Mathematical Programming (2023) 198:899–936
https://doi.org/10.1007/s10107-022-01797-5

FULL LENGTH PAPER

Series A

A globally convergent proximal Newton-type method in nonsmooth convex optimization

Boris S. Mordukhovich[1] · Xiaoming Yuan[2] · Shangzhi Zeng[3] · Jin Zhang[4]

Received: 16 November 2020 / Accepted: 23 February 2022 / Published online: 22 March 2022

Abstract

The paper proposes and justifies a new algorithm of the proximal Newton type to solve a broad class of nonsmooth composite convex optimization problems without strong convexity assumptions. Based on advanced notions and techniques of variational analysis, we establish implementable results on the global convergence of the proposed algorithm as well as its local convergence with superlinear and quadratic rates. For certain structured problems, the obtained local convergence conditions do not require the local Lipschitz continuity of the corresponding Hessian mappings that is a crucial assumption used in the literature to ensure a superlinear convergence of other algorithms of the proximal Newton type. The conducted numerical experiments of solving the l_1 regularized logistic regression model illustrate the possibility of applying the proposed algorithm to deal with practically important problems.

Keywords Nonsmooth convex optimization · Machine learning · Proximal Newton methods · Global and local convergence · Metric subregularity

Mathematics Subject Classification 90C25 · 49M15 · 49J53

Boris S. Mordukhovich: Partly supported by the USA National Science Foundation under grants DMS-1512846 and DMS-1808978, by the USA Air Force Office of Scientific Research grant #15RT04, and by Australian Research Council Under Grant DP-190100555.
Xiaoming Yuan: Supported by the General Research Fund 12302318 from the Hong Kong Research Grants Council.
Shangzhi Zeng: Supported by the Pacific Institute for the Mathematical Sciences (PIMS)
Jin Zhang: Supported by National Science Foundation of China 11971220, by Shenzhen Science and Technology Program (No. RCYX20200714114700072), and by the Stable Support Plan Program of Shenzhen Natural Science Fund (No. 20200925152128002).

✉ Boris S. Mordukhovich
boris@math.wayne.edu

✉ Jin Zhang
zhangj9@sustech.edu.cn

Extended author information available on the last page of the article

1 Introduction

In this paper we consider a class of optimization problems of the following type:

$$\min_{x\in\mathbb{R}^n} F(x) := f(x) + g(x), \tag{1}$$

where both functions $f, g\colon \mathbb{R}^n \to \bar{\mathbb{R}} := (-\infty, \infty]$ are proper, convex, and lower semicontinuous (l.s.c.), while being structurally different from each other. Namely, f is assumed to be twice continuously differentiable with the Lipschitz continuous gradient ∇f on its domain. On the other hand, g is merely continuous on its domain; see Assumption 1.1 for the precise formulations. It has been well recognized that model (1), known as a *composite convex optimization problem*, frequently appears in a variety of applications including, e.g., machine learning, signal processing, and statistics, where f is a *loss function* and g is a *regularizer*; we keep this terminology here. Note that problem (1) contains in fact implicit constraints written as $x \in \Omega := \operatorname{dom} g$.

It is typical in applications that problems of type (1) have a large size, which makes attractive to compute their solutions by employing first-order algorithms such as the *proximal gradient method* (PGM). Given each iterate x^k, the PGM constructs a new x^{k+1} by solving the following optimization subproblem, which approximates the smooth function f in (1) by the linear model:

$$\min_{x\in\mathbb{R}^n} l_k(x) + \frac{1}{2t}\|x - x_k\|^2 \ \text{ with } \ l_k(x) := f(x^k) + \nabla f(x^k)^T(x - x^k) + g(x), \tag{2}$$

where ^{T} indicates the matrix transposition, and where $t > 0$ represents the step size of PGM. As well known, the PGM applied to (1) generates a sequence of iterates that converges at least sublinearly of rate $O(1/k)$ (see, e.g., [1, 2]) and linearly with respect to the sequence of cost function values—provided that f is strongly convex; see e.g., [3]. Refined results on linear convergence of the PGM are derived under various error bound conditions as in [4–8].

When f is a twice continuously differentiable function, it is natural to expect algorithms having faster convergence rates by exploiting the Hessian $\nabla^2 f(x^k)$ of f at each iterate x^k and constructing the next iterate x^{k+1} as a solution to the following quadratic subproblem:

$$\min_{x\in\mathbb{R}^n} q_k(x) := f(x^k) + \nabla f(x^k)^T(x - x^k) + \frac{1}{2}(x - x^k)^T H_k(x - x^k) + g(x), \tag{3}$$

where H_k is an appropriate approximation of the Hessian $\nabla^2 f(x^k)$. Methods of this type to solve composite optimization problems (1) are unified under the name of *proximal Newton-type methods*; see, e.g., [9]. To the best of our knowledge, the origin of such methods to solve nonsmooth composite optimization problems given in form (1) can be traced back to the generalized proximal point method developed by Fukushima and Mine [10] who in turn considered it as an extension of Rockafellar's proximal point method [11] to find zeros of maximal monotone operators and subgradient inclusions associated with convex functions. On the other hand, the general

scheme of *successive quadratic approximations* to solve optimization-related problems is a common idea of Newton-type and quasi-Newton methods; see the books [12, 13] with their bibliographies. For particular subclasses of composite problems (1), the quadratic approximation scheme (3) contains special versions of the proximal Newton-type methods known as GLMNET [14], newGLMNET [15], QUIC [16], the Newton-LASSO method [17], the projected Newton-type algorithms [3, 18], etc.

