof can be found in most textbooks and is only given here for completeness.

Lemma 5. *Consider any* $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$, $0 < \beta < \infty$, $0 \le L < \infty$ *and* $h : \mathbb{R}^d \to \mathbb{R}$. *If*

$$\|\nabla h(\mathbf{y}) - \nabla h(\mathbf{x})\| \le L \|\mathbf{y} - \mathbf{x}\|^\beta, \ \forall \mathbf{y} \in [\mathbf{x}, \mathbf{z}],$$

then

$$h(\mathbf{y}) \le h(\mathbf{x}) + \langle \nabla h(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + L \|\mathbf{y} - \mathbf{x}\|^{\beta+1} / (\beta + 1), \ \forall \mathbf{y} \in [\mathbf{x}, \mathbf{z}].$$

Proof. For any $\mathbf{y} \in [\mathbf{x}, \mathbf{z}]$, using the mean value theorem, we have

$$h(\mathbf{y}) - h(\mathbf{x}) - \langle \nabla h(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle = \int_0^1 \langle \nabla h(\mathbf{x} + \tau(\mathbf{y} - \mathbf{x})), \mathbf{y} - \mathbf{x} \rangle \, d\tau - \langle \nabla h(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$$

$$\le \|\mathbf{y} - \mathbf{x}\| \int_0^1 \|\nabla h(\mathbf{x} + \tau(\mathbf{y} - \mathbf{x})) - \nabla h(\mathbf{x})\| \, d\tau$$

$$\le L \|\mathbf{y} - \mathbf{x}\|^{\beta+1} \int_0^1 \tau^\beta d\tau \le \frac{L}{\beta + 1} \|\mathbf{y} - \mathbf{x}\|^{\beta+1}. \quad \square$$

3.2.1. Exact update

The underlying sub-problem of Newton-MR, at k^{th} iteration, involves OLS problem of the form (4). Under the invexity assumption, the Hessian matrix can be indefinite and rank deficient, and as a result, the sub-problem (4) may contain infinitely many solutions. Indeed, the general solution of (4) is written as

$$\mathbf{p} = -[\mathbf{H}_k]^\dagger \mathbf{g}_k + \left(\mathbf{I} - \mathbf{H}_k [\mathbf{H}_k]^\dagger\right) \mathbf{q}, \quad \forall \mathbf{q} \in \mathbb{R}^d,$$

Algorithm 1 Exact Newton-MR.

Input:
  - Initial iterate $\mathbf{x}_0 \in \mathbb{R}^d$, Line-search parameter $0 < \rho < 1$, Termination tolerance $0 < \epsilon$

for $k = 0, 1, 2, \cdots$ until $\|\mathbf{g}_k\| \leq \epsilon$ **do**
    Find $\mathbf{p}_k$ using (20), i.e., set $\mathbf{p}_k = -[\mathbf{H}_k]^\dagger \mathbf{g}_k$
    Find step-size, α_k, such that (7) holds with ρ
    Set $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k$
end for
Output: $\mathbf{x}$ for which $\|\mathbf{g}(\mathbf{x})\| \leq \epsilon$

---

which implies that when $\mathbf{H}_k$ has a non-trivial null-space, the sub-problem (4) has infinitely many solutions. Among these, the one with the minimum norm is defined uniquely as

$$\min_{\mathbf{p} \in \mathbb{R}^d} \ \|\mathbf{p}\| \quad \text{s.t.} \quad \mathbf{p} \in \operatorname*{Arg\,min}_{\widehat{\mathbf{p}} \in \mathbb{R}^d} \|\mathbf{g}_k + \mathbf{H}_k \widehat{\mathbf{p}}\| , \tag{20}$$

which yields $\mathbf{p}_k = -[\mathbf{H}_k]^\dagger \mathbf{g}_k$. In our analysis below, among all possible solutions, we will choose the least norm solution, to get iterations of the form (5). The resulting Newton-MR variant with exact update (20) is depicted in Algorithm 1.

Theorem 1 gives the convergence guarantees of Algorithm 1.

Theorem 1 (*Convergence of Algorithm 1*). *Consider Assumptions 1 to 4. For the iterates of Algorithm 1, we have*

$$\|\mathbf{g}_{k+1}\|^2 \leq \left(1 - 2\rho\tau(\mathbf{x}_0) \|\mathbf{g}_k\|^{(1-\beta)/\beta}\right) \|\mathbf{g}_k\|^2 ,$$

where $\tau(\mathbf{x}_0) \triangleq \left((1 - \rho)(1 + \beta) \left(\nu(\mathbf{x}_0)\gamma(\mathbf{x}_0)\right)^{1+\beta}/L(\mathbf{x}_0)\right)^{1/\beta}$, ρ *is the line-search parameter of Algorithm 1,* $\mathbf{x}_0$ *is the initial iterate,* $(\beta, L(\mathbf{x}_0))$ *are as in Assumption 2,* $\gamma(\mathbf{x}_0)$ *is as in Assumption 3, and* $\nu(\mathbf{x}_0)$ *is defined in Assumption 4.*

Proof. We first note that, by (8), all iterates of Algorithm 1 remain in $\mathcal{X}_0$. Let $\mathbf{U}_k \in \mathbb{R}^{d \times r}$ be any orthogonal basis for the range of $\mathbf{H}_k$ and $r = \operatorname{rank}(\mathbf{H}_k) \leq d$. It follows that $\langle \mathbf{H}_k \mathbf{g}_k, \mathbf{p}_k \rangle = -\left\langle \mathbf{g}_k, \mathbf{H}_k [\mathbf{H}_k]^\dagger \mathbf{g}_k \right\rangle = -\|\mathbf{U}_k^\mathsf{T} \mathbf{g}_k\|^2$. Using (7), we get the reduction in the gradient norm as

$$\|\mathbf{g}_{k+1}\|^2 \leq \|\mathbf{g}_k\|^2 - 2\rho\alpha_k \|\mathbf{U}_k^\mathsf{T} \mathbf{g}_k\|^2 . \tag{21}$$

All that is left is to obtain a non-zero iteration independent lower-bound on α_k, i.e., $\alpha_k \geq \alpha > 0$, for which (21) holds. For this, using Assumption 2 and Lemma 5, with $\mathbf{x} = \mathbf{x}_k$, $\mathbf{z} = \mathbf{x}_k + \mathbf{p}_k$, $\mathbf{y} = \mathbf{x}_k + \alpha \mathbf{p}_k$, and $h(\mathbf{x}) = \|\mathbf{g}\|^2 / 2$, we get

$$\|\mathbf{g}_{k+1}\|^2 \leq \|\mathbf{g}_k\|^2 + 2\alpha \langle \mathbf{H}_k \mathbf{g}_k, \mathbf{p}_k \rangle + \frac{2\alpha^{\beta+1} L(\mathbf{x}_0)}{(\beta + 1)} \|\mathbf{p}_k\|^{\beta+1} . \tag{22}$$

Assumption 3 implies $\|\mathbf{p}_k\| = \left\|[\mathbf{H}_k]^\dagger \mathbf{g}_k\right\| \leq \|\mathbf{g}_k\| / [\gamma(\mathbf{x}_0)]^{\beta+1}$. Hence, it follows that

$$\|\mathbf{g}_{k+1}\|^2 \leq \|\mathbf{g}_k\|^2 - 2\alpha \|\mathbf{U}_k^\mathsf{T} \mathbf{g}_k\|^2 + \frac{2\alpha^{\beta+1} L(\mathbf{x}_0)}{[\gamma(\mathbf{x}_0)]^{\beta+1}(\beta+1)} \|\mathbf{g}_k\|^{\beta+1}.$$

Using the above as well as (21), we require for α to satisfy

$$\|\mathbf{g}_k\|^2 - 2\alpha \|\mathbf{U}_k^\mathsf{T} \mathbf{g}_k\|^2 + \frac{2\alpha^{\beta+1} L(\mathbf{x}_0)}{[\gamma(\mathbf{x}_0)]^{\beta+1}(\beta+1)} \|\mathbf{g}_k\|^{\beta+1} \leq \|\mathbf{g}_k\|^2 - 2\rho\alpha \|\mathbf{U}_k^\mathsf{T} \mathbf{g}_k\|^2.$$

But from (17a) and Assumption 4, this is implied if α satisfies $\alpha^{\beta+1} L(\mathbf{x}_0) \|\mathbf{g}_k\|^{\beta+1} \leq [\gamma(\mathbf{x}_0)]^{\beta+1}(\beta+1)(1-\rho)\nu(\mathbf{x}_0)\alpha \|\mathbf{g}_k\|^2$, which is given if $\alpha \leq ((1-\rho)(1+\beta)\nu(\mathbf{x}_0)[\gamma(\mathbf{x}_0)]^{1+\beta} / L(\mathbf{x}_0))^{1/\beta} \|\mathbf{g}_k\|^{(1-\beta)/\beta}$. This implies that any step-size returned from the line-search will be at least as large as the right-hand side. Now, using Assumption 4 and Lemma 4 again, we get $\|\mathbf{g}_{k+1}\|^2 \leq \|\mathbf{g}_k\|^2 - 2\rho\alpha \|\mathbf{U}_k^\mathsf{T} \mathbf{g}_k\|^2 \leq (1 - 2\rho\nu(\mathbf{x}_0)\alpha) \|\mathbf{g}_k\|^2.$ □

Remark 1. The convergence rate of Theorem 1, for $\beta \neq 1$, seems rather complicated. However, one can ensure that it is indeed meaningful, i.e., the rate is positive and strictly less than one. From Lemma 2 and Assumptions 3 and 4, it follows that $\forall \mathbf{x} \in \mathcal{X}_0$, we have

$$L(\mathbf{x}_0) \geq \left(\frac{2\beta}{\beta+1}\right)^\beta \frac{\|\mathbf{H}(\mathbf{x})\mathbf{g}(\mathbf{x})\|^{\beta+1}}{\|\mathbf{g}(\mathbf{x})\|^{2\beta}} = \left(\frac{2\beta}{\beta+1}\right)^\beta \frac{\|\mathbf{H}(\mathbf{x})(\mathbf{U}_\mathbf{x}\mathbf{U}_\mathbf{x}^\mathsf{T}\mathbf{g}(\mathbf{x}))\|^{\beta+1}}{\|\mathbf{g}(\mathbf{x})\|^{2\beta}}$$

$$\geq \left(\frac{2\beta}{\beta+1}\right)^\beta (\gamma(\mathbf{x}_0)\sqrt{\nu(\mathbf{x}_0)})^{\beta+1} \|\mathbf{g}(\mathbf{x})\|^{1-\beta},$$

where $\mathbf{U}_\mathbf{x} \in \mathbb{R}^{d \times r}$ is any orthogonal basis for the range of $\mathbf{H}(\mathbf{x})$. It can also be shown that

$$\rho\left((1+\beta)(1-\rho)\right)^{1/\beta} \leq \frac{\beta}{1+\beta}, \ \forall \rho \in (0,1), \ \forall \beta \in (0,\infty),$$

with equality holding at $\rho = \beta/(1+\beta)$. Hence, noting that $\nu(\mathbf{x}_0) \leq 1$, we obtain

$$L(\mathbf{x}_0) \geq 2^\beta \rho^\beta (1-\rho)(1+\beta) \left(\nu(\mathbf{x}_0)\gamma(\mathbf{x}_0)\right)^{\beta+1} \|\mathbf{g}(\mathbf{x})\|^{1-\beta}, \quad \forall \mathbf{x} \in \mathcal{X}_0,$$

which ensures

$$0 \leq 1 - 2\rho\tau(\mathbf{x}_0) \|\mathbf{g}_k\|^{(1-\beta)/\beta} < 1, \quad k = 0, 1, \dots. \tag{23}$$

Remark 2 (*Parallels between Newton-MR/Newton-CG and MINRES/CG*). In the traditional settings of strongly convex and smooth functions, the global convergence rate in Theorem 1 with $\beta = 1$ is identical to that of Newton-CG; see [81] for example. However, the former indicates the rate of reduction in the norm of the gradient of the function,

while the latter corresponds to the reduction in the function value itself. This relationship is reminiscent of the convergence guarantees of MINRES and CG for linear systems involving symmetric positive-definite matrices. Indeed, while having identical convergence rates, the former is stated in terms of the reduction in the norm of the residual of the iterates, while the latter indicates the reduction in the error of the iterates themselves. Also, recall that MINRES, unlike CG, can be applied beyond positive-definite matrices. In the same spirit, Newton-MR, unlike Newton-CG, is applicable beyond the traditional convex settings to invex functions.

Similarly as in the case of many Newton-type methods, we can obtain local convergence guarantees for Newton-MR with unit step-size, i.e., $\alpha_k = 1$. We further show that such result greatly generalizes the classical analysis of Newton-CG. We will show that local (super-)linear rate of convergence is possible under either Assumption 2 with $\beta > 1$, or Assumption 5.

Assumption 5. For some $0 < \beta < \infty$ and $0 \leq L_{\mathbf{H}} < \infty$, we have $\forall \mathbf{x} \in \mathbb{R}^d$, $\forall \mathbf{p} \in \text{Range}(\mathbf{H}(\mathbf{x}))$,

$$\left\langle \mathbf{g}(\mathbf{x} + \mathbf{p}), \left(\mathbf{H}(\mathbf{x} + \mathbf{p}) - \mathbf{H}(\mathbf{x}) \right) \mathbf{p} \right\rangle \leq L_{\mathbf{H}} \|\mathbf{g}(\mathbf{x} + \mathbf{p})\| \|\mathbf{p}\|^{1+\beta}. \tag{24}$$

Although, we do not know of a particular way to, a priori, verify Assumption 5, it is easy to see that (24) with $\beta = 1$ is implied by (11b), and hence weaker. In fact, unlike the usual local convergence analysis of Newton-type methods, analyzing iterations in terms of gradient norm allows us to weaken (11b) and instead consider (24).

Theorem 2 (*Error recursion of Algorithm 1 with $\alpha_k = 1$*). *Consider Assumptions 1, 3, and 4. Suppose $\mathbf{x}_k \in \mathcal{X}_0$ and consider one iteration of Algorithm 1 with $\alpha_k = 1$.*

(i) If Assumption 2 holds, then

$$\|\mathbf{g}_{k+1}\|^2 \leq \frac{2L(\mathbf{x}_0)}{(1+\beta)[\gamma(\mathbf{x}_0)]^{1+\beta}} \|\mathbf{g}_k\|^{1+\beta} + (1 - 2\nu(\mathbf{x}_0)) \|\mathbf{g}_k\|^2.$$

(ii) If Assumption 5 holds and $\mathbf{H}(\mathbf{x})$ is continuous, then

$$\|\mathbf{g}_{k+1}\| \leq \frac{L_{\mathbf{H}}}{(1+\beta)[\gamma(\mathbf{x}_0)]^{1+\beta}} \|\mathbf{g}_k\|^{1+\beta} + \sqrt{1 - \nu(\mathbf{x}_0)} \|\mathbf{g}_k\|.$$

Here, $L(\mathbf{x}_0), \gamma(\mathbf{x}_0), \nu(\mathbf{x}_0)$ and $L_{\mathbf{H}}$ are defined, respectively, in Assumptions 2, 3, 4, and 5. Also, β refers to the respective constants of Assumptions 2 and 5.

Proof. Recall that (20) and $\alpha_k = 1$ implies that $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{p}_k$, where we have $\mathbf{p}_k = -[\mathbf{H}(\mathbf{x}_k)]^{\dagger} \mathbf{g}(\mathbf{x}_k)$. Throughout the proof, let $\mathbf{U}_{\mathbf{x}}$ and $\mathbf{U}_{\mathbf{x}}^{\perp}$ denote any orthogonal bases

for Range($\mathbf{H}(\mathbf{x})$) and its orthogonal complement, respectively. (i) From (22) with $\alpha_k = 1$ and using Assumptions 3 and 4, we get

$$\|\mathbf{g}_{k+1}\|^2 \le \|\mathbf{g}_k\|^2 - 2\|\mathbf{U}_k^\mathsf{T}\mathbf{g}_k\|^2 + \frac{2L(\mathbf{x}_0)}{(\beta+1)[\gamma(\mathbf{x}_0)]^{\beta+1}}\|\mathbf{g}_k\|^{\beta+1}$$

$$\le (1 - 2\nu(\mathbf{x}_0))\|\mathbf{g}_k\|^2 + \frac{2L(\mathbf{x}_0)}{(1+\beta)[\gamma(\mathbf{x}_0)]^{1+\beta}}\|\mathbf{g}_k\|^{1+\beta}.$$

(ii) Since by assumption $\mathbf{g}(\mathbf{x})$ is continuously differentiable, using mean-value theorem for vector-valued functions ([31, Theorem 7.9-1(d)]) for $\mathbf{g}_{k+1} = \mathbf{g}(\mathbf{x}_k + \mathbf{p}_k)$, we have

$$\|\mathbf{g}_{k+1}\|^2 = \langle \mathbf{g}_{k+1}, \mathbf{g}_{k+1}\rangle = \left\langle \mathbf{g}_{k+1}, \mathbf{g}_k + \int_0^1 [\mathbf{H}(\mathbf{x}_k + t\mathbf{p}_k)\mathbf{p}_k]\,\mathrm{d}t\right\rangle. \qquad (25)$$

Note that

$$\mathbf{g}(\mathbf{x}_k) = \mathbf{H}(\mathbf{x}_k)[\mathbf{H}(\mathbf{x}_k)]^\dagger \mathbf{g}(\mathbf{x}_k) + \mathbf{U}_\mathbf{x}^\perp [\mathbf{U}_\mathbf{x}^\perp]^\mathsf{T} \mathbf{g}(\mathbf{x}_k) = -\mathbf{H}(\mathbf{x}_k)\mathbf{p}_k + \mathbf{U}_\mathbf{x}^\perp [\mathbf{U}_\mathbf{x}^\perp]^\mathsf{T}\mathbf{g}(\mathbf{x}_k).$$

Hence, it follows that

$$\|\mathbf{g}_{k+1}\|^2 = \left\langle \mathbf{g}_{k+1}, -\mathbf{H}(\mathbf{x}_k)\mathbf{p}_k + \int_0^1 [\mathbf{H}(\mathbf{x}_k + t\mathbf{p}_k)\mathbf{p}_k]\,\mathrm{d}t\right\rangle + \left\langle \mathbf{g}_{k+1}, \mathbf{U}_\mathbf{x}^\perp [\mathbf{U}_\mathbf{x}^\perp]^\mathsf{T}\mathbf{g}(\mathbf{x}_k)\right\rangle$$

$$\le \left\langle \mathbf{g}_{k+1}, \int_0^1 [(\mathbf{H}(\mathbf{x}_k + t\mathbf{p}_k) - \mathbf{H}(\mathbf{x}_k))\mathbf{p}_k]\,\mathrm{d}t\right\rangle + \|\mathbf{g}_{k+1}\|\left\|[\mathbf{U}_\mathbf{x}^\perp]^\mathsf{T}\mathbf{g}(\mathbf{x}_k)\right\|$$

$$\le \int_0^1 t^{-1} [\langle \mathbf{g}_{k+1}, (\mathbf{H}(\mathbf{x}_k + t\mathbf{p}_k) - \mathbf{H}(\mathbf{x}_k))t\mathbf{p}_k\rangle]\,\mathrm{d}t + \sqrt{1-\nu(\mathbf{x}_0)}\|\mathbf{g}_{k+1}\|\|\mathbf{g}_k\|,$$

where the inequality follows by Lemma 4. Using Assumptions 3 and 5, we get

$$\|\mathbf{g}_{k+1}\|^2 \le L_\mathbf{H}\|\mathbf{g}_{k+1}\|\|\mathbf{p}_k\|^{1+\beta}\int_0^1 t^\beta \mathrm{d}t + \sqrt{1-\nu(\mathbf{x}_0)}\|\mathbf{g}_{k+1}\|\|\mathbf{g}_k\|,$$

and hence,

$$\|\mathbf{g}_{k+1}\| \le \frac{L_\mathbf{H}}{(1+\beta)}\|\mathbf{p}_k\|^{1+\beta} + \sqrt{1-\nu(\mathbf{x}_0)}\|\mathbf{g}_k\|$$

$$\le \frac{L_\mathbf{H}}{(1+\beta)}\left\|[\mathbf{H}_k]^\dagger \mathbf{g}_k\right\|^{1+\beta} + \sqrt{1-\nu(\mathbf{x}_0)}\|\mathbf{g}_k\|$$

$$\le \frac{L_\mathbf{H}}{(1+\beta)[\gamma(\mathbf{x}_0)]^{1+\beta}}\|\mathbf{g}_k\|^{1+\beta} + \sqrt{1-\nu(\mathbf{x}_0)}\|\mathbf{g}_k\|. \qquad \square$$

Remark 3. Theorem 2-(ii) implies convergence in gradient norm with a local rate that is linear when $\nu(\mathbf{x}_0) < 1$, super-linear when $\nu(\mathbf{x}_0) = 1$, and quadratic when $\nu(\mathbf{x}_0) = \beta = 1$. For example, when $\nu(\mathbf{x}_0) < 1$, for any given $\sqrt{1 - \nu(\mathbf{x}_0)} < c < 1$, if $\|\mathbf{g}_k\| \leq \left((c - \sqrt{1 - \nu(\mathbf{x}_0)})(1 + \beta)[\gamma(\mathbf{x}_0)]^{1+\beta}/L_{\mathbf{H}} \right)^{1/\beta}$, we get $\|\mathbf{g}_{k+1}\| \leq c\|\mathbf{g}_k\|$. In other words, the local convergence rate is *problem-independent*, which is a similar characteristic to that of exact Newton-CG; see [81]. If $\nu(\mathbf{x}_0) = \beta = 1$, then Theorem 2-(ii) implies that

$$\|\mathbf{g}_{k+1}\| \leq \frac{L_{\mathbf{H}}}{2\gamma(\mathbf{x}_0)^2} \|\mathbf{g}_k\|^2, \tag{26}$$

which greatly resembles the quadratic convergence of Newton-CG, but in terms of $\|\mathbf{g}_k\|$ in lieu of $\|\mathbf{x}_k - \mathbf{x}^\star\|$. For strongly-convex objectives, (26) coincides exactly with the well-known bound on $\|\mathbf{g}\|$ in the literature, e.g., see [18, Eqn. (9.33)]. For comparison, the local convergence rate of the Newton-type method proposed in [68] for strongly-convex objectives is superlinear.

3.2.2. Inexact update

Clearly, in almost all practical application, it is rather unreasonable to assume that (20) can be solved exactly. Approximations to (20), in the similar literature, are typically done by requiring

$$\|\mathbf{H}_k\mathbf{p}_k + \mathbf{g}_k\| \leq \theta \|\mathbf{g}_k\|, \tag{27}$$

for some appropriate $0 \leq \theta < 1$. In other words, we can simply require that an approximate solution $\mathbf{p}_k$ is, at least, better than "$\mathbf{p} = \mathbf{0}$" by some factor. Such inexactness conditions have long been used in the analysis of Newton-CG, e.g., see [17,21,76,81]. However, Newton-MR allows for further relaxation of this condition. Indeed, as we will later see in this section, we can further loosen (27) by merely requiring that an approximate solution satisfies

$$\langle \mathbf{H}_k\mathbf{p}_k, \mathbf{g}_k \rangle \leq -\frac{(1 - \theta)}{2} \|\mathbf{g}_k\|^2. \tag{28}$$

It is easy to see that (27) implies (28). In our theoretical analysis below, we will employ (15) and hence we need to ensure that for the update directions, $\mathbf{p}_k$, we have $\mathbf{p}_k \in \text{Range}(\mathbf{H}_k)$. In this light, we introduce the following relaxation of (27)

$$\text{Find} \quad \mathbf{p} \in \text{Range}(\mathbf{H}_k) \quad \text{s.t.} \quad \mathbf{p} \text{ satisfies } (28). \tag{29}$$

The inexactness condition in (29) involves two criteria for an approximate solution $\mathbf{p}$, namely feasibility of $\mathbf{p}$ in (28) and that $\mathbf{p} \in \text{Range}(\mathbf{H}_k)$. To seamlessly enforce the latter, recall that the tth iteration of MINRES-QLP can be described as follows [29, Table 5.1 and Eqn (3.1)]

$$\mathbf{p}^{(t)} = \arg\min_{\mathbf{p} \in \mathbb{R}^d} \|\mathbf{p}\|, \quad \text{subject to} \quad \mathbf{p} \in \underset{\widehat{\mathbf{p}} \in \mathcal{K}_t(\mathbf{H}, \mathbf{g})}{\text{Arg}\min} \|\mathbf{H}\widehat{\mathbf{p}} + \mathbf{g}\|, \tag{30}$$

where $\mathcal{K}_t(\mathbf{A}, \mathbf{b}) = \text{Span}\left\{\mathbf{b}, \mathbf{A}\mathbf{b}, \ldots, \mathbf{A}^{t-1}\mathbf{b}\right\}$ denotes the tth-Krylov sub-space generated by $\mathbf{A}$ and $\mathbf{b}$ with $t \leq r \triangleq \text{rank}(\mathbf{A})$. If $\nu(\mathbf{x}_0) = 1$ in Assumption 4, then we necessarily have $\mathbf{g} \in \text{Range}(\mathbf{H})$, and hence $\mathbf{p}^{(t)} \in \text{Range}(\mathbf{H}), \forall t$. Otherwise, when $\nu(\mathbf{x}_0) < 1$ in Assumption 4, which implies $\mathbf{g} \notin \text{Range}(\mathbf{H})$, we cannot necessarily expect to have $\mathbf{p}^{(t)} \in \text{Range}(\mathbf{H})$. However, one can easily remedy this by modifying MINRES-QLP iterations to incorporate $\mathcal{K}_t(\mathbf{H}, \mathbf{Hg})$ instead of $\mathcal{K}_t(\mathbf{H}, \mathbf{g})$. Compared with (30), this essentially boils down to performing one additional matrix-vector product to compute the vector $\mathbf{Hg}$, which is then normalized and used as the initial vector within the Lanczos process; see [22,54] for similar modifications applied to MINRES. Clearly, this will not change the guarantees of MINRES-QLP regarding the monotonicity of the residuals as well as the final solution at termination, i.e., $\mathbf{p}^\dagger = -[\mathbf{H}]^\dagger \mathbf{g}$. Since for all t, we have $\mathbf{p}_t \in \text{Range}(\mathbf{H})$, it follows that any feasible $\mathbf{p}_t$ satisfies (29). We note that, in general, the Krylov subspace methods that define their iterates in $\mathbf{H} \cdot \mathcal{K}_t(\mathbf{H}, \mathbf{g})$, as opposed to $\mathcal{K}_t(\mathbf{H}, \mathbf{g})$, can have a slower convergence, e.g., see [43]. However, in our experience, such a side-effect can be negligible given the relatively small number of iterations that is often required to satisfy the inexactness condition (28), as opposed to (27); see also Remark 4.

For the analysis of Newton-type methods, to the best of our knowledge, (29) has never been considered before and constitutes the most relaxed inexactness condition on the sub-problems in the similar literature.

Remark 4 *((28) is more relaxed than (27))*. For a fixed θ, the relative residual of any solution to (28) is often much larger than that required by (27). Indeed, from (30) and [29, Lemma 3.3], we have $\langle \mathbf{p}_k, \mathbf{H}_k (\mathbf{H}_k\mathbf{p}_k + \mathbf{g}_k)\rangle = 0$. Now, from (28) we get

$$(1 + \theta) \|\mathbf{g}_k\|^2$$
$$\geq 2 \langle \mathbf{g}_k, \mathbf{H}_k\mathbf{p}_k + \mathbf{g}_k\rangle \qquad\qquad\qquad\quad (\text{From (28)})$$
$$= 2 \langle \mathbf{g}_k, \mathbf{H}_k\mathbf{p}_k + \mathbf{g}_k\rangle + 2\langle \mathbf{H}_k\mathbf{p}_k, (\mathbf{H}_k\mathbf{p}_k + \mathbf{g}_k)\rangle \qquad (\langle \mathbf{p}_k, \mathbf{H}_k (\mathbf{H}_k\mathbf{p}_k + \mathbf{g}_k)\rangle = 0)$$
$$= 2 \langle \mathbf{g}_k + \mathbf{H}_k\mathbf{p}_k, \mathbf{H}_k\mathbf{p}_k + \mathbf{g}_k\rangle = 2\|\mathbf{H}_k\mathbf{p}_k + \mathbf{g}_k\|^2,$$

which implies

$$\|\mathbf{H}_k\mathbf{p}_k + \mathbf{g}_k\| \leq \sqrt{\frac{1+\theta}{2}} \|\mathbf{g}_k\|. \tag{31}$$

Hence, (28) is equivalent to requiring the relative residual condition albeit with the tolerance of $\sqrt{(1+\theta)/2}$. This in turn implies that for a given θ, we can satisfy (29) in roughly less than *half* as many iterations as what is needed to ensure the more stringent (27).

Algorithm 2 Inexact Newton-MR.

Input:
- Initial iterate $\mathbf{x}_0 \in \mathbb{R}^d$, Inexactness tolerance θ as in Condition 1, Line-search parameter $0 < \rho < 1$, Termination tolerance $0 < \epsilon$

for $k = 0, 1, 2, \cdots$ until $\|\mathbf{g}_k\| \leq \epsilon$ **do**
    Find $\mathbf{p}_k$ using (29) with inexactness tolerance θ
    Find step-size, α_k, such that (7) holds with ρ
    Set $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k$
end for
Output: x for which $\|\mathbf{g}(\mathbf{x})\| \leq \epsilon$

The inexactness tolerance has to be chosen with regards to Assumption 4. Indeed, under Assumption 4, for the exact solution, $\mathbf{p}_k = -[\mathbf{H}_k]^\dagger \mathbf{g}_k$, we have $\langle \mathbf{H}_k \mathbf{p}_k, \mathbf{g}_k \rangle = -\langle \mathbf{H}_k [\mathbf{H}_k]^\dagger \mathbf{g}_k, \mathbf{g}_k \rangle \leq -\nu(\mathbf{x}_0) \|\mathbf{g}_k\|^2$. Hence, it suffices to choose θ in (28) such that $\theta \geq 1 - 2\nu(\mathbf{x}_0)$.

Condition 1 (*Inexactness tolerance* θ). *The inexactness tolerance in,* θ, *in* (28) *is chosen such that* $\theta \in [1 - 2\nu(\mathbf{x}_0), 1)$, *where* $\nu(\mathbf{x}_0)$ *is as in Assumption 4.*

Of course, because of $\nu(\mathbf{x}_0)$, the interval in Condition 1 also depends on $\mathbf{x}_0$, which consequently affect the range of possible choices for θ. In this sense, the proper notation for inexactness tolerance is $\theta(\mathbf{x}_0)$. However, to simplify our results, we drop the dependence of $\theta(\mathbf{x}_0)$ on $\mathbf{x}_0$. Also, note that if $\nu(\mathbf{x}_0) > 1/2$, then we can take θ to be negative.