Observe further that, due to the convexity of both functions f and g with f being smooth, problem (1) can be equivalently written as the *generalized equation*

$$0 \in \nabla f(x) + \partial g(x) \tag{4}$$

in the sense of Robinson [19], where $\partial g(x)$ is the subdifferential of g at x, and where the used subdifferential sum rule does not require any qualification conditions due to the smoothness of f; see, e.g., [20, Proposition 1.30]. Then subproblem (3) for constructing the new iterate x^{k+1} in the proximal Newton method for (4) reduces to solving the following *partially linearized* generalized equation at the given iterate x^k:

$$0 \in \nabla f(x^k) + H_k(x - x^k) + \partial g(x). \tag{5}$$

Various results on the local superlinear and quadratic convergence of iterative sequences $\{x^k\}$ for (5) are obtained in the literature in the framework of quasi-Newton methods for generalized equations under different kinds of regularity conditions imposed on ∂F from (1); see, e.g., the books [12, 13, 21] with the references and discussions therein. In particular, Fischer [22] proposes an iterative procedure to solve generalized equations and proves local superlinear and quadratic convergence of iterates under a certain Lipschitz stability property of the corresponding perturbed solution map. More specifically, paper [22] develops a quasi-Newton algorithm to solve (1) in the framework of (5) that exhibits a local superlinear/quadratic convergence in the setting where g is the indicator function of a box constraint, and where H_k in (3) is taken as the regularized Hessian $H_k := \nabla^2 f(x^k) + \alpha_k I$ with $\{\alpha_k\}$ being a positive vanishing sequence satisfying certain conditions. The main assumptions of [22] include the local Lipschitz continuity of the Hessian $\nabla^2 f(x)$ and the upper Lipschitz continuity/calmness of the perturbed solution map (1) at the points in question.

However, how to build a reasonable *globalization* of the local scheme given by (3) has not been completely resolved yet. Various globalizations of the proximal Newton method can be found in the literature, see, e.g., [9, 23–25]. Unfortunately, all these works require f to be *strongly convex*. In particular, paper by Byrd et al. [23], which addresses the special case of problem (1) with $g := \lambda\|x\|_1$ and $\lambda > 0$, proposes implementable inexactness conditions and backtracking line search procedures to design a globally convergent proximal Newton method, but the local superlinear and quadratic convergence results therein are established under the strong convexity assumption on f. Quite recently, in [26], the inexactness conditions and backtracking line search procedures of [23] is applied to develop a proximal Newton method for (1) with proving its local convergence of superlinear and quadratic rates by using the Luo-Tseng error bound condition [5] instead of the strong convexity assumption in [23]. However, the convergence results in [26] have a crucial flaw. To achieve a local quadratic

convergence rate, the authors of [26] require that parameters of their method satisfy a certain condition involving the constant in the error bound, which is extremely challenging to estimate. Note to this end that the strong convexity assumption has not been imposed by using some other Newton-type algorithms such as the one based on the forward-backward envelope (FBE), which is different from the proximal Newton-type method developed below; see, e.g., [27] and the references therein.

In this paper we design a new globally convergent proximal Newton-type algorithm to solve composite convex optimization problems of class (1) under the following *standing assumptions* on the given data without requiring the strong convexity of the loss function f:

Assumption 1.1 Impose the following properties of the loss function and the regularizer in (1):

(1) Both functions $f, g : \mathbb{R}^n \to (-\infty, \infty]$ are proper, l.s.c., and convex.
(2) The effective domain of the loss function $\operatorname{dom} f := \{x \mid f(x) < \infty\}$ is open, and $f(x)$ is twice continuously differentiable on a closed set $\Omega \supset \operatorname{dom} f$.
(3) The regularizer $g(x)$ is continuous on its domain and $\varnothing \neq \operatorname{dom} g \subset \operatorname{dom} f$.
(4) The gradient $\nabla f(x)$ is Lipschitz continuous on a closed set Ω from (ii) with Lipschitz constant $L_1 > 0$.
(5) Problem (1) has a nonempty solution set $\mathcal{X}^* := \arg\min_{x\in\mathbb{R}^n} F(x)$ with the optimal value F^*.

Basic convex analysis tells us that the imposed assumptions (2) and (3) ensure the fulfillment of the subdifferential sum rule $\partial F(x) = \nabla f(x) + \partial g(x)$ for all $x \in \operatorname{dom} g$; see, e.g., [28, Corollary 2.45].

Our *main contributions* can be summarized as follows:

(1) We develop a *globally convergent* proximal Newton-type algorithm to solve (1) with an *implementable inexact condition* for subproblem (3) and a new reasonable *backtracking line search* strategy. Our line search procedure does not require any restrictive assumptions. It is shown in this way that if the subgradient mapping ∂F is *metrically subregular* at some limiting point of the iterative sequence, the backtracking line search procedure accepts a unit step size when the iterates are close to the solution. Furthermore, we prove that the proposed proximal Newton-type algorithm exhibits a *local convergence* with the *quadratic convergence rate*. Numerical experiments are performed to solve the l_1 *regularized logistic regression* problem that illustrate the efficiency of the proposed algorithm.
(2) We establish novel local convergence results for the proposed algorithm under the *metric q-subregularity* assumption imposed on the subgradient mapping ∂F for any positive number $q > \frac{1}{2}$. If $q < 1$, the obtained results require *less restrictive assumptions* in comparison with the case of metric subregularity ($q = 1$) to ensure a superlinear convergence of iterates, while for $q > 1$ we achieve a convergence rate that is *higher than quadratic*.
(3) When the loss function f in (1) satisfies additional structural assumptions, we obtain a local superlinear convergence rate of our proposed algorithm *without imposing the Lipschitz continuity* of the Hessian matrix $\nabla^2 f(x)$. The latter assumption is crucial for establishing a fast convergence of the previously known algorithms of the proximal Newton type.

The rest of the paper is organized as follows. Section 2 briefly overviews the notions and results of variational analysis needed for the subsequent material. An concrete example is also given to justify our motivation of using metric subregularity instead of the Luo-Tseng error bound condition. In Sect. 3 we present our proximal Newton-type algorithm and establish its global convergence. In Sect. 4, for the cases where $q \in (0, 1]$ and $q > 1$, we separately derive local fast convergence results under the metric q-subregularity of ∂F. Specially, local superlinear and quadratic convergence of the proposed algorithm under metric subregularity are given. Section 5 is devoted to problem (1) with a certain structure of the loss function f and establishes in this case a superlinear convergence of the proposed algorithm without the Lipschitz continuity of the loss function Hessian. Finally, Sect. 6 conducts and analyzes numerical experiments to solve the practically important l_1 regularized logistic regression problem by implementing the designed proximal Newton-type method.