Replacing the exact update (20) in Algorithm 1 with the inexact condition (29) and considering Condition 1, we get an inexact variant of Newton-MR, depicted in Algorithm 2, and Theorem 3 provides its convergence properties.

Theorem 3 (*Convergence of Algorithm 2*). *Consider Assumptions 1 to 4. For the iterates of Algorithm 2, we have*

$$\|\mathbf{g}_{k+1}\|^2 \leq \left(1 - 2\rho\widehat{\tau}(\mathbf{x}_0) \|\mathbf{g}_k\|^{(1-\beta)/\beta}\right) \|\mathbf{g}_k\|^2 , \tag{32}$$

where $\widehat{\tau}(\mathbf{x}_0) \triangleq \tau(\mathbf{x}_0) [(1 - \theta) / (2\nu(\mathbf{x}_0))]^{(1+\beta)/\beta}$, *and* ρ *is the line-search parameter of Algorithm 2,* $\nu(\mathbf{x}_0)$ *is defined in Assumption 4,* $\tau(\mathbf{x}_0)$ *is as defined in Theorem 1 and* θ *is given in Condition 1.*

Proof. Similar to the proof of Theorem 1, we get (22). By [30, Lemma 3.3 and Section 6.6], we know that $\|\mathbf{H}_k \mathbf{p}_k\|$ is monotonically non-decreasing in MINRES-type solvers, i.e., $\|\mathbf{H}_k \mathbf{p}_k\| \leq \left\|\mathbf{H}_k [\mathbf{H}_k]^\dagger \mathbf{g}_k\right\| \leq \|\mathbf{g}_k\|$. Since $\mathbf{p}_k \in \text{Range}(\mathbf{H}_k)$, from Assumption 3 it follows that

$$\|\mathbf{p}_k\| \leq \frac{1}{\gamma(\mathbf{x}_0)} \|\mathbf{H}_k \mathbf{p}_k\| \leq \frac{1}{\gamma(\mathbf{x}_0)} \|\mathbf{g}_k\| . \tag{33}$$

This, in turn, implies that

$$\left\|\mathbf{g}_{k+1}\right\|^2 \le \left\|\mathbf{g}_k\right\|^2 + 2\alpha \left\langle \mathbf{H}_k \mathbf{g}_k, \mathbf{p}_k \right\rangle + \frac{2L(\mathbf{x}_0)\alpha^{\beta+1}}{(\beta+1)[\gamma(\mathbf{x}_0)]^{\beta+1}} \left\|\mathbf{g}_k\right\|^{\beta+1} .$$

Now, as in the proof of Theorem 1, to obtain a lower bound on the step-size returned from the line-search, using the above and (7), we consider α satisfying $2L(\mathbf{x}_0)\alpha^{\beta+1} \left\|\mathbf{g}_k\right\|^{\beta+1} \le (\beta+1)[\gamma(\mathbf{x}_0)]^{\beta+1}\alpha(1-\rho)(1-\theta) \left\|\mathbf{g}_k\right\|^2$. This, in turn, implies that the step-size returned from the line-search (7) must be such that $\alpha \ge \left((1-\theta)(1-\rho)(1+\beta)[\gamma(\mathbf{x}_0)]^{1+\beta}/(2L(\mathbf{x}_0))\right)^{1/\beta} \left\|\mathbf{g}_k\right\|^{(1-\beta)/\beta}$. With this lower-bound on the step-size, we obtain the desired result by noting that $\left\|\mathbf{g}_k\right\|^2 + 2\rho\alpha \left\langle \mathbf{p}_k, \mathbf{H}_k \mathbf{g}_k \right\rangle \le (1 - \rho\alpha(1-\theta)) \left\|\mathbf{g}_k\right\|^2$.   □

Recalling that $1 - \theta \le 2\nu(\mathbf{x}_0)$ from Condition 1, we can obtain a bound similar to (23) with $\widehat{\tau}(\mathbf{x}_0)$ replacing $\tau(\mathbf{x}_0)$, for the rate given by Theorem 3.

Remark 5. For $\theta = 1 - 2\nu(\mathbf{x}_0)$, i.e., when sub-problems are solved exactly, Theorem 3 coincides with Theorem 1. More generally, for any $\nu(\mathbf{x}_0) \in (0, 1]$ and $\theta \in [1 - 2\nu(\mathbf{x}_0), 1)$, we get a rate similar, up to a constant factor of $(1-\theta)/(2\nu(\mathbf{x}_0))$, to that of the exact update in Theorem 1, i.e., *the effects of the problem-related quantities such as $L(\mathbf{x}_0)$ and $\gamma(\mathbf{x}_0)$ remain the same*. This is in contrast to Newton-CG, for which in order to obtain a rate similar to that of the exact algorithm, one has to solve the linear system to a high enough accuracy, i.e., $\theta \le \sqrt{\kappa}$, where κ is the condition number of the problem. Otherwise, the dependence of the convergence rate on the problem-related quantities is significantly worsened; see [81, Theorem 2].

Similar local convergence results as in Theorem 2 can also be obtained for the case where the update direction, $\mathbf{p}_k$, is obtained approximately. Note that, again as in Remark 5, when $\theta = 1 - 2\nu(\mathbf{x}_0)$, the results of Theorem 4 coincide with those of Theorem 2.

Theorem 4 (*Error recursion of Algorithm 2 with $\alpha_k = 1$*). *Under the same assumptions of Theorem 2, suppose $\mathbf{x}_k \in \mathcal{X}_0$, and consider one iteration of Algorithm 2 with $\alpha_k = 1$.*

(i) If Assumption 2 holds, then

$$\left\|\mathbf{g}_{k+1}\right\|^2 \le \frac{2L(\mathbf{x}_0)}{(1+\beta)[\gamma(\mathbf{x}_0)]^{1+\beta}} \left\|\mathbf{g}_k\right\|^{1+\beta} + \theta \left\|\mathbf{g}_k\right\|^2 .$$

(ii) If Assumption 5 holds and $\mathbf{H}(\mathbf{x})$ is continuous, then

$$\left\|\mathbf{g}_{k+1}\right\| \le \frac{L_{\mathbf{H}}}{(1+\beta)[\gamma(\mathbf{x}_0)]^{1+\beta}} \left\|\mathbf{g}_k\right\|^{1+\beta} + \sqrt{\frac{1+\theta}{2}} \left\|\mathbf{g}_k\right\| .$$

Here, $L(\mathbf{x}_0), \gamma(\mathbf{x}_0)$ and $L_{\mathbf{H}}$ are defined, respectively, in Assumptions 2, 3, and 5, β refers to the respective constants of Assumptions 2 and 5, and θ is an inexactness tolerance chosen according to Condition 1.

Proof. The proof is very similar to that of Theorem 2. (i) This follows immediately from (22) with $\alpha_k = 1$ and using (28) coupled with (33). (ii) We first replace $\mathbf{g}_k = -\mathbf{H}_k \mathbf{p}_k + (\mathbf{H}_k \mathbf{p}_k + \mathbf{g}_k)$ in (25). As in the proof Theorem 2-(ii), from (31), we get $\|\mathbf{g}_{k+1}\| \leq L_{\mathbf{H}} \|\mathbf{p}_k\|^{1+\beta} / (1+\beta) + \sqrt{(1+\theta)/2} \|\mathbf{g}_k\|$, which using (33) gives the result. $\square$

Remark 6. Here, as in Remark 3, one can obtain local linear convergence rate that is problem-independent. For example, from Theorem 4-(ii), it follows that if $\sqrt{(1+\theta)/2} < c < 1$ and the gradient is small enough, we get $\|\mathbf{g}_{k+1}\| \leq c \|\mathbf{g}_k\|$; see [58, Section 6] for a similar statement in the context of more general nonlinear equations. This implies that local convergence in the norm of the gradient is very fast even with a very crude solution of the sub-problem. We also note that if $\nu(\mathbf{x}_0) = \beta = 1$, and $\theta_k \in \mathcal{O}(\|\mathbf{g}_k\|^2) - 1$, then Theorem 4-(ii) implies a quadratic convergence rate.

Finally, we can obtain iteration complexity to find a solution satisfying $\|\mathbf{g}(\mathbf{x})\| \leq \epsilon$ for a desired $\epsilon > 0$.

Corollary 1 *(Iteration complexity of Algorithm 2). Under the same assumptions as in Theorem 3, consider finding an iterate for which $\|\mathbf{g}_k\| \leq \epsilon$ for a desired $\epsilon > 0$. Then, we have the following two cases:*

(i) With $\beta \geq 1$ in Assumption 2 then $k \in \mathcal{O}(\log(1/\epsilon))$.
(ii) With $0 < \beta < 1$ in Assumption 2, then $k \in \mathcal{O}\left(\epsilon^{(\beta-1)/\beta} \log(1/\epsilon)\right)$.

Proof. Recall that we always have $\|\mathbf{g}_k\| \leq \|\mathbf{g}^{(k-1)}\| \leq \ldots \leq \|\mathbf{g}_0\|$. Dropping the dependence of $\widehat{\tau}(\mathbf{x}_0)$ on $\mathbf{x}_0$, we have the following.

(i) For $\beta \geq 1$, from (32), we have $\|\mathbf{g}_k\|^2 \leq \left(1 - 2\rho\widehat{\tau} \|\mathbf{g}_0\|^{(1-\beta)/\beta}\right)^k \|\mathbf{g}_0\|^2$, which, in order to obtain $\|\mathbf{g}_k\|^2 \leq \epsilon$, implies that we must have

$$k \geq \log\left(\epsilon/\|\mathbf{g}_0\|\right) / \log\left(1 - 2\rho\widehat{\tau} \|\mathbf{g}_0\|^{(1-\beta)/\beta}\right).$$

Now, noting that $-\log(1 - 1/x) \in \mathcal{O}(1/x)$, we obtain the result.

(ii) For $\beta < 1$, as long as $\|\mathbf{g}_k\|^2 \geq \epsilon$, from (32), we have

$$\|\mathbf{g}_k\|^2 \leq \left(1 - 2\rho\widehat{\tau} \|\mathbf{g}_k\|^{(1-\beta)/\beta}\right)^k \|\mathbf{g}_0\|^2 \leq \left(1 - 2\rho\widehat{\tau}\epsilon^{(1-\beta)/\beta}\right)^k \|\mathbf{g}_0\|^2.$$

Now, we obtain the result as above. $\square$

Corollary 1 implies that (inexact) Newton-MR is guaranteed to converge for functions, which traditionally have not been considered suitable candidates for Newton-type methods. For example, Algorithm 2 can still be applied for optimization of a twice continuously differentiable objective for which we have $\beta \ll 1$ in Assumption 2. Such functions can be extremely non-smooth, and as a consequence, they have been labeled, rather inaccurately, as cases where the application of curvature might be useless, e.g., [4,5].

3.2.3. Convergence under generalized Polyak-Łojasiewicz inequality

The set of global optima of an invex function, $f(\mathbf{x})$, is characterized by the zeros of the gradient, $\mathbf{g}(\mathbf{x})$. If, in addition, the distance to the optimal set, in terms of iterates and/or their respective function values, is also somehow related to $\mathbf{g}(\mathbf{x})$, then one can obtain rates at which the iterates and/or their objective values approach optimality. In this section, we consider an important sub-class of invex problems, which allows us to do that.

Definition 2 *(Generalized Polyak-Łojasiewicz inequality).* Let $\mathcal{X} \subseteq \mathbb{R}^d$ be an open set. A differentiable function f on $\mathcal{X}$ is said to satisfy the generalized Polyak-Łojasiewicz (GPL) inequality on $\mathcal{X}$ if there exist $1 < \eta < \infty$ and $0 < \mu < \infty$ such that

$$f(\mathbf{x}) - \inf_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) \leq \left(\frac{1}{\mu} \|\mathbf{g}(\mathbf{x})\|^{\eta} \right)^{1/(\eta-1)}, \qquad \forall \mathbf{x} \in \mathcal{X}. \tag{34}$$

The class of functions satisfying (38) is denoted by $\mathcal{F}_{\eta,\mu}^{GPL}$.

It is clear that $\mathcal{F}_{\eta,\mu}^{GPL} \subset \mathcal{F}$ (cf. Definition 1). Most often in the literature, Polyak-Łojasiewicz (PL) inequality is referred to as (34) with $\eta = 2$, e.g., [57], which excludes many functions. The generalized notion of Polyak-Łojasiewicz in (34) with any $1 < \eta < \infty$ encompasses many of such functions. For example, the weakly convex function $f(x) = x^4$ clearly violates the typical PL (i.e., with $\eta = 2$), but it indeed satisfies (34) with $\eta = 4$ and $\mu = 256$. In fact, any polynomial function of the form $f(x) = \sum_{i=1}^{p} a_i x^{2i}$ with $a_i \geq 0, i = 1, \ldots, p$ satisfies (34).

But "how large is the class $\mathcal{F}_{\eta,\mu}^{GPL}$ in Definition 2"? We aim to shed light on this question by considering other classes of functions that are, in some sense, equivalent to $\mathcal{F}_{\eta,\mu}^{GPL}$. In doing so, we draw similarities from various relaxations of strong convexity. More specifically, to alleviate the restrictions imposed by making strong convexity assumptions, several authors have introduced relaxations under which desirable convergence guarantees of various algorithms are maintained; e.g., the *quadratic growth condition* [3], the *restricted secant inequality* [89], and the *error bounds* [65]. The relationships among these classes of functions have also been established; see [57,72,83,98,99]. We now give natural extensions of these conditions to invex functions and show that $\mathcal{F}_{\eta,\mu}^{GPL}$ is an equivalent class of functions.

Let us define the set of optimal points of the problem (1) as

$$\mathcal{X}^\star \triangleq \left\{ \mathbf{x}^\star \in \mathbb{R}^d \mid f(\mathbf{x}^\star) \leq f(\mathbf{x}), \ \forall \mathbf{x} \in \mathbb{R}^d \right\}. \tag{35}$$

Further, assume that $\mathcal{X}^\star$ is non-empty, and denote the optimal value of (1) by $f^\star$. Recall that when f is invex, $\mathcal{X}^\star$ need not be a convex set, but it is clearly closed; see [69, p. 14].

Definition 3 *(Generalized functional growth property).* A differentiable function f is said to satisfy the generalized functional growth (GFG) property if there exist $1 < \eta < \infty$, $0 < \mu < \infty$ and a vector-valued function $\boldsymbol{\phi} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$, such that

$$f(\mathbf{x}) - f^\star \geq \mu \min_{\mathbf{y} \in \mathcal{X}^\star} \|\boldsymbol{\phi}(\mathbf{y}, \mathbf{x})\|^\eta, \qquad \forall \mathbf{x} \in \mathbb{R}^d. \tag{36}$$

The class of functions satisfying (36) with the same $\boldsymbol{\phi}$ is denoted by $\mathcal{F}_{\boldsymbol{\phi}, \eta, \mu}^{GFG}$.