2 Preliminaries from variational analysis

Here we recall and discuss some material from variational analysis that is broadly used in what follows. The reader can find more details and references in the books [20, 21, 29].

Throughout the paper, we use the standard notation. Recall that $\mathbb{R}^n$ signifies an n-dimensional Euclidean space with the inner product $\langle\cdot,\cdot\rangle$ and the norm denoted by $\|\cdot\|$, while the 1-norm is signified by $\|\cdot\|_1$. For any matrix $A \in \mathbb{R}^{m\times n}$ we have $\|A\| := \max_{x\neq 0} \frac{\|Ax\|}{\|x\|}$ with $\tilde{\sigma}_{\min}(A)$ standing for the smallest nonzero singular value of A. The symbols $\mathbb{B}_r(x)$ and $\overline{\mathbb{B}}_r(x)$ denote the open and the closed Euclidean norm ball centered at x with radius $r > 0$, respectively, while we use $\mathbb{B}$ and $\overline{\mathbb{B}}$ for the corresponding unit balls around the origin. Given a nonempty subset $\Omega \subset \mathbb{R}^n$, denote by $\operatorname{bd}\Omega$ its boundary and consider the associated distance function $\operatorname{dist}(x;\Omega) := \inf\{\|x-y\| \,\big|\, y \in \Omega\}$ and the indicator function $\delta_\Omega(x)$ equal 0 if $x \in \Omega$ and ∞ otherwise. The graph of a set-valued mapping/multifunction $\Psi\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is given by $\operatorname{gph}\Psi := \{(x,\upsilon) \in \mathbb{R}^n \times \mathbb{R}^m \mid \upsilon \in \Psi(x)\}$, and the inverse to Ψ is $\Psi^{-1}(\upsilon) := \{x \in \mathbb{R}^n \mid \upsilon \in \Psi(x)\}$.

The following fundamental properties of set-valued mappings are employed in the paper to establish fast local convergence results for the proposed proximal Newton-type algorithm.

Definition 2.1 Let $\Psi\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ be a set-valued mapping, let $(\bar{x},\bar{\upsilon}) \in \operatorname{gph}\Psi$, and let $q > 0$.

(1) We say that Ψ is METRICALLY q- SUBREGULAR at $(\bar{x},\bar{\upsilon})$ with modulus $\kappa > 0$ if there is $\epsilon > 0$ such that

$$\operatorname{dist}\big(x;\Psi^{-1}(\bar{\upsilon})\big) \le \kappa \operatorname{dist}\big(\bar{\upsilon};\Psi(x)\big)^q \quad \text{for all } \ x \in \mathbb{B}_\epsilon(\bar{x}). \tag{6}$$

(2) Ψ is said to be METRICALLY SUBREGULAR at $(\bar{x},\bar{\upsilon})$ if $q = 1$ in (6).

(3) We say that Ψ is STRONGLY METRICALLY q- SUBREGULAR at $(\bar{x}, \bar{v})$ with modulus $\kappa > 0$ if there exists $\epsilon > 0$ such that

$$\|x - \bar{x}\| \le \kappa \operatorname{dist}\big(\bar{v}; \Psi(x)\big)^q \quad \text{for all} \quad x \in \mathbb{B}_\epsilon(\bar{x}).$$

The metric subregularity property has been well recognized and applied in variational analysis and optimization numerical aspects. The reader can find more information and references in [20, 21] with the commentaries and the bibliographies therein. In this paper we employ metric subregularity of *subgradient mappings*, which form a remarkable class of multifunctions with special properties. Various sufficient conditions and characterizations of this property of subgradient mappings are given in [30–32] in terms of certain second-order growth conditions imposed on the function in question.

The metric q-subregularity of order $q \in (0, 1)$, known also as *Hölder metric subregularity*, is much less investigated, while some verifiable conditions for the fulfillment of this property can be found in, e.g., [33–35]. Note that the Hölder metric subregularity is clearly a weaker assumption in comparison with the standard metric subregularity property.

The case of *higher-order metric subregularity* with $q > 1$ in (6) is largely open in the literature. One of the reasons for this is that the corresponding *metric q-regularity* property with $q > 1$ does not make sense, since it holds only for constant mappings. Nevertheless, it is shown in [36] that the higher-order metric subregularity is a useful property in variational analysis and optimization. This property is characterized for subgradient mappings in [36] via a higher-order growth condition, and its *strong* version from Definition 2.1(iii) is applied therein to the convergence analysis of quasi-Newton methods for generalized equations.

Next we consider the *proximal mapping*

$$\operatorname{Prox}_g(u) := \operatorname{argmin}\Big\{g(x) + \frac{1}{2}\|x - u\|^2 \;\Big|\; x \in \mathbb{R}^n\Big\}, \quad u \in \mathbb{R}^n, \tag{7}$$

associated with a proper function $g\colon \mathbb{R}^n \to \bar{\mathbb{R}}$. A crucial role of proximal mappings has been well recognized not only in proximal Newton-type algorithms (see, e.g., [9, 23]), but also in other second-order methods of numerical optimization. In particular, we refer the reader to the very recent papers [37, 38], where the proximal mappings are used for designing superlinearly convergent Newton-type algorithms to find tilt-stable local minimizers of nonconvex extended-real-valued functions and to solve subgradient inclusions in a large generality. If g is l.s.c. and convex, then the proximal mapping (7) is single-valued and nonexpansive on $\mathbb{R}^n$, i.e., Lipschitz continuous with constant one; see, e.g., [29, Theorem 12.12].

It is important to emphasize that in many practical models of type (1) arising, in particular, in machine learning and statistics, the proximal mapping associated with the regularizer term g (e.g., when g is the l_1-norm, the group Lasso regularizer, etc.) can be easily computed. This is the case of the l_1 regularized logistic regression problem in our applications developed in Sect. 6.