Definition 4 *(Generalized restricted secant inequality).* For a vector-valued function $\boldsymbol{\phi} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ define $\mathcal{Y}_{\boldsymbol{\phi}}^\star(\mathbf{x}) \triangleq \{\mathbf{y}^\star \in \mathcal{X}^\star \mid \|\boldsymbol{\phi}(\mathbf{y}^\star, \mathbf{x})\| \leq \|\boldsymbol{\phi}(\mathbf{y}, \mathbf{x})\|, \ \forall \mathbf{y} \in \mathcal{X}^\star\}$. A differentiable function f is said to satisfy the generalized restricted secant (GRS) inequality if there exist $1 < \eta < \infty$, $0 < \mu < \infty$ and a mapping $\boldsymbol{\phi} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$, such that

$$\min_{\mathbf{y} \in \mathcal{Y}_{\boldsymbol{\phi}}^\star(\mathbf{x})} \langle -\mathbf{g}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{y}, \mathbf{x}) \rangle \geq \mu \min_{\mathbf{y} \in \mathcal{X}^\star} \|\boldsymbol{\phi}(\mathbf{y}, \mathbf{x})\|^\eta, \qquad \forall \mathbf{x} \in \mathbb{R}^d. \tag{37}$$

The class of functions satisfying (37) with the same $\boldsymbol{\phi}$ is denoted by $\mathcal{F}_{\boldsymbol{\phi}, \eta, \mu}^{GRS}$.

It is easy to see that when f is convex and $\boldsymbol{\phi}(\mathbf{y}, \mathbf{x}) = \mathbf{y} - \mathbf{x}$, we have $\mathcal{Y}_{\boldsymbol{\phi}}^\star(\mathbf{x}) = \{[\mathbf{x}]_{\mathcal{X}^\star}\}$, where $[\mathbf{x}]_{\mathcal{X}^\star} \triangleq \arg\min_{\mathbf{y} \in \mathcal{X}^\star} \|\mathbf{y} - \mathbf{x}\|$ is the unique orthogonal projection of $\mathbf{x}$ onto the set of optimal solutions $\mathcal{X}^\star$ (which, in this case, is convex). Hence, the generalized condition (37) coincides with, and hence is a generalization of, the usual definition of restricted secant inequality for convex functions with $\eta = 2$; see [57,83,98,99].

Definition 5 *(Generalized error bound property).* A differentiable function f is said to satisfy the generalized error bound (GEB) property if there exist $1 < \eta < \infty$, $0 < \mu < \infty$ and a vector-valued function $\boldsymbol{\phi} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$, such that

$$\|\mathbf{g}(\mathbf{x})\| \geq \mu \min_{\mathbf{y} \in \mathcal{X}^\star} \|\boldsymbol{\phi}(\mathbf{y}, \mathbf{x})\|^{\eta - 1}, \qquad \forall \mathbf{x} \in \mathbb{R}^d. \tag{38}$$

The class of functions satisfying (38) with the same $\boldsymbol{\phi}$ is denoted by $\mathcal{F}_{\boldsymbol{\phi}, \eta, \mu}^{GEB}$.

Lemma 6 establishes a loose notion of equivalence among the classes of functions mentioned above, as they relate to invexity. However, when restricted to a particular class

of invex functions for a given ϕ, Lemma 6 shows that $\mathcal{F}_{\eta,\mu}^{GPL} \cap \mathcal{F}_\phi$ can indeed be larger. In contrast, for when $\phi(\mathbf{y},\mathbf{x}) = \mathbf{y}-\mathbf{x}$, equivalence between these classes has been established in [57]. This is in particular so since, unlike [57], here we have made no assumptions on the smoothness of f such as Lipschitz-continuity of its gradient. This is a rather crucial distinction as for our main results in this paper, we make smoothness assumptions that are less strict that those typically found in the literature; see Assumption 2 in Section 3.1.

Lemma 6.

(i) *For any ϕ, we have*

$$\mathcal{F}_{\phi,\eta,\mu}^{GFG} \cap \mathcal{F}_\phi \subseteq \mathcal{F}_{\phi,\eta,\mu}^{GRS} \cap \mathcal{F}_\phi \subseteq \mathcal{F}_{\phi,\eta,\mu}^{GEB} \cap \mathcal{F}_\phi \subseteq \mathcal{F}_{\eta,\mu}^{GPL} \cap \mathcal{F}_\phi.$$

(ii) *There exists a $\widehat{\phi}$, for which*

$$\mathcal{F}_{\eta,\mu}^{GPL} \equiv \mathcal{F}_{\eta,\mu}^{GPL} \cap \mathcal{F}_{\widehat{\phi}} \subseteq \mathcal{F}_{\widehat{\phi},\eta,\mu}^{GFG} \cap \mathcal{F}_{\widehat{\phi}}.$$

(iii) *These classes are equivalent in the sense that*

$$\bigcup_\phi \left\{ \mathcal{F}_{\phi,\eta,\mu}^{GFG} \cap \mathcal{F}_\phi \right\} \equiv \bigcup_\phi \left\{ \mathcal{F}_{\phi,\eta,\mu}^{GRS} \cap \mathcal{F}_\phi \right\} \equiv \bigcup_\phi \left\{ \mathcal{F}_{\phi,\eta,\mu}^{GEB} \cap \mathcal{F}_\phi \right\} \equiv \mathcal{F}_{\eta,\mu}^{GPL}.$$

Proof.   **(i)** We start by showing that, for any $f \in \mathcal{F}_\phi$, (36) implies (37). Consider any $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{y}^\star \in \mathcal{Y}_\phi^\star(\mathbf{x})$. By (36) and invexity as well as noticing that $\mathcal{Y}_\phi^\star(\mathbf{x}) \subseteq \mathcal{X}^\star \Rightarrow f(\mathbf{y}^\star) = f^\star$, we have $\mu \min_{\mathbf{y}\in\mathcal{X}^\star} \|\phi(\mathbf{y},\mathbf{x})\|^\eta \le f(\mathbf{x}) - f(\mathbf{y}^\star) \le -\langle \phi(\mathbf{y}^\star,\mathbf{x}), \mathbf{g}(\mathbf{x})\rangle$. Since the last inequality holds for all $\mathbf{y}^\star \in \mathcal{Y}_\phi^\star(\mathbf{x})$, we can minimize the right-hand side over all $\mathcal{Y}_\phi^\star(\mathbf{x})$ to get

$$\mu \min_{\mathbf{y}\in\mathcal{X}^\star} \|\phi(\mathbf{y},\mathbf{x})\|^\eta \le \min_{\mathbf{y}\in\mathcal{Y}_\phi^\star(\mathbf{x})} \langle -\phi(\mathbf{y},\mathbf{x}), \mathbf{g}(\mathbf{x})\rangle,$$

which is exactly (37). A simple application of Cauchy-Schwarz inequality on (37) and noting that $\min_{\mathbf{y}\in\mathcal{Y}_\phi^\star(\mathbf{x})} \|\phi(\mathbf{y},\mathbf{x})\| = \min_{\mathbf{y}\in\mathcal{X}^\star} \|\phi(\mathbf{y},\mathbf{x})\|$, it follows that $\mu \min_{\mathbf{y}\in\mathcal{X}^\star} \|\phi(\mathbf{y},\mathbf{x})\|^\eta \le \min_{\mathbf{y}\in\mathcal{X}^\star} \|\phi(\mathbf{y},\mathbf{x})\| \|\mathbf{g}(\mathbf{x})\|$. If $\mathbf{x}$ is such that $\min_{\mathbf{y}\in\mathcal{X}^\star} \|\phi(\mathbf{y},\mathbf{x})\| = 0$, then the inequality (38) trivially holds, otherwise dividing both sides by $\min_{\mathbf{y}\in\mathcal{X}^\star} \|\phi(\mathbf{y},\mathbf{x})\|$ gives (38).

   To get (34) from (38), note that by invexity, for any $\mathbf{y} \in \mathcal{X}^\star$, we have $f(\mathbf{x}) - f^\star \le \|\phi(\mathbf{y},\mathbf{x})\| \|\mathbf{g}(\mathbf{x})\|$, hence $f(\mathbf{x}) - f^\star \le \min_{\mathbf{y}\in\mathcal{X}^\star} \|\phi(\mathbf{y},\mathbf{x})\| \|\mathbf{g}(\mathbf{x})\|$, which using (38) gives (34).

(ii) We will show that (34) implies that there exists a $\widehat{\phi}$ for which (36) holds. Indeed, for any invex function f, we can always define a corresponding $\widehat{\phi}$ as

$$\widehat{\phi}(\mathbf{y}, \mathbf{x}) = \begin{cases} 0, & \mathbf{g}(\mathbf{x}) = \mathbf{0} \\ \left((f(\mathbf{y}) - f(\mathbf{x}))/\|\mathbf{g}(\mathbf{x})\|^2 \right) \mathbf{g}(\mathbf{x}), & \mathbf{g}(\mathbf{x}) \neq \mathbf{0} \end{cases}.$$

Now suppose $f \in \mathcal{F}_{\eta,\mu}^{GPL}$. For any $\mathbf{x} \in \mathcal{X}^\star$, the inequality (36) trivially holds. Suppose $\mathbf{x} \notin \mathcal{X}^\star$, which implies $\mathbf{g}(\mathbf{x}) \neq \mathbf{0}$. For any $\mathbf{y} \in \mathcal{X}^\star$, we have

$$\left\| \widehat{\phi}(\mathbf{y}, \mathbf{x}) \right\|^\eta = \frac{|f(\mathbf{y}) - f(\mathbf{x})|^\eta}{\|\mathbf{g}(\mathbf{x})\|^\eta} = \frac{(f(\mathbf{x}) - f^\star)^\eta}{\|\mathbf{g}(\mathbf{x})\|^\eta} \leq \frac{1}{\mu} (f(\mathbf{x}) - f^\star),$$

where the last inequality follows since by (34), we have $(f(\mathbf{x}) - f^\star)^{(1-\eta)} \geq \mu/\|\mathbf{g}(\mathbf{x})\|^\eta, \; \forall \mathbf{x} \in \mathcal{X}$.

(iii) This isThe is a simple implication of the first two parts. $\quad \square$

If $f \in \mathcal{F}_{\eta,\mu}^{GPL}$, where $\mathcal{F}_{\eta,\mu}^{GPL}$ is the class of Generalized Polyak-Łojasiewicz (GPL) functions as defined in Definition 2, we immediately have the following convergence guarantee of Algorithm 2 in terms of objective value $f(\mathbf{x}_k)$. Similar results for Algorithm 1 can also easily be obtained.

Theorem 5 *(Convergence of Algorithm 2 under GPL inequality* (34)*). Under the same assumptions as in Theorem 3, if $f(\mathbf{x})$ satisfies GPL inequality* (34)*, there are constants $0 < C < \infty$, $0 < \zeta < 1$ and $\omega > 0$, such that for the iterates of Algorithm 2 we have the following.*

(i) *With $\beta \geq 1$ in Assumption 2, we get R-linear convergence as*

$$f(\mathbf{x}_k) - \inf_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) \leq C\zeta^k,$$

where $C = C(\mathbf{x}_0, \eta, \mu)$, and $0 < \zeta = \zeta(\mathbf{x}_0, \beta, \eta, \rho, \widehat{\tau}(\mathbf{x}_0)) < 1$.

(ii) *With $0 < \beta < 1$ in Assumption 2, we get R-sublinear convergence as*

$$f(\mathbf{x}_k) - \inf_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) \leq C \left(\frac{1}{k+1} \right)^\omega,$$

where $C = C(\mathbf{x}_0, \eta, \mu, \rho, \beta, \widehat{\tau}(\mathbf{x}_0))$, and $\omega = \omega(\eta, \beta) > 0$.

Here, ρ is the line-search parameter of Algorithm 2, η and μ are as in Definition 2, and $\widehat{\tau}(\mathbf{x}_0)$ is as in Theorem 3.

Proof. **(i)** For notational simplicity, we drop the dependence of $\widehat{\tau}(\mathbf{x}_0)$ on $\mathbf{x}_0$. Consider $\beta \geq 1$. From (32) and (34), we have

$$f(\mathbf{x}_k) - \inf_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) \leq \frac{1}{\mu^{1/(\eta-1)}} \left(\|\mathbf{g}_k\|^2 \right)^{\eta/(2(\eta-1))}$$

$$\leq \frac{1}{\mu^{1/(\eta-1)}} \left(\left(1 - 2\rho\widehat{\tau} \left\|\mathbf{g}_0\right\|^{(1-\beta)/\beta}\right)^k \left\|\mathbf{g}_0\right\|^2 \right)^{\eta/(2(\eta-1))}$$

$$= \underbrace{\frac{\left\|\mathbf{g}_0\right\|^{\eta/(\eta-1)}}{\mu^{1/(\eta-1)}}}_{C} \underbrace{\left(\left(1 - 2\rho\widehat{\tau} \left\|\mathbf{g}_0\right\|^{(1-\beta)/\beta}\right)^{\eta/(2(\eta-1))} \right)}_{\zeta}^k,$$

where $\widehat{\tau}$ is as in Theorem 3, and by (23) we also have that $0 < \zeta < 1$.

(ii) Now take $0 < \beta < 1$. From (32), we have

$$\left\|\mathbf{g}_{k+1}\right\|^2 \leq \left(1 - 2\rho\widehat{\tau} \left\|\mathbf{g}_k\right\|^{(1-\beta)/\beta}\right) \left\|\mathbf{g}_k\right\|^2 = \left\|\mathbf{g}_k\right\|^2 - 2\rho\widehat{\tau} \left\|\mathbf{g}_k\right\|^{(1+\beta)/\beta},$$

which, implies $\left\|\mathbf{g}_{k+1}\right\|^2 \leq \left\|\mathbf{g}_0\right\|^2 - 2\rho\widehat{\tau} \sum_{j=0}^{k} \left\|\mathbf{g}^j\right\|^{(1+\beta)/\beta}$. Rearranging the last inequality gives $2\rho\widehat{\tau} \sum_{j=0}^{k} \left\|\mathbf{g}^j\right\|^{(1+\beta)/\beta} \leq \left\|\mathbf{g}_0\right\|^2 - \left\|\mathbf{g}_{k+1}\right\|^2 \leq \left\|\mathbf{g}_0\right\|^2$. From (8), we get $(k+1)2\rho\widehat{\tau} \left\|\mathbf{g}^k\right\|^{(1+\beta)/\beta} \leq \left\|\mathbf{g}_0\right\|^2$, which gives

$$\left\|\mathbf{g}^k\right\| \leq \left((2\rho\widehat{\tau})^{-1} \left\|\mathbf{g}_0\right\|^2\right)^{\beta/(1+\beta)} \left(\frac{1}{k+1}\right)^{\beta/(1+\beta)}.$$

Using (34), we get

$$f(\mathbf{x}_k) - \inf_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$$

$$\leq \frac{1}{\mu^{1/(\eta-1)}} \left\|\mathbf{g}_k\right\|^{\eta/((\eta-1))}$$

$$\leq \frac{1}{\mu^{1/(\eta-1)}} \left(\left((2\rho\widehat{\tau})^{-1} \left\|\mathbf{g}_0\right\|^2\right)^{\beta/(1+\beta)} \left(\frac{1}{k+1}\right)^{\beta/(1+\beta)} \right)^{\eta/((\eta-1))}$$

$$= \underbrace{\frac{1}{\mu^{1/(\eta-1)}} \left((2\rho\widehat{\tau})^{-1} \left\|\mathbf{g}_0\right\|^2\right)^{\frac{\beta\eta}{(\eta-1)(1+\beta)}}}_{C} \left(\frac{1}{k+1}\right)^{\overbrace{\frac{\beta\eta}{(\eta-1)(1+\beta)}}^{\omega}}. \quad \square$$

The following Lemma shows that, under some assumptions, the norm of the gradient at each point can be estimated using the distance of the point to the optimality set $\mathcal{X}^\star$.

Lemma 7. *Under Assumptions 2 to 4, for any $\mathbf{x}_0 \in \mathbb{R}^d$, there exists a constant $0 \leq L_{\mathbf{g}}(\mathbf{x}_0) < \infty$, such that*

$$\left\|\mathbf{g}(\mathbf{x})\right\| \leq L_{\mathbf{g}}(\mathbf{x}_0) \min_{\mathbf{x}^\star \in \mathcal{X}^\star} \left\|\mathbf{x} - \mathbf{x}^\star\right\|, \quad \forall \mathbf{x} \in \mathcal{X}_0, \tag{39}$$

where $\mathcal{X}^\star$ is as in (35).

Proof. By Assumption 2, since $\mathcal{X}^\star \subseteq \mathcal{X}_0$, we have $\|\mathbf{H}(\mathbf{x})\mathbf{g}(\mathbf{x})\| = \|\mathbf{H}(\mathbf{x})\mathbf{g}(\mathbf{x}) - \mathbf{H}(\mathbf{x}^\star)\mathbf{g}(\mathbf{x}^\star)\| \leq L(\mathbf{x}_0)\|\mathbf{x} - \mathbf{x}^\star\|$, $\forall (\mathbf{x}^\star, \mathbf{x}) \in \mathcal{X}^\star \times \mathbb{R}^d$. Since the left-hand side of the above inequality is independent of $\mathbf{x}_0$, we get $\|\mathbf{H}(\mathbf{x})\mathbf{g}(\mathbf{x})\| \leq L\|\mathbf{x} - \mathbf{x}^\star\|$, $\forall (\mathbf{x}^\star, \mathbf{x}) \in \mathcal{X}^\star \times \mathbb{R}^d$, where $L \triangleq \inf_{\mathbf{x}_0 \in \mathbb{R}^d} L(\mathbf{x}_0)$. Now, from Assumptions 3 and 4, it follows that for any $(\mathbf{x}, \mathbf{x}^\star) \in \mathcal{X}_0 \times \mathcal{X}^\star$, we have

$$\|\mathbf{g}(\mathbf{x})\| \leq \frac{1}{\nu(\mathbf{x}_0)}\left\|[\mathbf{H}(\mathbf{x})]^\dagger \mathbf{H}(\mathbf{x})\mathbf{g}(\mathbf{x})\right\| \leq \frac{1}{\nu(\mathbf{x}_0)\gamma(\mathbf{x}_0)}\|\mathbf{H}(\mathbf{x})\mathbf{g}(\mathbf{x})\| \leq \frac{L}{\nu(\mathbf{x}_0)\gamma(\mathbf{x}_0)}\|\mathbf{x} - \mathbf{x}^\star\|.$$

The result follows by taking the minimum of the right-hand side over $\mathcal{X}^\star$. □

Using Lemma 7, from Theorem 4-(ii), with any $\mathbf{x}^\star \in \mathcal{X}^\star$, we get

$$\|\mathbf{g}_{k+1}\| \leq \frac{L_\mathbf{H}[L_\mathbf{g}(\mathbf{x}_0)]^{1+\beta}}{(1+\beta)[\gamma(\mathbf{x}_0)]^{1+\beta}}\|\mathbf{x}_k - \mathbf{x}^\star\|^{1+\beta} + \sqrt{\frac{1+\theta}{2}}L_\mathbf{g}(\mathbf{x}_0)\|\mathbf{x}_k - \mathbf{x}^\star\|,$$

where $\gamma(\mathbf{x}_0), \nu(\mathbf{x}_0), L_\mathbf{H}, \theta$, and $L_\mathbf{g}(\mathbf{x}_0)$ are the constants defined, respectively in (14), (16), (24), (28) and (39). Now, suppose we have (38) with $\phi(\mathbf{y}, \mathbf{x}) = \mathbf{y} - \mathbf{x}$ (perhaps only locally). It follows that

$$\min_{\mathbf{x}^\star \in \mathcal{X}^\star}\|\mathbf{x}_{k+1} - \mathbf{x}^\star\|^{\eta-1} \leq \frac{L_\mathbf{H}[L_\mathbf{g}(\mathbf{x}_0)]^{1+\beta}}{(1+\beta)[\gamma(\mathbf{x}_0)]^{1+\beta}\mu}\|\mathbf{x}_k - \mathbf{x}^\star\|^{1+\beta}$$
$$+ \frac{\sqrt{(1+\theta)/2}L_\mathbf{g}(\mathbf{x}_0)}{\mu}\|\mathbf{x}_k - \mathbf{x}^\star\|.$$

We can also obtain a similar inequality using Theorem 4-(i). We can gather the above in the following corollary.

Theorem 6. *Consider Assumptions 1, 3 and 4. Suppose that we have (38) with $\phi(\mathbf{y}, \mathbf{x}) = \mathbf{y} - \mathbf{x}$. For the iterates of Algorithm 2 using update directions from (29) and with $\alpha_k = 1$, we have*

$$\min_{\mathbf{x}^\star \in \mathcal{X}^\star}\|\mathbf{x}_{k+1} - \mathbf{x}^\star\|^{c_0(\eta-1)} \leq c_1 \min_{\mathbf{x}^\star \in \mathcal{X}^\star}\|\mathbf{x}_k - \mathbf{x}^\star\|^{1+\beta} + c_2 \min_{\mathbf{x}^\star \in \mathcal{X}^\star}\|\mathbf{x}_k - \mathbf{x}^\star\|^{c_0},$$

where

(i) if Assumption 2 holds, then

$$c_0 = 2, \quad c_1 = \frac{2L(\mathbf{x}_0)[L_\mathbf{g}(\mathbf{x}_0)]^{1+\beta}}{(1+\beta)[\gamma(\mathbf{x}_0)]^{1+\beta}\mu^2}, \quad c_2 = \frac{\theta L_\mathbf{g}^2(\mathbf{x}_0)}{\mu^2},$$

(ii) and if Assumption 5 holds and $\mathbf{H}(\mathbf{x})$ is continuous, then

$$c_0 = 1, \quad c_1 = \frac{L_\mathbf{H}[L_\mathbf{g}(\mathbf{x}_0)]^{1+\beta}}{(1+\beta)[\gamma(\mathbf{x}_0)]^{1+\beta}\mu}, \quad c_2 = \frac{L_\mathbf{g}(\mathbf{x}_0)\sqrt{(1+\theta)/2}}{\mu}.$$

Here, $L(\mathbf{x}_0), \gamma(\mathbf{x}_0), L_{\mathbf{H}}$ and θ are defined, respectively, in Assumptions *2, 3* and *5* and Condition *1*, $\mathcal{X}^\star$ is as in (35), $L_{\mathbf{g}}(\mathbf{x}_0)$ is as in Lemma *7*, μ and η are as in Definition *5*, and β refers to the respective constants of Assumptions *2* and *5*.

Remark 7. Assuming (38) with $\phi(\mathbf{y}, \mathbf{x}) = \mathbf{y} - \mathbf{x}$ (locally), is weaker than requiring to have isolated (local) minimum. For example, suppose we have $\nu(\mathbf{x}_0) = 1$ and $\eta = 2$. By setting $\theta_k \in \mathcal{O}\left(\min_{\mathbf{x}^\star \in \mathcal{X}^\star} \|\mathbf{x}_k - \mathbf{x}^\star\|^{2\beta}\right) - 1$, from Theorem 6-(ii), we get super-linear convergence to the *set of optimal solutions*, $\mathcal{X}^\star$. In fact, the rate is quadratic for $\beta = 1$ as a special case, which matches those from many prior works, e.g., [35,62,101]. Clearly, convergence in terms of $\min_{\mathbf{x}^\star \in \mathcal{X}^\star} \|\mathbf{x}_k - \mathbf{x}^\star\|$ relaxes the notion of isolated minimum, which is at times used for the local convergence analysis of Newton-CG. Hence, error bound conditions similar to that used in Theorem 6 have been extensively used in similar literature to establish convergence to non-isolated (local) minima [44,45,62,63,95,101, 102].

4. Numerical experiments

In this section, we evaluate the empirical performance of inexact Newton-MR (Algorithm 2) as compared with several other Newton-type optimization methods on some machine learning problems. Empirical comparisons to a variety of first-order methods are left for future work.

Optimization methods   In the following performance evaluations, we compare Newton-MR with some widely used optimization methods as listed below; see [76] for the details of each of these algorithms.

 - *Newton-CG* with Armijo line-search.
 - *L-BFGS* using strong Wolfe conditions.
 - *Nonlinear-CG* using strong Wolfe conditions. In all of experiments, among the many variants of nonlinear-CG, FR-PR variant [76, Chapter 5] performed better on almost all instances.
 - *Gauss-Newton* with Armijo line-search.
 - *Trust-region* with CG-Steihaug sub-problem solver as described in [32, Algorithms 6.1.1] and [76, Algorithm 7.2].

The main hyper-parameters for these methods, used in our simulations, can be found in Table 1. In all of our experiments, we run each method until the norm of the gradient falls below 1E-10, maximum number of iterations are reached, or the algorithm fails to make progress. The latter case is detected when Newton-CG encounters a negative curvature and its CG inner iterations are terminated, or when, for any method, the maximum number of corresponding line-search has been reached and no step-size has been accepted. These scenarios are depicted by a cross "×" on all the plots.

Table 1

Hyper-parameters used in optimization methods. The initial trial step-size for all methods, with the exception of nonlinear-CG is set to $\alpha = 1$. For nonlinear-CG, we use the strategy described in [76, Section 3.5]. The initial trust-region size is also set to one.

CG/MINRES-QLP/Steihaug inexactness tolerance	History size of L-BFGS	Armijo line-search parameter	Wolfe curvature condition parameter	Maximum line-search iterations	Trust-region parameters
0.01	20	10^{-4}	0.9 (L-BFGS) 0.1 (Nonlinear-CG)	1,000	[32, (6.1.6)]

Table 2

Complexity measure per iteration for each of the algorithms. N_s and N_l denote, respectively, the total number of iterations for the corresponding inner solver and that resulting from performing the line search or the trust-region radius update.

Newton-MR	L-BFGS	Newton-CG	Nonlinear-CG	Gauss-Newton	TR CG-Steihaug
$2+2\times N_s+2\times N_l$	$2+2\times N_l$	$2+2\times N_s+N_l$	$2+2\times N_l$	$2+2\times N_s+N_l$	$2+(2\times N_s+1)\times N_l$

Remark 8. Although Newton-CG is not meant to be used on problems where Hessian can become singular or indefinite, we run plain Newton-CG on our examples without any modifications, e.g., we do not attempt to employ the encountered negative curvature as in [82,96]. As a result, we terminate its iterations if CG fails. This is a judicious choice to highlight a significant difference between Newton-MR and Newton-CG. Specifically, this serves to distinguish between cases where the trajectory of Newton-CG remains in regions that are locally strongly-convex and those where it enters areas with high degree of weak-convexity or non-convexity. For example, in Section 4.2, more often than not, Newton-CG fails at the very outset with $\mathbf{x}_0$, whereas Newton-MR makes continuous progress.

Performance evaluation measure   In all of our experiments, we plot the objective value vs. the total number of oracle calls of function, gradient and Hessian-vector product. This is so since measuring "wall-clock" time can be greatly affected by individual implementation details. In contrast, counting the number of oracle calls, as an implementation and system independent unit of complexity, is most appropriate and fair. More specifically, after computing each function value, computing the corresponding gradient is equivalent to one additional function evaluation. Our implementations are Hessian-free, i.e., we merely require Hessian-vector products instead of using the explicit Hessian. For this, each Hessian-vector product amounts to two additional function evaluations, as compared with gradient evaluation. The number of such oracle calls per iteration for all the algorithms is given in Table 2.

Table 3
Data sets used for example of Section 4.1. All data sets are publicly available from [27,38].

Name	n	p	C	$d = (C-1) \times p$	Convexity of (40)
20 Newsgroups	10,142	53,975	20	1,025,525	Weak
cifar10	50,000	3,072	10	27,648	Strict
covetype,	435,759	54	7	324	Strict
gisette	6,000	5,000	2	5,000	Strict
mnist	60,000	784	10	7,065	Strict
UJIIndoorLoc	19,937	520	5	2,080	Strict

4.1. Softmax regression

Here, we consider the softmax cross-entropy minimization problem without regularization, which is used in machine learning for multi-class classification applications. More specifically, we have

$$
f(\mathbf{x}) \triangleq \mathcal{L}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{C-1}) = \sum_{i=1}^{n} \left(\log \left(1 + \sum_{c'=1}^{C-1} e^{\langle \mathbf{a}_i, \mathbf{x}_{c'} \rangle} \right) - \sum_{c=1}^{C-1} \mathbf{1}(b_i = c) \langle \mathbf{a}_i, \mathbf{x}_c \rangle \right),
$$
(40)

where $\{\mathbf{a}_i, b_i\}_{i=1}^n$ with $\mathbf{a}_i \in \mathbb{R}^p$, $b_i \in \{0, 1, \ldots, C\}$ denote the training data, C is the total number of classes for each input data $\mathbf{a}_i$ and $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{C-1})$. Note that, in this case, we have $d = (C-1) \times p$. It can be shown that, depending on the data, (40) is either strictly convex or merely weakly convex. In either case, however, by a similar analysis as that in Example 4, one can show that $\nu(\mathbf{x}) = 1$. Hence, it follows that $\mathbf{g}(\mathbf{x}_k) \in \text{Range}(\mathbf{H}(\mathbf{x}_k))$ and CG iterations within Newton-CG are well-defined. The datasets use for our experiments in this section are listed in Table 3 and the performance of each method is depicted in Fig. 1.

4.2. Gaussian mixture model

Here, we consider an example involving a mixture of Gaussian densities where the goal is to recover some mixture parameters such as the mean vectors and the mixture weights. Although, this problem is generally not invex, it has been shown in [67] that it indeed exhibits features that are close to being invex, e.g., small regions of saddle points and large regions containing global minimum. Nonetheless, the non-convexity of this problem results in a Hessian matrix that can become indefinite and/or singular across iterations.
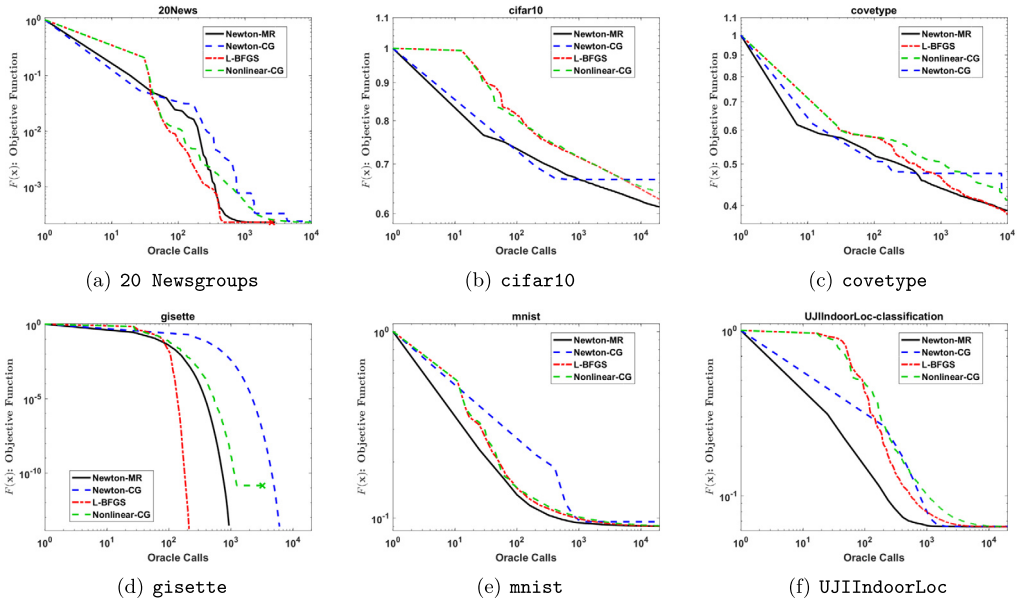
Fig. 1. Objective value, $f(\mathbf{x})$ vs. number of oracle calls as in Table 2 for datasets of Table 3 on the model problem (40). As for this convex problem, Gauss-Newton is mathematically equivalent to Newton-CG, it is naturally excluded from comparisons. All algorithms are always initialized at $\mathbf{x}_0 = \mathbf{0}$. As it can be clearly seen, Newton-MR almost always performs very competitively, and in particular, it greatly outperforms Newton-CG.