Having (7), define further the *prox-gradient mapping* associated with (1) by

$$\mathcal{G}(x) := x - \mathrm{Prox}_g\big(x - \nabla f(x)\big), \quad x \in \mathbb{R}^n, \tag{8}$$

and present some properties of (8) used in what follows. Note that $\mathcal{G}(x)$ is generally defined in terms of a positive parameter $L > 0$ as $\mathcal{G}_L(x) := x - \mathrm{Prox}_{\frac{1}{L}g}\big(x - \frac{1}{L}\nabla f(x)\big)$. In order to concentrate on the main idea, we simply set $L = 1$ throughout this paper. All of the results in this paper can be easily extended to the case with a given positive L. Thanks to the convexity, we have that $\mathcal{G}(x) = 0$ if and only if $x \in \mathcal{X}^*$. The following proposition provides an upper estimate for $\|\mathcal{G}(x)\|$ by $\mathrm{dist}(0; \nabla f(x) + \partial g(x))$ and shows that $\mathcal{G}(x)$ is Lipschitz continuous. It can be seen as a direct combination of [39, Theorem 3.5] and [40, Lemma 10.10].

Proposition 2.1 *Let ∇f be Lipschitz continuous with modulus L_1 on $\mathbb{R}^n$. Then we have the estimates*

$$\begin{aligned} \|\mathcal{G}(x)\| &\le \mathrm{dist}(0; \nabla f(x) + \partial g(x)) \ \text{ for all } \ x \in \mathrm{dom}\, f, \\ \|\mathcal{G}(x) - \mathcal{G}(y)\| &\le (2 + L_1)\|x - y\| \ \text{ for any } \ x, y \in \mathrm{dom}\, f. \end{aligned}$$

The next proposition is a combination of [39, Theorems 3.4 and 3.5].

Proposition 2.2 *Let ∇f be Lipschitz continuous with modulus L_1 around $\bar{x}$, and let the mapping $\nabla f(x) + \partial g(x)$ be metrically subregular at $(\bar{x}, 0)$, i.e., there exist numbers $\epsilon, \kappa > 0$ such that*

$$\mathrm{dist}(x; \mathcal{X}^*) \le \kappa\, \mathrm{dist}\big(0; \nabla f(x) + \partial g(x)\big) \ \text{ for all } \ x \in \mathbb{B}_\epsilon(\bar{x}).$$

Then whenever $x \in \mathbb{B}_\epsilon(\bar{x})$ we have the estimate

$$\mathrm{dist}(x; \mathcal{X}^*) \le (1 + \kappa)(1 + L_1)\|\mathcal{G}(x)\|.$$

The following proposition gives a reverse statement to Proposition 2.2 while providing an estimate of the norm of (8) via the distance to the solution set of the convex composite problem (1).

Proposition 2.3 *Let ∇f be Lipschitz continuous with modulus L_1 on $\mathbb{R}^n$. Then we have the estimate*

$$\|\mathcal{G}(x)\| \le (2 + L_1)\mathrm{dist}(x; \mathcal{X}^*) \ \text{ for all } \ x \in \mathrm{dom}\, f.$$

Proof Observe first that the mapping $\mathcal{G}(x)$ is well-defined and single-valued for all $x \in \mathrm{dom}\, f$ due to the aforementioned result of [29]. It easily follows from Assumption 1.1 that the nonempty solution set $\mathcal{X}^*$ is closed and convex; hence each point $x \in \mathbb{R}^n$ has the unique projection $\pi_x \in \mathcal{X}^*$ onto $\mathcal{X}^*$. Note that $\mathcal{G}(\pi_x) = \pi_x - \mathrm{Prox}_g(\pi_x - \nabla f(\pi_x)) = 0$ for $\pi_x \in \mathcal{X}^*$. Thus we verify the claim of the proposition by

$$\|\mathcal{G}(x)\| = \|\mathcal{G}(x) - \mathcal{G}(\pi_x)\| \le (2 + L_1)\|x - \pi_x\|, \quad x \in \mathrm{dom}\, f,$$

where the inequality holds since $\mathcal{G}(x)$ is $(2+L_1)$-Lipschitz continuous by Proposition 2.1. □

Next we obtain an extension of Proposition 2.2 to case where the subgradient mapping $\nabla f + \partial g$ in (1) satisfies the Hölder subregularity property in the point in question.

Proposition 2.4 *Let ∇f be Lipschitz continuous with modulus L_1 around $\bar{x}$, and let the mapping $\nabla f(x) + \partial g(x)$ be metrically q-subregular at $(\bar{x}, 0)$ with $q \in (0, 1]$, i.e., there exist $\epsilon_1, \kappa_1 > 0$ such that*

$$\operatorname{dist}(x; \mathcal{X}^*) \le \kappa_1 \operatorname{dist}\big(0; \nabla f(x) + \partial g(x)\big)^q \quad \text{for all} \quad x \in \mathbb{B}_{\epsilon_1}(\bar{x}).$$

Then we find constants $\epsilon_2, \kappa_2 > 0$ that ensure the estimate

$$\operatorname{dist}(x; \mathcal{X}^*) \le \kappa_2 \|\mathcal{G}(x)\|^q \quad \text{whenever} \quad x \in \mathbb{B}_{\epsilon_2}(\bar{x}). \tag{9}$$

Proof By (8) we have the inclusions

$$\begin{aligned} &\mathcal{G}(x) \in \nabla f(x) + \partial g\big(x - \mathcal{G}(x)\big) \ \text{ and} \\ &\mathcal{G}(x) + \nabla f\big(x - \mathcal{G}(x)\big) - \nabla f(x) \in \nabla f\big(x - \mathcal{G}(x)\big) + \partial g\big(x - \mathcal{G}(x)\big) \end{aligned}$$

for all $x \in \operatorname{dom} f$. When $x \in \mathbb{B}_{\epsilon_1}(\bar{x}) \cap \operatorname{dom} f$, it follows from the imposed assumption that

$$\begin{aligned} \operatorname{dist}\big(x - \mathcal{G}(x); \mathcal{X}^*\big) &\le \kappa_1 \operatorname{dist}\big(0; \mathcal{G}(x) + \nabla f\big(x - \mathcal{G}(x)\big) - \nabla f(x)\big)^q \\ &\le \kappa_1 (1 + L_1)^q \|\mathcal{G}(x)\|^q, \end{aligned}$$

which leads us to the resulting estimates for such x:

$$\begin{aligned} \operatorname{dist}(x; \mathcal{X}^*) &\le \operatorname{dist}\big(x - \mathcal{G}(x); \mathcal{X}^*\big) + \|\mathcal{G}(x)\| \\ &\le (1 + \kappa_1 (1 + L_1)^q) \max\big\{\|\mathcal{G}(x)\|, \|\mathcal{G}(x)\|^q\big\}. \end{aligned}$$