For simplicity, we consider a mixture model with two Gaussian components as

$$f(\mathbf{x}) \triangleq \mathcal{L}(x_0, \mathbf{x}_1, \mathbf{x}_2) = -\sum_{i=1}^{n} \log \Big(\omega(x_0)\Phi\left(\mathbf{a}_i; \mathbf{x}_1, \boldsymbol{\Sigma}_1\right) + (1 - \omega(x_0))\Phi\left(\mathbf{a}_i; \mathbf{x}_2, \boldsymbol{\Sigma}_2\right) \Big),$$
(41)

where Φ denotes the density of the p-dimensional standard normal distribution, $\mathbf{a}_i \in \mathbb{R}^p$ are the data points, $\mathbf{x}_1 \in \mathbb{R}^d, \mathbf{x}_2 \in \mathbb{R}^p, \boldsymbol{\Sigma}_1 \in \mathbb{R}^{p \times p}$, and $\boldsymbol{\Sigma}_2 \in \mathbb{R}^{p \times p}$ are the corresponding mean vectors and the covariance matrices of two the Gaussian distributions, $x_0 \in \mathbb{R}$, and $\omega(t) = 1/(1 + e^{-t})$ to ensure that the mixing weight lies within $[0, 1]$. Note that, here, $\mathbf{x} \triangleq [x_0, \mathbf{x}_1^\mathsf{T}, \mathbf{x}_2^\mathsf{T}]^\mathsf{T} \in \mathbb{R}^{2p+1}$ and $d = 2p + 1$.

We run the experiments 500 times, and plot the performance profile [39,51] of each method; see Fig. 2. For each run, we generate $1,000$ random data points, generated from the mixture distribution (41) with $p = 100$, and ground truth parameters as $x_0^\star \sim \mathcal{U}_1[0, 1], \mathbf{x}_1^\star \sim \mathcal{U}_p[-1, 1]$, and $\mathbf{x}_2^\star \sim \mathcal{U}_p[3, 4]$, where $\mathcal{U}_p[a, b]$ denotes the distribution of a p-dimensional random vector whose independent components are drawn uniformly over the interval $[a, b]$. Covariance matrices are constructed randomly, at each run, with controlled condition number, such that they are not axis-aligned. For this, we first randomly generate two $p \times p$ matrices, $\mathbf{W}_1, \mathbf{W}_2$, whose elements are iid drawn standard normal distribution. We then find corresponding orthogonal basis, $\mathbf{Q}_1, \mathbf{Q}_2$, using QR factorizations. This is then followed by forming $\boldsymbol{\Sigma}_i^{-1} = \mathbf{Q}_i^\mathsf{T} \mathbf{D} \mathbf{Q}_i$ where $\mathbf{D}$ is a diagonal
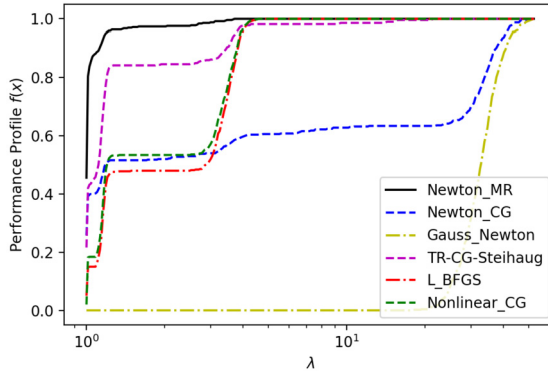
Fig. 2. Performance profile for 500 runs of various methods for solving (41) as detailed in Section 4.2. For a given λ in the x-axis, the corresponding value on the y-axis is the proportion of times that a given solver's performance lies within a factor λ of the best possible performance over all runs. All algorithms are always initialized at random using a multivariate normal distribution with mean zero and identity covariance. As it can be seen, Newton-MR, greatly outperforms alternative second-order methods in this example.

matrix whose diagonal entries are chosen equidistantly from the interval $[0, 100]$, i.e., the condition number of each $\mathbf{\Sigma}_i$ is 100.

The performance profiles of all the methods are gathered in Fig. 2. As it can be seen, Newton-MR greatly outperforms all alternative second-order methods in this example. The trust-region method also shows a reasonable performance. The L-BFGS and Nonlinear-CG methods both have comparable performances. It is worthwhile to highlight that Newton-CG fails to converge in many of the runs. This is because, for this non-convex problem, the Hessian matrix can become indefinite and, as a result, the CG iterations can fail before the underlying inexactness condition is met. Furthermore, the Gauss-Newton method performed extremely poorly, which can be an indication that the underlying Gauss-Newton matrix is perhaps a low quality approximation to the true Hessian in such a model.

4.3. Nonlinear least-squares

The examples of Sections 4.1 and 4.2 show that, when the problem is (approximately) invex, the inexact Newton-MR method, depicted in Algorithm 2, can be a highly effective optimization algorithm. We now study a potential drawback of using the plain Newton-MR method for the optimization of the objectives that are highly non-invex. We set out to show that for such functions, Algorithms 1 and 2, by virtue of the monotonic reduction of the gradient norm, run the risk of converging to undesirable saddle points or local maxima. For this, following [93], we consider the simple, yet illustrative, non-linear least squares problems arising from the task of binary classification with squared

loss.[2] More specifically, given the data $\{\mathbf{a}_i, b_i\}_{i=1}^n$ with $\mathbf{a}_i \in \mathbb{R}^d, b_i \in \{0, 1\}$, we consider the regularized objective

$$f(\mathbf{x}) = \sum_{i=1}^n \left(b_i - \phi(\langle \mathbf{a}_i, \mathbf{x}\rangle)\right)^2 + \psi(\mathbf{x}), \tag{42}$$

where $\phi(z) = 1/(1 + e^{-z})$ is the sigmoid function and $\psi(\mathbf{x})$ is some regularization term. To make the objective highly non-invex, we add a non-convex regularization in the form of $\psi(\mathbf{x}) = \sum_{i=1}^d x_i^2/(1 + x_i^2)$, where x_i is the ith component of the vector $\mathbf{x}$.

We compare the performance of the inexact variant of Newton-MR, Algorithm 2, the Gauss-Newton method, as well as the trust-region algorithm with CG-Steihaug solver. The Gauss-Newton method has traditionally been a method of choice for non-linear least-squares problems (of course without a non-convex regularization). The trust-region algorithm can take advantage of the negative curvature directions that arise as part of the Steihaug-CG iterations, and it is an effective method for most-non-convex problems. The approximate Hessian for the Gauss-Newton method is taken as $\mathbf{J_r^\mathsf{T}}(\mathbf{x}) \cdot \mathbf{J_r}(\mathbf{x}) + \nabla^2\psi(\mathbf{x})$, where $\mathbf{J_r}(\mathbf{x}) \in \mathbb{R}^{n \times d}$ is the Jacobian of the sigmoid mapping within the squared loss. Since this matrix is not necessarily positive semi-definite (due to non-convexity of $\psi(\mathbf{x})$), the CG inner iterations might fail for this problem in which case we terminate the Gauss-Newton method. We initialize the algorithms using $\mathbf{x}_0 = \mathbf{0}$, $\mathbf{x}_0 = \mathbf{1}$, and one drawn from the multivariate normal distribution with mean zero and identity covariance. We evaluate the performance of the methods on three datasets from Table 3, namely 20 Newsgroups, mnist and UJIIndoorLoc. For our binary classification setting, we have relabeled the odd and even classes, respectively, as 0 and 1. The results are depicted in Fig. 3.

It is clear that, when the optimization landscape is somewhat less rugged near the initialization, e.g., the origin on most of the datasets, Newton-MR (as well as Gauss-Newton) can perform reasonably well. However, near more "hostile" regions with a high degree of non-convexity, Newton-MR fails to find a good solution and can converge to saddle points or local maxima. In such highly non-convex regions, the matrix used within the Gauss-Newton iterations can also become indefinite, leading to the underlying CG iterations to fail in which case the Gauss-Newton method is terminated (indicated by "×"). In contrast, by leveraging the negative curvature directions, the trust-region method can reliably converge to a local minima.

Experiments such as those depicted in Fig. 3 highlight the need for further refinement of the vanilla Newton-MR method studied in this paper to design more general purpose variants, that just like trust-region, can leverage the negative curvature directions that could arise as part of the inner solver. This way, one can extend Algorithms 1 and 2 far beyond invex problems for arbitrary non-convex objectives. We leave this important venture to future work.

---

[2] Logistic loss, the "standard" loss used in this task, leads to a convex objective. We use squared loss to obtain a nonconvex objective.
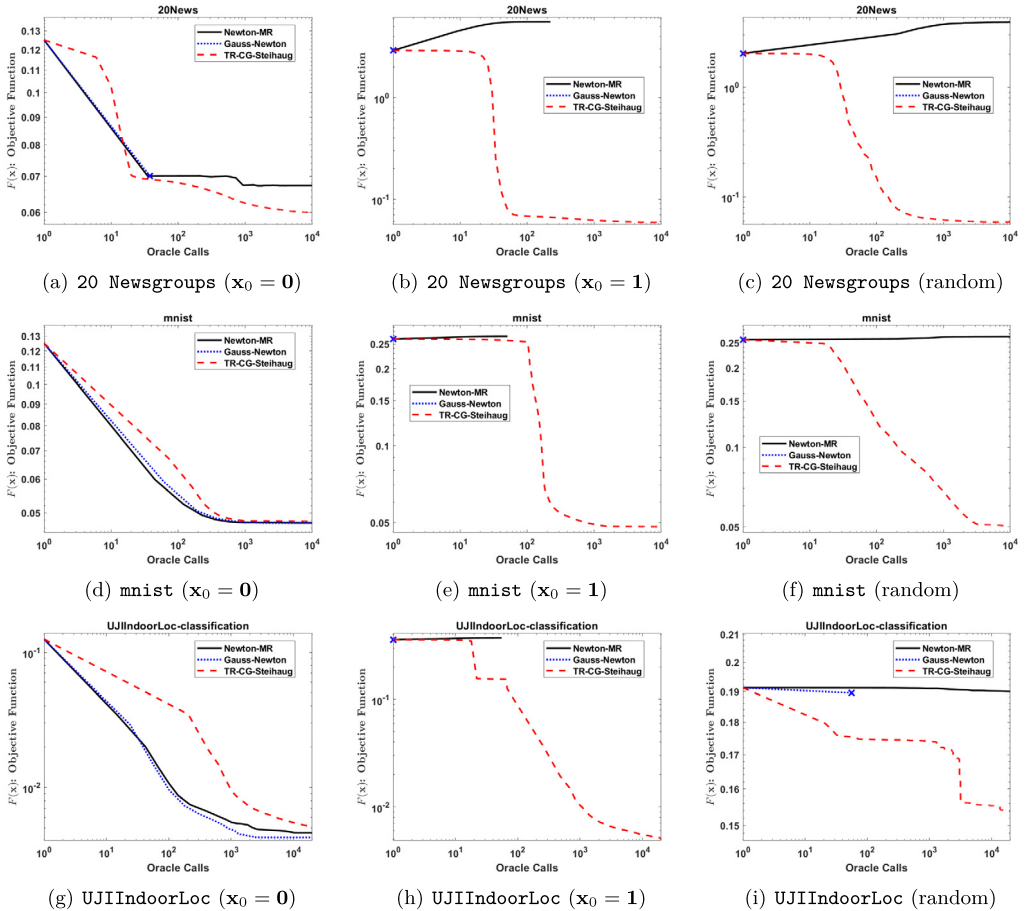
Fig. 3. Comparison of the Newton-MR, Gauss-Newton, and trust-region algorithms for a highly non-invex optimization problem, namely the problem (42) on three datasets from Table 3. We consider three initialization strategies: $\mathbf{x}_0 = \mathbf{0}$, $\mathbf{x}_0 = \mathbf{1}$, and one that is randomly drawn from the multivariate normal distribution with mean zero and identity covariance. Since this problem is highly non-invex, with unfortunate initialization, Newton-MR can converge to saddle points or even local maxima, whereas the trust-region algorithm, by leveraging the negative curvature direction arising as part of the CG-Steihaug iterations, can navigate its way out of these undesirable regions. The mark "$\times$" indicates the Gauss-Newton iteration where the CG method fails due to the indefiniteness of the underlying approximate Hessian matrix.

5. Conclusions

Motivated to extend the simplicity of the iterations of Newton-CG to beyond strongly convex problems, we consider an alternative algorithm, called Newton-MR, which can be readily used for unconstrained optimization of invex objectives. The iterations of Newton-MR merely involve ordinary least-squares problems, which are (approximately) solved by minimum residual iterative methods, followed by Armijo-type line-search to determine the step-size. We show that under mild assumptions and weak inexactness

conditions for sub-problems, we can obtain a variety of local/global convergence results for Newton-MR.

Declaration of competing interest

Acknowledgements

Appendix A. More on Assumption 4

This section contains some more examples and insights relating to Assumption 4.

A.1. Assumption 4 with $\nu < 1$

Example 4 gives a function that satisfies Assumption 4 with $\nu = 1$. We now provide a non-trivial example of a function that satisfies Assumption 4 with $\nu < 1$. Consider the fractional programming problem of the form

$$\min_{(x_1, x_2) \in \mathcal{X}} f(x_1, x_2) = \frac{a x_1^2}{b - x_2},$$

where $\mathcal{X} = \big\{ (x_1, x_2) \mid x_1 \in \mathbb{R}, \ x_2 \in (-\infty, b) \cup (b, \infty) \big\}$. The Hessian matrix and the gradient of f can, respectively, be written as

$$\mathbf{g}(\mathbf{x}) = \begin{pmatrix} \frac{-2a x_1}{x_2 - b} \\ \frac{a x_1^2}{(x_2 - b)^2} \end{pmatrix}, \quad \mathbf{H}(\mathbf{x}) = \begin{pmatrix} \frac{-2a}{x_2 - b} & \frac{2a x_1}{(x_2 - b)^2} \\ \frac{2a x_1}{(x_2 - b)^2} & \frac{-2a x_1^2}{(x_2 - b)^3} \end{pmatrix}.$$

Hence, for a given $\mathbf{x}$, the range and the null-space of $\mathbf{H}(\mathbf{x})$ are, respectively, given by

$$\mathbf{u_x} = \begin{pmatrix} -1 \\ \frac{x_1}{x_2 - b} \end{pmatrix}, \quad \mathbf{u_x^{\perp}} = \begin{pmatrix} \frac{x_1}{x_2 - b} \\ 1 \end{pmatrix}.$$

Now, we get

$$|\langle \mathbf{u_x}, \mathbf{g}(\mathbf{x}) \rangle| = \left| \frac{2ax_1}{x_2 - b} + \frac{ax_1^3}{(x_2 - b)^3} \right|, \quad |\langle \mathbf{u_x^\perp}, \mathbf{g}(\mathbf{x}) \rangle| = \frac{ax_1^2}{(x_2 - b)^2},$$

and

$$\frac{|\langle \mathbf{u_x}, \mathbf{g}(\mathbf{x}) \rangle|}{|\langle \mathbf{u_x^\perp}, \mathbf{g}(\mathbf{x}) \rangle|} = \left| \frac{2(x_2 - b)}{x_1} + \frac{x_1}{(x_2 - b)} \right| \geq 2\sqrt{2}, \quad \forall \mathbf{x} \in \mathcal{X},$$

where the lower bound is obtained by considering the minimum of the function $h(y) = 2/y + y$ for $y = x_1/(x_2 - b)$. Hence, it follows that $\nu = 8/9$. Clearly, f is unbounded below and, admittedly, this example is of little interest in optimization. Nonetheless, it serves as a non-trivial example of functions that satisfy Assumption 4.

A.2. Composition of functions and Assumption 4

We now consider Assumption 4 in the context of the composition of functions. More specifically, consider

$$f(\mathbf{x}) = h(\mathbf{r}(\mathbf{x})), \tag{43}$$

where $\mathbf{r} : \mathbb{R}^d \to \mathbb{R}^p$ and $h : \mathbb{R}^p \to \mathbb{R}$ are smooth functions. We have

$$\mathbf{g}(\mathbf{x}) \triangleq \nabla f(\mathbf{x}) = \mathbf{J_r^\mathsf{T}}(\mathbf{x}) \cdot \nabla h(\mathbf{r}(\mathbf{x})),$$

$$\mathbf{H}(\mathbf{x}) \triangleq \nabla^2 f(\mathbf{x}) = \mathbf{J_r^\mathsf{T}}(\mathbf{x}) \cdot \nabla^2 h(\mathbf{r}(\mathbf{x})) \cdot \mathbf{J_r}(\mathbf{x}) + \partial \mathbf{J_r}(\mathbf{x}) \cdot \nabla h(\mathbf{r}(\mathbf{x})),$$

where $\mathbf{J_r}(\mathbf{x}) \in \mathbb{R}^{p \times d}$ and $\partial \mathbf{J_r}(\mathbf{x}) \in \mathbb{R}^{d \times d \times p}$ are, respectively, the Jacobian and the tensor of all second-order partial derivatives of $\mathbf{r}$, and $\nabla h(\mathbf{r}(\mathbf{x})) \in \mathbb{R}^p$, and $\nabla^2 h(\mathbf{r}(\mathbf{x})) \in \mathbb{R}^{p \times p}$ are the gradient and Hessian of h, respectively. Lemma 8 gives a sufficient condition for f to satisfy Assumption 4.

Lemma 8. *Suppose Assumption 3 holds, $\nabla^2 h(\mathbf{r}(\mathbf{x}))$ is full rank, and also $\mathrm{rank}(\mathbf{H}(\mathbf{x})) \geq \mathrm{rank}(\mathbf{J_r}(\mathbf{x}))$. If for any $\mathbf{x}_0 \in \mathbb{R}^d$, there is some $\nu(\mathbf{x}_0) \in (0, 1]$, such that*

$$\|\partial \mathbf{J_r}(\mathbf{x}) \cdot \nabla h(\mathbf{r}(\mathbf{x}))\| \leq \frac{\gamma(\mathbf{x}_0)\sqrt{1 - \nu(\mathbf{x}_0)}}{2}, \quad \forall \mathbf{x} \in \mathcal{X}_0, \tag{44}$$

then Assumption 4 holds with $\nu(\mathbf{x}_0)$. Here, $\gamma(\mathbf{x}_0)$ is as in Assumption 3.

Proof. Let $\mathbf{x}_0 \in \mathbb{R}^d$ be given and take any $\mathbf{x} \in \mathcal{X}_0$. Define $\mathbf{B} \triangleq \mathbf{A} + \mathbf{E}$ where

$$\mathbf{B} = \mathbf{H}(\mathbf{x}), \quad \mathbf{A} = \mathbf{J_r^\mathsf{T}}(\mathbf{x}) \cdot \nabla^2 h(\mathbf{r}(\mathbf{x})) \cdot \mathbf{J_r}(\mathbf{x}), \quad \text{and} \quad \mathbf{E} = \partial \mathbf{J_r}(\mathbf{x}) \cdot \nabla h(\mathbf{r}(\mathbf{x})).$$

Let $\mathbf{U_A}$ and $\mathbf{U_B}$ be, respectively, arbitrary orthogonal bases for $\mathrm{Range}(\mathbf{A})$ and $\mathrm{Range}(\mathbf{B})$ and denote $\mathbf{U_A^\perp}$ and $\mathbf{U_B^\perp}$ as their respective orthogonal complement.

By assumption, we have $r_{\mathbf{A}} = \text{rank}(\mathbf{A}) \leq \text{rank}(\mathbf{B}) = r_{\mathbf{B}}$. Let $\widehat{\mathbf{U}}_{\mathbf{A}} \in \mathbb{R}^{d \times r_{\mathbf{B}}}$ be $\mathbf{U}_{\mathbf{A}}$ augmented by any $r_{\mathbf{B}} - r_{\mathbf{A}}$ vectors from $\mathbf{U}_{\mathbf{A}}^{\perp}$ such that $\text{rank}(\widehat{\mathbf{U}}_{\mathbf{A}}) = \text{rank}(\mathbf{U}_{\mathbf{B}})$. By the matrix perturbation theory applied to $\mathbf{B} = \mathbf{A} + \mathbf{E}$, [77, Theorem 19], we have

$$\left\| \widehat{\mathbf{U}}_{\mathbf{A}} \widehat{\mathbf{U}}_{\mathbf{A}}^{\intercal} - \mathbf{U}_{\mathbf{B}} \mathbf{U}_{\mathbf{B}}^{\intercal} \right\| \leq \frac{2 \left\| \mathbf{E} \right\|}{\sigma_{r_{\mathbf{B}}}(\mathbf{B})} \leq \frac{2 \left\| \mathbf{E} \right\|}{\gamma(\mathbf{x}_0)} \leq \sqrt{1 - \nu(\mathbf{x}_0)},$$

where $\sigma_{r_{\mathbf{B}}}(\mathbf{B})$ is the smallest non-zero singular value of $\mathbf{B}$. This, in turn, implies (see [50, Theorem 2.5.1])

$$\left\| [\mathbf{U}_{\mathbf{B}}^{\perp}]^{\intercal} \widehat{\mathbf{U}}_{\mathbf{A}} \right\|^2 = \left\| \widehat{\mathbf{U}}_{\mathbf{A}} \widehat{\mathbf{U}}_{\mathbf{A}}^{\intercal} - \mathbf{U}_{\mathbf{B}} \mathbf{U}_{\mathbf{B}}^{\intercal} \right\|^2 \leq 1 - \nu(\mathbf{x}_0).$$

Now since $\nabla^2 h(\mathbf{r}(\mathbf{x}))$ is full rank, then $\text{Range}(\mathbf{A}) = \text{Range}(\mathbf{J}_{\mathbf{r}}^{\intercal}(\mathbf{x}))$. Therefore $\mathbf{g}(\mathbf{x}) = \mathbf{U}_{\mathbf{A}} \mathbf{U}_{\mathbf{A}}^{\intercal} \mathbf{g}(\mathbf{x})$ and hence $\mathbf{U}_{\mathbf{A}}^{\perp} [\mathbf{U}_{\mathbf{A}}^{\perp}]^{\intercal} \mathbf{g}(\mathbf{x}) = \mathbf{0}$, which implies $\mathbf{g}(\mathbf{x}) = \widehat{\mathbf{U}}_{\mathbf{A}} \widehat{\mathbf{U}}_{\mathbf{A}}^{\intercal} \mathbf{g}(\mathbf{x})$ and $\| \mathbf{g}(\mathbf{x}) \| = \left\| \widehat{\mathbf{U}}_{\mathbf{A}}^{\intercal} \mathbf{g}(\mathbf{x}) \right\|$. Now we have

$$\left\| [\mathbf{U}_{\mathbf{B}}^{\perp}]^{\intercal} \mathbf{g}(\mathbf{x}) \right\|^2 = \left\| [\mathbf{U}_{\mathbf{B}}^{\perp}]^{\intercal} \widehat{\mathbf{U}}_{\mathbf{A}} \widehat{\mathbf{U}}_{\mathbf{A}}^{\intercal} \mathbf{g}(\mathbf{x}) \right\|^2 \leq \left\| [\mathbf{U}_{\mathbf{B}}^{\perp}]^{\intercal} \widehat{\mathbf{U}}_{\mathbf{A}} \right\|^2 \left\| \widehat{\mathbf{U}}_{\mathbf{A}}^{\intercal} \mathbf{g}(\mathbf{x}) \right\|^2$$

$$\leq \left(1 - \nu(\mathbf{x}_0) \right) \| \mathbf{g}(\mathbf{x}) \|^2 = \left(1 - \nu(\mathbf{x}_0) \right) \left(\| \mathbf{U}_{\mathbf{B}}^{\intercal} \mathbf{g}(\mathbf{x}) \|^2 + \left\| [\mathbf{U}_{\mathbf{B}}^{\perp}]^{\intercal} \mathbf{g}(\mathbf{x}) \right\|^2 \right).$$

Rearranging the terms above, we obtain

$$\left\| [\mathbf{U}_{\mathbf{B}}^{\perp}]^{\intercal} \mathbf{g}(\mathbf{x}) \right\|^2 \leq \frac{1 - \nu(\mathbf{x}_0)}{\nu(\mathbf{x}_0)} \| \mathbf{U}_{\mathbf{B}}^{\intercal} \mathbf{g}(\mathbf{x}) \|^2,$$

which completes the proof. $\square$

The requirement for $\nabla^2 h(\mathbf{r}(\mathbf{x}))$ being full rank is satisfied in many applications, e.g., nonlinear least squares where $h(\mathbf{z}) = \| \mathbf{z} \|^2 / 2$, statistical parameter estimations using negative log-likelihood where $h(z) = -\log(z)$, or multi-class classifications using cross-entropy loss where $h(\mathbf{z}) = -\sum_{i=1}^{p} y_i \log(z_i)$ for some $\sum_{i=1}^{p} y_i = 1$ and $y_i \geq 0$. The requirement for $\text{rank}(\mathbf{H}(\mathbf{x})) \geq \text{rank}(\mathbf{J}_{\mathbf{r}}(\mathbf{x}))$, however, might not hold for many problems. Nonetheless, Lemma 8 can give a qualitative guide to understanding the connection between Newton-MR and Gauss-Newton, when both are applied to $f(\mathbf{x}) = h(\mathbf{r}(\mathbf{x}))$.

Connection to Gauss-Newton Suppose f in (43) has isolated local minima, e.g., if $\mathbf{J}_{\mathbf{r}}(\mathbf{x}^{\star})$ is full-rank at a local minimum $\mathbf{x}^{\star}$. Define $\mathbf{S}(\mathbf{x}) \triangleq \partial \mathbf{J}_{\mathbf{r}}(\mathbf{x}) \cdot \nabla h(\mathbf{r}(\mathbf{x}))$. It is well-known that the convergence of Gauss-Newton is greatly affected by the magnitude of $\| \mathbf{S}(\mathbf{x}^{\star}) \|$; see [87, Section 7.2]. Indeed, when $\| \mathbf{S}(\mathbf{x}^{\star}) \|$ is large, Gauss-Newton can perform poorly, whereas small values of $\| \mathbf{S}(\mathbf{x}^{\star}) \|$ imply Gauss-Newton matrix is a good approximation to the full Hessian. In fact, local quadratic convergence of Gauss-Newton can only be guaranteed when $\| \mathbf{S}(\mathbf{x}^{\star}) \| = 0$; otherwise, the convergence can degrade significantly. For

example, if $\|\mathbf{S}(\mathbf{x}^\star)\|$ is too large, the Gauss-Newton method may not even converge, while if $\|\mathbf{S}(\mathbf{x}^\star)\|$ is small enough relative to $\left\|\mathbf{J}_\mathbf{r}^\mathsf{T}(\mathbf{x}^\star) \cdot \nabla^2 h(\mathbf{r}(\mathbf{x}^\star)) \cdot \mathbf{J}_\mathbf{r}(\mathbf{x}^\star)\right\|$, the convergence is linear.

Lemma 8 allows us to, at least very qualitatively and superficially, compare the performance of Gauss-Newton with Newton-MR for the minimization of (43). Roughly speaking, Lemma 8 relates $\|\mathbf{S}(\mathbf{x}^\star)\|$ to ν, which directly affects the performance of Newton-MR, e.g., see Theorem 1 where larger values of ν imply faster convergence for Newton-MR. More specifically, suppose $\|\mathbf{S}(\mathbf{x}^\star)\|$ is small. Assume that $\mathbf{S}(\mathbf{x}^\star)$ is smooth enough such that $\|\mathbf{S}(\mathbf{x})\|$ is also small for any $\mathbf{x}$ in some compact neighborhood of $\mathbf{x}^\star$, denoted by $\mathcal{C}$. Let $\mathbf{x}_0$ be the point in this compact neighborhood such that $\|\mathbf{g}(\mathbf{x})\| \leq \|\mathbf{g}(\mathbf{x}_0)\|, \ \forall \mathbf{x} \in \mathcal{C}$. With $\mathcal{X}_0$ defined by this $\mathbf{x}_0$, consider (44). Since the left hand side is assumed small, $\nu(\mathbf{x}_0)$ can be taken large, which directly implies a better convergence rate for Newton-MR.

In this light, we can expect Newton-MR to perform well whenever Gauss-Newton exhibits good behaviors. Our empirical evaluations of Section 4.2 indeed support this, while hinting at a possibility that the converse might not necessarily be true. Nonetheless, to make a fair and more accurate theoretical comparison, one needs to analyze the Gauss-Newton method under the assumptions of this paper (for example by drawing upon the ideas in [42] related to the minimum-norm solution). This endeavor is left for future work.

References

[1] H.-B. An, Z.-Z. Bai, A globally convergent Newton-GMRES method for large sparse systems of nonlinear equations, Appl. Numer. Math. 57 (3) (2007) 235–252.

[2] L. Angelani, R. Di Leonardo, G. Ruocco, A. Scala, F. Sciortino, Saddles in the energy landscape probed by supercooled liquids, Phys. Rev. Lett. 85 (25) (2000) 5356.

[3] M. Anitescu, Degenerate nonlinear programming with a quadratic growth condition, SIAM J. Optim. 10 (4) (2000) 1116–1135.

[4] Y. Arjevani, O. Shamir, Oracle complexity of second-order methods for finite-sum problems, in: International Conference on Machine Learning, 2017, pp. 205–213.

[5] Y. Arjevani, O. Shamir, R. Shiff, Oracle complexity of second-order methods for smooth convex optimization, Math. Program. (2017) 1–34.

[6] R. Barboni, G. Peyré, F.-X. Vialard, Global convergence of ResNets: from finite to infinite width using linear parameterization, arXiv preprint, arXiv:2112.05531, 2021.

[7] R. Bassily, M. Belkin, S. Ma, On exponential convergence of SGD in non-convex over-parametrized learning, arXiv preprint, arXiv:1811.02564, 2018.

[8] R. Behling, A. Fischer, A unified local convergence analysis of inexact constrained Levenberg–Marquardt methods, Optim. Lett. 6 (5) (2012) 927–940.

[9] R. Behling, D.S. Gonçalves, S.A. Santos, Local convergence analysis of the Levenberg–Marquardt framework for nonzero-residue nonlinear least-squares problems under an error bound condition, J. Optim. Theory Appl. 183 (3) (2019) 1099–1122.

[10] S. Bellavia, C. Cartis, N.I.M. Gould, B. Morini, Ph.L. Toint, Convergence of a regularized Euclidean residual algorithm for nonlinear least-squares, SIAM J. Numer. Anal. 48 (1) (2010) 1–29.