Applying now Proposition 2.3 tells us that, whenever $\operatorname{dist}(x; \mathcal{X}^*) \le 1/(2 + L_1)$ and $x \in \operatorname{dom} f$, we get

$$\|\mathcal{G}(x)\| \le (2 + L_1) \operatorname{dist}(x; \mathcal{X}^*) \le 1.$$

Letting $\epsilon_2 := \min\{1/(2 + L_1), \epsilon_1\}$ and remembering that $q \le 1$ bring us to the inequality

$$\operatorname{dist}(x; \mathcal{X}^*) \le (1 + \kappa_1 (1 + L_1)^q) \|\mathcal{G}(x)\|^q \quad \text{for all} \quad x \in \mathbb{B}_{\epsilon_2}(\bar{x}),$$

which verifies (9) with $\kappa_2 := (1 + \kappa_1 (1 + L_1)^q)$ and thus completes the proof of the proposition. □

Finally, we establish a sufficient condition for the metric q-subregularity of the subgradient mapping $\nabla f(x) + \partial g(x)$. Recall first the following characterization of metric q-subregularity, which is a direct specification of [36, Theorem 3.4] in the convex case.

Proposition 2.5 *Let $\bar{x} \in \operatorname{dom} F$, $\bar{v} \in \partial F(\bar{x})$, and $q > 0$. Then we have the equivalent statements:*

(1) ∂F is metrically q-subregular at $(\bar{x}, \bar{v})$.
(2) There are two positive numbers ϵ and c such that

$$F(x) \geq F(\bar{x}) + \langle \bar{v}, x - \bar{x} \rangle + c \cdot \operatorname{dist}\big(x; (\partial F)^{-1}(\bar{v})\big)^{\frac{1+q}{q}} \quad \text{for all} \quad x \in \mathbb{B}_\epsilon(\bar{x}).$$

The aforementioned sufficient conditions for the metric q-subregularity of the subgradient mapping $\nabla f(x) + \partial g(x)$ as formulated as follows.

Proposition 2.6 *Let $\bar{x} \in \mathcal{X}^*$, and let $q > 0$. Suppose ∂g is strongly metrically q-subregular at $(\bar{x}, -\nabla f(\bar{x}))$. Then $\nabla f(x) + \partial g(x)$ is metrically q-subregular at $(\bar{x}, 0)$.*

Proof Since ∂g is strongly metrically q-subregular at $(\bar{x}, -\nabla f(\bar{x}))$, we have $(\partial g)^{-1}(-\nabla f(\bar{x})) = \{\bar{x}\}$, and Proposition 2.5 gives us positive numbers ϵ and c such that

$$g(x) \geq g(\bar{x}) + \langle -\nabla f(\bar{x}), x - \bar{x} \rangle + c\|x - \bar{x}\|^{\frac{1+q}{q}} \quad \text{for all} \quad x \in \mathbb{B}_\epsilon(\bar{x}).$$

The convexity of f implies that

$$f(x) \geq f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle \quad \text{for all} \quad x \in \mathbb{R}^n.$$

Summing up the above two inequalities gives us

$$F(x) \geq F(\bar{x}) + c\|x - \bar{x}\|^{\frac{1+q}{q}} \geq F(\bar{x}) + c \cdot \operatorname{dist}(x; \mathcal{X}^*)^{\frac{1+q}{q}} \quad \text{for all} \quad x \in \mathbb{B}_\epsilon(\bar{x}).$$

Then the conclusion of the proposition immediately follows from Proposition 2.5. □

Metric subregularity is a weaker assumption than the Luo-Tseng error bound condition. We conclude this section by giving a specific example verifying this statement.

Example 1 Consider the following problem of composite convex optimization:

$$min_{x \in \mathbb{R}^n} f(x) + g(x) \quad \text{with} \quad f(x) := c^T x \quad \text{and} \quad g(x) := \|x\|,$$

where $c \in \mathbb{R}^n$ is such that $\|c\| = 1$. Problems of this type frequently appear in statistical learning models. It can be easily calculated that the optimal value is $F^* = 0$ and the optimal solution is $\mathcal{X}^* = \{-\gamma c \mid \gamma \geq 0\}$. We know that for any $\bar{x} \in \mathcal{X}^*$ the mapping

$$\nabla f(x) + \partial g(x) = c + \partial \|x\|$$

is metrically subregular at $(\bar{x}, 0)$ since $\partial\|x\|$ is metrically subregular on its graph; see, e.g., [8, Lemma 4].

On the other hand, the Luo-Tseng error bound fails. Indeed, if this condition holds, then there exist $\kappa, \epsilon, \eta > 0$ such that for any x satisfying $F(x) \le \eta$ and $\|\mathcal{G}(x)\| \le \epsilon$ we get

$$\operatorname{dist}(x; \mathcal{X}^*) \le \kappa\|\mathcal{G}(x)\|.$$

Let $c_k \to c$ with $\|c_k\| = 1, k \in \mathbb{N}$. Setting $x_k := \gamma_k(-c_k)$ with $\gamma_k := \frac{1}{\sqrt{-c^T c_k + 1}} \to \infty$ gives us

$$\begin{aligned} F(x_k) &= \gamma_k\left(-c^T c_k + \|c_k\|\right) = \sqrt{-c^T c_k + 1} \to 0,\\ c - c_k &= c + \frac{x_k}{\|x_k\|} \in \nabla f(x_k) + \partial g(x_k). \end{aligned}$$

It follows from Proposition 2.1 that we have the estimate

$$\|\mathcal{G}(x_k)\| \le \operatorname{dist}\big(0; \nabla f(x) + \partial g(x)\big) \le \|c - c_k\| \to 0.$$

However, letting $\theta_k := \arccos(\frac{c^T(c-c_k)}{\|c-c_k\|}) \to \pi/2$ tells us that

$$\frac{\operatorname{dist}(x_k; \mathcal{X}^*)}{\|\mathcal{G}(x_k)\|} \ge \frac{\gamma_k\sqrt{1 - (c^T c_k)^2}}{\|c - c_k\|} = \gamma_k \sin(\theta_k) \to \infty,$$

which clearly contradicts the Luo-Tseng error bound condition.