[11] S. Bellavia, S. Gratton, E. Riccietti, A Levenberg–Marquardt method for large nonlinear least-squares problems with dynamic accuracy in functions and gradients, Numer. Math. 140 (3) (2018) 791–825.

[12] S. Bellavia, M. Macconi, B. Morini, A hybrid Newton-GMRES method for solving nonlinear equations, in: International Conference on Numerical Analysis and Its Applications, Springer, 2000, pp. 68–75.

[13] S. Bellavia, B. Morini, A globally convergent Newton-GMRES subspace method for systems of nonlinear equations, SIAM J. Sci. Comput. 23 (3) (2001) 940–960.

[14] A. Ben-Israel, B. Mond, What is invexity?, ANZIAM J. 28 (1) (1986) 1–9.

[15] A.S. Berahas, R. Bollapragada, J. Nocedal, An investigation of Newton-sketch and subsampled Newton methods, Optim. Methods Softw. 35 (4) (2020) 661–680.

[16] T. Bianconcini, G. Liuzzi, B. Morini, M. Sciandrone, On the use of iterative methods in cubic regularization for unconstrained optimization, Comput. Optim. Appl. 60 (1) (2015) 35–57.

[17] R. Bollapragada, R.H. Byrd, J. Nocedal, Exact and inexact subsampled Newton methods for optimization, IMA J. Numer. Anal. 39 (2) (2019) 545–578.

[18] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.

[19] P.N. Brown, Y. Saad, Hybrid Krylov methods for nonlinear systems of equations, SIAM J. Sci. Stat. Comput. 11 (3) (1990) 450–481.

[20] P.N. Brown, Y. Saad, Convergence theory of nonlinear Newton–Krylov algorithms, SIAM J. Optim. 4 (2) (1994) 297–330.

[21] R.H. Byrd, G.M. Chin, W. Neveitt, J. Nocedal, On the use of stochastic Hessian information in optimization methods for machine learning, SIAM J. Optim. 21 (3) (2011) 977–995.

[22] D. Calvetti, B. Lewis, L. Reichel, L-Curve for the MINRES Method, Advanced Signal Processing Algorithms, Architectures, and Implementations X, vol. 4116, International Society for Optics and Photonics, 2000, pp. 385–396.

[23] A. Cambini, L. Martein, Generalized Convexity and Optimization: Theory and Applications, Lecture Notes in Economics and Mathematical Systems, Springer Berlin Heidelberg, 2008.

[24] Y. Carmon, J.C. Duchi, Gradient descent efficiently finds the cubic-regularized non-convex Newton step, arXiv preprint, arXiv:1612.00547, 2016.

[25] C. Cartis, N.I.M. Gould, Ph.L. Toint, Adaptive cubic regularisation methods for unconstrained optimization. Part I: motivation, convergence and numerical results, Math. Program. 127 (2) (2011) 245–295.

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

[27] C.-C. Chang, C.-J. Lin LIBSVM, A library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) 27, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[28] X. Chen, T. Yamamoto, Newton-like methods for solving underdetermined nonlinear equations with nondifferentiable terms, J. Comput. Appl. Math. 55 (3) (1994) 311–324.

[29] S.-C.T. Choi, C.C. Paige, M.A. Saunders, MINRES-QLP: a Krylov subspace method for indefinite or singular symmetric systems, SIAM J. Sci. Comput. 33 (4) (2011) 1810–1836.

[30] S.-C.T. Choi, M.A. Saunders, Algorithm 937: MINRES-QLP for symmetric and Hermitian linear equations and least-squares problems, ACM Trans. Math. Softw. 40 (2) (2014) 16.

[31] P. Ciarlet, Linear and Nonlinear Functional Analysis with Applications, Society for Industrial and Applied Mathematics, 2013.

[32] A.R. Conn, N.I.M. Gould, Ph.L. Toint, Trust Region Methods, Vol. 1, SIAM, 2000.

[33] R. Crane, F. Roosta, Invexifying regularization of non-linear least-squares problems, arXiv preprint, arXiv:2111.11027, 2021.

[34] F.E. Curtis, D.P. Robinson, C.W. Royer, S.J. Wright, Trust-region Newton-CG with strong second-order complexity guarantees for nonconvex optimization, SIAM J. Optim. 31 (1) (2021) 518–544.

[35] R.S. Dembo, S.C. Eisenstat, T. Steihaug, Inexact Newton methods, SIAM J. Numer. Anal. 19 (2) (1982) 400–408.

[36] J.E. Dennis Jr, R.B. Schnabel, Numerical Methods for Unconstrained Optimization and Nonlinear Equations, SIAM, 1996.

[37] S. Dereich, S. Kassing, Convergence of stochastic gradient descent schemes for Lojasiewicz-landscapes, arXiv preprint, arXiv:2102.09385, 2021.

[38] D. Dheeru, E. Karra Taniskidou, UCI Machine Learning Repository, 2017.

[39] E.D. Dolan, J.J. Moré, Benchmarking optimization software with performance profiles, Math. Program. 91 (2) (2002) 201–213.

[40] S.C. Eisenstat, H.F. Walker, Globally convergent inexact Newton methods, SIAM J. Optim. 4 (2) (1994) 393–422.

[41] S.C. Eisenstat, H.F. Walker, Choosing the forcing terms in an inexact Newton method, SIAM J. Sci. Comput. 17 (1) (1996) 16–32.

[42] J. Eriksson, P. Wedin, M. Gulliksson, I. Söderkvist, Regularization methods for uniformly rank-deficient nonlinear least-squares problems, J. Optim. Theory Appl. 127 (1) (2005) 1–26.

[43] R. Estrin, D. Orban, M. Saunders, Euclidean-norm error bounds for SYMMLQ and CG, SIAM J. Matrix Anal. Appl. 40 (1) (2019) 235–253.

[44] J. Fan, The modified Levenberg-Marquardt method for nonlinear equations with cubic convergence, Math. Comput. 81 (277) (2012) 447–466.

[45] J.-y. Fan, Y.-x. Yuan, On the quadratic convergence of the Levenberg-Marquardt method without nonsingularity assumption, Computing 74 (1) (2005) 23–39.

[46] C.G. Frye, Critical Point Finding with Newton-MR by Analogy to Computing Square Roots. ArXiv preprint, arXiv:1906.05273, 2019.

[47] C.G. Frye, J. Simon, N.S. Wadia, A. Ligeralde, M.R. DeWeese, K.E. Bouchard, Critical point-finding methods reveal gradient-flat regions of deep network losses, Neural Comput. 33 (6) (2021) 1469–1497.

[48] C.G. Frye, N.S. Wadia, M.R. DeWeese, K.E. Bouchard, Numerically recovering the critical points of a deep linear autoencoder, arXiv preprint, arXiv:1901.10603, 2019.

[49] G. Golub, C. Van Loan, Matrix Computations, Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, 2013.

[50] G.H. Golub, C.F. Van Loan, Matrix Computations, vol. 3, JHU Press, 2012.

[51] N. Gould, J. Scott, A note on performance profiles for benchmarking software, ACM Trans. Math. Softw. 43 (2) (2016) 15.

[52] N.I.M. Gould, S. Lucidi, M. Roma, Ph.L. Toint, Solving the trust-region subproblem using the Lanczos method, SIAM J. Optim. 9 (2) (1999) 504–525.

[53] S. Gratton, A.S. Lawless, N.K. Nichols, Approximate Gauss–Newton methods for nonlinear least squares problems, SIAM J. Optim. 18 (1) (2007) 106–132.

[54] M. Hanke, Conjugate Gradient Type Methods for Ill-Posed Problems, Routledge, 2017.

[55] M.A. Hanson, On sufficiency of the Kuhn-Tucker conditions, J. Math. Anal. Appl. 80 (2) (1981) 545–550.

[56] J.H. Hubbard, B.B. Hubbard, Vector Calculus, Linear Algebra, and Differential Forms, Matrix Editions 5th edition, 2015.

[57] H. Karimi, J. Nutini, M. Schmidt, Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2016, pp. 795–811.

[58] C.T. Kelley, Iterative Methods for Linear and Nonlinear Equations, SIAM, 1995.

[59] S. Kim, L. Madden, E. Dall'Anese, Convergence of the inexact online gradient and proximal-gradient under the Polyak-Łojasiewicz condition, arXiv preprint, arXiv:2108.03285, 2021.

[60] S. Kylasa, F. Roosta, M.W. Mahoney, A. Grama GPU, Accelerated sub-sampled Newton's method for convex classification problems, in: Proceedings of the 2019 SIAM International Conference on Data Mining (SDM-19), SIAM, 2019, pp. 702–710.

[61] F. Lenders, C. Kirches, A. Potschka, trlib: a vector-free implementation of the GLTR method for iterative solution of the trust region problem, Optim. Methods Softw. 33 (3) (2018) 420–449.

[62] D.-H. Li, M. Fukushima, L. Qi, N. Yamashita, Regularized Newton methods for convex minimization problems with singular solutions, Comput. Optim. Appl. 28 (2) (2004) 131–147.

[63] Y.-J. Li, D.-H. Li, Truncated regularized Newton method for convex minimizations, Comput. Optim. Appl. 43 (1) (2009) 119–131.

[64] C. Liu, L. Zhu, M. Belkin, Loss landscapes and optimization in over-parameterized non-linear systems and neural networks, Appl. Comput. Harmon. Anal. (2022), https://doi.org/10.1016/j.acha.2021.12.009.

[65] Z.-Q. Luo, P. Tseng, Error bounds and convergence analysis of feasible descent methods: a general approach, Ann. Oper. Res. 46 (1) (1993) 157–178.

[66] J.W. McIver Jr, A. Komornicki, Structure of transition states in organic reactions. General theory and an application to the cyclobutene-butadiene isomerization using a semiempirical molecular orbital method, J. Am. Chem. Soc. 94 (8) (1972) 2625–2633.

[67] S. Mei, Y. Bai, A. Montanari, The landscape of empirical risk for non-convex losses, arXiv preprint, arXiv:1607.06534, 2016.

[68] K. Mishchenko, Regularized Newton method with global $O(1/k^2)$ convergence, arXiv preprint, arXiv:2112.02089, 2021.

[69] S.K. Mishra, G. Giorgi, Invexity and Optimization, Vol. 88, Springer Science & Business Media, 2008.

[70] B. Morini, Convergence behaviour of inexact Newton methods, Math. Comput. 68 (228) (1999) 1605–1613.

[71] V. Muthukumar, A. Narang, V. Subramanian, M. Belkin, D. Hsu, A. Sahai, Classification vs regression in overparameterized regimes: does the loss function matter?, J. Mach. Learn. Res. 22 (222) (2021) 1–69.

[72] I. Necoara, Y. Nesterov, F. Glineur, Linear convergence of first order methods for non-strongly convex optimization, Math. Program. (2018) 1–39.

[73] Y. Nesterov, Introductory Lectures on Convex Optimization, Vol. 87, Springer Science & Business Media, 2004.

[74] Y. Nesterov, Modified Gauss–Newton scheme with worst case guarantees for global performance, Optim. Methods Softw. 22 (3) (2007) 469–483.

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

[76] J. Nocedal, S. Wright, Numerical Optimization, Springer Science & Business Media, 2006.

[77] S. O'Rourke, V. Vu, K. Wang, Random perturbation of low rank matrices: improving classical bounds, Linear Algebra Appl. 540 (2018) 26–59.

[78] S. Oymak, M. Soltanolkotabi, Overparameterized nonlinear learning: gradient descent takes the shortest path?, in: International Conference on Machine Learning, PMLR, 2019, pp. 4951–4960.

[79] C.C. Paige, M.A. Saunders, Solution of sparse indefinite systems of linear equations, SIAM J. Numer. Anal. 12 (4) (1975) 617–629.

[80] B. Polyak, A. Tremba, Solving underdetermined nonlinear equations by Newton-like method, arXiv preprint, arXiv:1703.07810, 2017.

[81] F. Roosta-Khorasani, M.W. Mahoney, Sub-sampled Newton methods, Math. Program. 174 (1) (2019) 293–326.

[82] C.W. Royer, M. O'Neill, S.J. Wright, A Newton-CG algorithm with complexity guarantees for smooth unconstrained optimization, Math. Program. (2019) 1–38.

[83] F. Schöpfer, Linear convergence of descent methods for the unconstrained minimization of restricted strongly convex functions, SIAM J. Optim. 26 (3) (2016) 1883–1911.

[84] S. Shalev-Shwartz, S. Ben-David, Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press, 2014.

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

[86] L. Su, P. Yang, On learning over-parameterized neural networks: a functional approximation perspective, Adv. Neural Inf. Process. Syst. 32 (2019).

[87] W. Sun, Y.-X. Yuan, Optimization Theory and Methods: Nonlinear Programming, Vol. 1, Springer Science & Business Media, 2006.

[88] H. Taheri, R. Pedarsani, C. Thrampoulidis, Fundamental limits of ridge-regularized empirical risk minimization in high dimensions, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 2773–2781.

[89] P. Tseng, S. Yun, Block-coordinate gradient descent method for linearly constrained nonsmooth separable optimization, J. Optim. Theory Appl. 140 (3) (2009) 513.

[90] S. Vaswani, F. Bach, M. Schmidt, Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron, in: The 22nd International Conference on Artificial Intelligence and Statistics, PMLR, 2019, pp. 1195–1204.

[91] H.F. Walker, L.T. Watson, Least-change secant update methods for underdetermined systems, SIAM J. Numer. Anal. 27 (5) (1990) 1227–1262.

[92] P. Xu, F. Roosta, M.W. Mahoney, Newton-type methods for non-convex optimization under inexact Hessian information, Math. Program. 184 (1) (2020) 35–70.

[93] P. Xu, F. Roosta, M.W. Mahoney, Second-order optimization for non-convex machine learning: an empirical study, in: Proceedings of the 2020 SIAM International Conference on Data Mining, SIAM, 2020, pp. 199–207.

[94] P. Xu, J. Yang, F. Roosta-Khorasani, C. Ré, M.W. Mahoney, Sub-sampled Newton methods with non-uniform sampling, in: Advances in Neural Information Processing Systems, 2016, pp. 3000–3008.

[95] N. Yamashita, M. Fukushima, On the rate of convergence of the Levenberg-Marquardt method, in: Topics in Numerical Analysis, Springer, 2001, pp. 239–249.

[96] Z. Yao, P. Xu, F. Roosta, S.J. Wright, M.W. Mahoney, Inexact Newton-CG algorithms with complexity guarantees, arXiv preprint, arXiv:2109.14016, 2021.

[97] Y.-X. Yuan, Recent advances in numerical methods for nonlinear equations and nonlinear least squares, Numer. Algebra Control Optim. 1 (1) (2011) 15–34.

[98] H. Zhang, The restricted strong convexity revisited: analysis of equivalence to error bound and quadratic growth, Optim. Lett. 11 (4) (2017) 817–833.

[99] H. Zhang, L. Cheng, Restricted strong convexity and its applications to convergence analysis of gradient-type methods in convex optimization, Optim. Lett. 9 (5) (2015) 961–979.

[100] R. Zhao, J. Fan, Global complexity bound of the Levenberg–Marquardt method, Optim. Methods Softw. 31 (4) (2016) 805–814.

[101] G. Zhou, L. Qi, On the convergence of an inexact Newton-type method, Oper. Res. Lett. 34 (6) (2006) 647–652.

[102] G. Zhou, K.-C. Toh, Superlinear convergence of a Newton-type algorithm for monotone equations, J. Optim. Theory Appl. 125 (1) (2005) 205–221.