3 The new algorithm and its global convergence

In this section we describe the proposed proximal Newton-type algorithm to solve the class of composite convex optimization problems (1) with justifying its global convergence under the standing assumptions.

Given a current iterate x^k for each $k = 0, 1, \ldots$, we select a *positive semidefinite matrix* B_k as an arbitrary approximation of the Hessian $\nabla^2 f(x^k)$ satisfying the *standing boundedness assumption*:

$$\text{there exists } M \ge 0 \text{ such that } \|B_k\| \le M \text{ whenever } k = 0, 1, \ldots. \tag{10}$$

If the gradient mapping ∇f is uniformly Lipschitz continuous along the sequence of iterates with constant L_1, then (10) holds for $B_k = \nabla^2 f(x^k)$ with $M = L_1$. In the general case of B_k, pick any constants $c > 0$ and $\rho \in (0, 1]$ and, using the prox-regular mapping (8), consider the positive number $\alpha_k := c\|\mathcal{G}(x^k)\|^\rho$ and define the *quasi-Newton approximation* of the Hessian of f at x^k by

$$H_k := B_k + \alpha_k I \text{ for all } k = 0, 1, \ldots, \tag{11}$$

which is a positive definite matrix. Then similarly to [23], but with the different approximation (11), denote

$$r_k(x) := x - \mathrm{Prox}_g\Big(x - \nabla f(x^k) - H_k(x - x^k)\Big) \tag{12}$$

and select $\hat{x}^k$ as an *approximate minimizer* of the quadratic *subproblem* for (1) given by

$$\min_{x\in\mathbb{R}^n} q_k(x) := f(x^k) + \nabla f(x^k)^T(x - x^k) + \frac{1}{2}(x - x^k)H_k(x - x^k) + g(x) \tag{13}$$

with the *residual* number $\|r_k(\hat{x}^k)\|$ measuring the approximate optimality of $\hat{x}^k$ in (13). Observing that $\|r_k(\hat{x}^k)\| = 0$ if and only if $\hat{x}^k$ is an *exact* solution to subproblem (13), we use the nonnegative number $\|r_k(\hat{x}^k)\|$ with $r_k(x)$ taken from (12) as the *optimality measure* of $\hat{x}^k$ in subproblem (13). Adapting the scheme of [23] in our new setting, let us impose the following two estimates as *inexact conditions* for choosing $\hat{x}^k$ as an approximate solution to subproblem (13):

$$\|r_k(\hat{x}^k)\| \le \eta_k\|\mathcal{G}(x^k)\| \ \text{ and } \ q_k(\hat{x}^k) \le q_k(x^k) \tag{14}$$

with the parameter $\eta_k := \nu\min\{1, \|\mathcal{G}(x^k)\|^{\varrho}\}$ defined via (12) and some numbers $\nu \in [0, 1)$ and $\varrho > 0$. Since any exact solution to subproblem (13) always fulfills the two inexact conditions in (14), a point $\hat{x}^k$ satisfying (14) always exists. For many application problems, the mapping Prox_g is easy to calculate, thus inexact conditions (14) can be readily verified. In the case where $\mathrm{dist}\big(0; \partial q_k(\hat{x}^k)\big)$ is easy to estimate, because $\|r_k(\hat{x}^k)\| \le \mathrm{dist}\big(0; \partial q_k(\hat{x}^k)\big)$ always holds (see, e.g., [39, Theorem 3.5]), we can use $\mathrm{dist}\big(0; \partial q_k(\hat{x}^k)\big) \le \eta_k\|\mathcal{G}(x^k)\|$ for verifying the first inexact condition in (14).

Using the above constructions and the line search procedure inspired by [41, 42], we are ready to propose the proximal Newton-type algorithm designed as in Algorithm 1. In the proposed Algorithm 1, we add Step 4, that is inspired by [41, 42], to check whether the prox-gradient residual $\|\mathcal{G}(\hat{x}^k)\|$ of the inexact solution $\hat{x}^k$ to subproblem (3) decreases to be under a given fixed ratio, which is smaller than one, times the previous value. If it does, then we will update the next iterate x^{k+1} by using the Newton direction d^k with a unit step size to let $x^{k+1} = \hat{x}^k$ and skip the backtracking line search in Step 5. Otherwise, we use the conventional backtracking line search procedure in Step 5 to find a conservative step size for updating the next iterate. It is shown in Theorem 4.1 below that Step 4 always gives us a unit step size when the iterate x^k is close to the solution under the metric q-subregularity condition.

In the rest of this section, we show that the proposed algorithm *globally converges* under the mild standing assumptions, which are imposed above and will not be repeated. Let us start with the following lemma providing a subgradient estimate for subproblem (13) at the approximate solution.

Algorithm 1 Proximal Newton-type method

1: Choose $x^0 \in \mathbb{R}^n$, $0 < \theta, \sigma, \gamma < 1$, $C > F(x^0)$, $\bar{\alpha}, c > 0$, and $\rho \in (0, 1]$.
2: **for** $k = 0, 1, \ldots$ **do**

1. Update the approximation of the Hessian matrix B_k.
2. Form the quadratic model (3) with $H_k := B_k + \alpha_k I$ and $\alpha_k := \min\left\{\bar{\alpha}, c\|\mathcal{G}(x^k)\|^\rho\right\}$.
3. Obtain an inexact solution $\hat{x}^k$ of (3) satisfying the conditions in (14).
4. If $k = 0$, let $\vartheta_1 := \mathcal{G}(x^0)$ and go to Step 5. For $k \geq 1$, if $\|\mathcal{G}(\hat{x}^k)\| \leq \sigma\vartheta_k$ and $F(\hat{x}^k) \leq C$, let $t_k := 1$, $\vartheta_{k+1} := \|\mathcal{G}(\hat{x}^k)\|$, and go to Step 6. Otherwise, let $\vartheta_{k+1} := \vartheta_k$ and go to Step 5.
5. Perform a backtracking line search along the direction $d^k := \hat{x}^k - x^k$ by setting $t_k := \gamma^{m_k}$, where m_k is the smallest nonnegative integer m such that

$$F(x^k + \gamma^m d^k) \leq F(x^k) - \theta\alpha_k\gamma^m\|d^k\|^2. \tag{15}$$

6. Set $x^{k+1} := x^k + t_k d^k$.

3: **end for**

Lemma 3.1 *Given an approximate solution $\hat{x}^k$ to (13), there exists a vector $e_k \in \mathbb{R}^n$ such that*

$$\begin{aligned} e_k &\in \nabla f(x^k) + H_k(\hat{x}^k - x^k) + \partial g(\hat{x}^k - e_k) \ \text{ and } \ \|e_k\| \\ &\leq \nu \min\left\{\|\mathcal{G}(x^k)\|, \|\mathcal{G}(x^k)\|^{1+\varrho}\right\}. \end{aligned} \tag{16}$$

Proof Let $e_k := r_k(\hat{x}^k) = \hat{x}^k - \text{Prox}_g(\hat{x}^k - \nabla f(x^k) - H_k(\hat{x}^k - x^k))$. Then we have

$$e_k \in \nabla f(x^k) + H_k(\hat{x}^k - x^k) + \partial g(\hat{x}^k - e_k),$$

which follow from (7). Using finally the inexact conditions (14) for $\hat{x}^k$, we verify the claim of the lemma. □

The next lemma provides elaborations on Step 5 of the proposed algorithm with the decreasing of the cost function in (1) by the backtracking line search.

Lemma 3.2 *Let t_k be chosen by the backtracking line search in Step 5 of Algorithm 1 at iteration k. Then we have the step size estimate*

$$t_k \geq \frac{\gamma(1-\theta)\alpha_k}{L_1} \tag{17}$$

with the cost function decrease satisfying

$$F(x^{k+1}) - F(x^k) \leq -\frac{\gamma\theta(1-\theta)}{2L_1}\left(\frac{(1-\nu)\alpha_k}{1+M+\alpha_k}\right)^2\|\mathcal{G}(x^k)\|^2. \tag{18}$$

Proof Since $\hat{x}^k$ is an inexact solution to (3) obeying the conditions in (14), it follows that

$$0 \geq q_k(\hat{x}^k) - q_k(x^k) = l_k(\hat{x}^k) - l_k(x^k) + \frac{1}{2}(\hat{x}^k - x^k)^T H_k(\hat{x}^k - x^k),$$

where l_k is the linear part of q_k defined in (2). This yields

$$l_k(x^k) - l_k(\hat{x}^k) \geq \frac{1}{2}(\hat{x}^k - x^k)^T H_k(\hat{x}^k - x^k) \geq \frac{1}{2}\alpha_k \|\hat{x}^k - x^k\|^2. \tag{19}$$

By $\mathcal{G}(x^k) = x^k - \mathrm{Prox}_g(x^k - \nabla f(x^k))$ we deduce from the stationary and subdifferential sum rules that

$$\mathcal{G}(x^k) \in \nabla f(x^k) + \partial g\big(x^k - \mathcal{G}(x^k)\big).$$

Furthermore, Lemma 3.1 gives us the condition $e_k \in \nabla f(x^k) + H_k(\hat{x}^k - x^k) + \partial g(\hat{x}^k - e_k)$ for $\hat{x}^k$ with e_k satisfying the estimate $\|e_k\| \leq \nu\|\mathcal{G}(x^k)\|$. The monotonicity of the subgradient mapping ∂g ensures that

$$\big\langle \mathcal{G}(x^k) + H_k(\hat{x}^k - x^k) - e_k, x^k - \mathcal{G}(x^k) - \hat{x}^k + e_k \big\rangle \geq 0,$$

which therefore leads us to the inequality

$$\begin{aligned}\|\mathcal{G}(x^k) - e_k\|^2 &\leq \|\mathcal{G}(x^k)\|^2 - 2\langle e_k, \mathcal{G}(x^k)\rangle + \|e_k\|^2 + (\hat{x}^k - x^k)^T H_k(\hat{x}^k - x^k)\\ &\leq \big\langle \mathcal{G}(x^k) - e_k, x^k - \hat{x}^k + H_k(x^k - \hat{x}^k)\big\rangle\\ &\leq \big\|\mathcal{G}(x^k) - e_k\big\| \cdot \big\|x^k - \hat{x}^k + H_k(x^k - \hat{x}^k)\big\|.\end{aligned}$$

Using again the condition $\|e_k\| \leq \nu\|\mathcal{G}(x^k)\|$ together with $\|B_k\| \leq M$ from (10) results in

$$\|\mathcal{G}(x^k)\| \leq \|\mathcal{G}(x^k) - e_k\| + \|e_k\| \leq (1 + M + \alpha_k)\|\hat{x}^k - x^k\| + \nu\|\mathcal{G}(x^k)\|.$$

Remembering the choice of $\nu \in [0, 1)$, we estimate the prox-gradient mapping (8) at the iterate x^k by

$$\|\mathcal{G}(x^k)\| \leq \frac{1 + M + \alpha_k}{1 - \nu}\|\hat{x}^k - x^k\|. \tag{20}$$

Next let us show that the backtracking line search along the direction $d^k = \hat{x}^k - x^k$ in Step 5 is well-defined and the proposed step size ensures a sufficient decrease in the cost function F. It follows from the Lipschitz continuity of ∇f that

$$f(x^k + \tau d^k) \leq f(x^k) + \tau \nabla f(x^k)^T d^k + \frac{L_1}{2}\tau^2\|d^k\|^2 \ \text{ for any } \ \tau \geq 0,$$

and thus we deduce from the definition of l_k in (2) that

$$F(x^k) - F(x^k + \tau d^k) \geq l_k(x^k) - l_k(x^k + \tau d^k) - \frac{L_1}{2}\tau^2\|d^k\|^2.$$

This implies by the convexity of g that

$$l_k(x^k) - l_k(x^k + \tau d^k) \geq \tau\big(l_k(x^k) - l_k(x^k + d^k)\big).$$

Combining the latter with (19) and using the choice of $\theta \in (0, 1)$ yield the relationships

$$\begin{aligned} F(x^k) - F(x^k + \tau d^k) - \frac{\theta\alpha_k\tau}{2}\|d^k\|^2 &\geq l_k(x^k) \\ &\quad - l_k(x^k + \tau d^k) - \frac{L_1}{2}\tau^2\|d^k\|^2 - \frac{\theta\alpha_k\tau}{2}\|d^k\|^2 \\ &\geq \tau\big(l_k(x^k) - l_k(x^k + d^k)\big) - \frac{L_1}{2}\tau^2\|d^k\|^2 - \frac{\theta\alpha_k\tau}{2}\|d^k\|^2 \\ &\geq (1-\theta)\tau\frac{\alpha_k}{2}\|d^k\|^2 - \frac{L_1}{2}\tau^2\|d^k\|^2 \\ &= \frac{\tau}{2}\|d^k\|^2\big((1-\theta)\alpha_k - L_1\tau\big). \end{aligned} \tag{21}$$

This tells us that the backtracking line search criterion (15) fulfills when $0 < \tau \leq \frac{(1-\theta)\alpha_k}{L_1}$, and thus the step size t_k satisfies the claimed condition (17). Substituting now $\tau := t_k \geq \frac{\gamma(1-\theta)\alpha_k}{L_1}$ into (21) and employing the estimate of $\|\mathcal{G}(x^k)\|^2$ from (20), we arrive at the inequalities

$$\begin{aligned} F(x^k) - F(x^k + t_k d^k) &\geq \frac{\theta\alpha_k t_k}{2}\|d^k\|^2 \\ &\geq \frac{\gamma\theta(1-\theta)\alpha_k^2}{2L_1}\left(\frac{1-\nu}{1+M+\alpha_k}\right)^2\|\mathcal{G}(x^k)\|^2, \end{aligned}$$

which verify the decreasing condition (18) and thus completes the proof of the lemma. □

Now we are ready to prove the global convergence of Algorithm 1. Define the sets

$$K := \big\{0, 1, \ldots\big\} \quad \text{and} \quad K_0 := \{0\} \cup \big\{k+1 \in K \;\big|\; \text{Step 5 is not applied at iteration } k\big\}. \tag{22}$$

Theorem 3.1 *Let $\{x^k\}$ be the sequence of iterates generated by Algorithm 1 with an arbitrarily chosen starting point $x^0 \in \mathbb{R}^n$ under the standing assumptions made. Then we have the residual condition*

$$\liminf_{k\to\infty} \|\mathcal{G}(x^k)\| = 0 \tag{23}$$

along the prox-gradient mapping (8). Furthermore, the boundedness of $\{x^k\}$ yields the convergence to the optimal value $\lim_{k\to\infty} F(x^k) = F^*$ and ensures that any limiting point of $\{x^k\}$ is a global minimizer in (1).

Proof First we consider the case where the set K_0 is infinite. We can reorganize K_0 in such a way that $0 = k_0 < k_1 < k_2 < \ldots$. It follows from Step 4 of Algorithm 1 that the estimate

$$\|\mathcal{G}(x^{k_{\ell+1}})\| \le \sigma \|\mathcal{G}(x^{k_\ell})\| \ \text{ whenever } \ \ell = 0, 1, \ldots$$

holds with the chosen number $\sigma \in (0, 1)$ in the algorithm, and therefore we get

$$0 \le \liminf_{k\to\infty} \|\mathcal{G}(x^k)\| \le \limsup_{\ell\to\infty} \|\mathcal{G}(x^{k_\ell})\| \le \lim_{\ell\to\infty} \sigma^\ell \|\mathcal{G}(x^{k_0})\| = 0,$$

which clearly yields (23). The continuity of $\mathcal{G}(\cdot)$ ensures that $\|\mathcal{G}(\bar{x})\| = 0$ for a limiting point $\bar{x}$ of the sequence $\{x^k\}_{k\in K_0}$, and thus $\bar{x} \in \mathcal{X}^*$. Consider now any limiting point $\bar{x}$ of the entire sequence of iterates $\{x^k\}_{k\in K}$. If there exists $\bar{k}$ such that $k \in K_0$ for all $k \ge \bar{k}$, it is easy to see that $\bar{x}$ is a global minimizer of (1). Otherwise, for any $k \notin K_0$, denote by $k_\ell \in K_0$ the largest number satisfying $k_\ell < k$, and hence we get the following estimate from Step 5:

$$F^* \le F(x^k) \le F(x^{k-1}) \le \ldots \le F(x^{k_\ell}).$$

When the sequence $\{x^k\}_{k\in K_0}$ is bounded, since any limiting point of $\{x^k\}_{k\in K_0}$ is a global minimizer of (1) as already shown, it follows that $\lim_{\ell\to\infty} F(x^{k_\ell}) = F^*$. This readily verifies by the constructions above that $\lim_{k\to\infty} F(x^k) = F^*$, and thus any limiting point of $\{x^k\}_{k\in K}$ provides a global minimum to (1).

Next we consider the case where K_0 is finite and denote $\bar{k} := \max_{k\in K_0} k$. It follows from Lemma 3.2 that for any $k > \bar{k}$ we get

$$F(x^{k+1}) - F(x^k) \le -\frac{\gamma\theta(1-\theta)}{2L_1}\left(\frac{(1-\nu)\alpha_k}{1+M+\alpha_k}\right)^2 \|\mathcal{G}(x^k)\|^2,$$

which therefore tells us that

$$\sum_{k=\bar{k}}^{\infty} \frac{\gamma\theta(1-\theta)}{2L_1}\left(\frac{(1-\nu)\alpha_k}{1+M+\alpha_k}\right)^2 \|\mathcal{G}(x^k)\|^2 \le F(x^{\bar{k}}) - F^* \le 0.$$

The latter implies in turn that

$$\lim_{k\to\infty} \frac{\gamma\theta(1-\theta)}{2L_1}\left(\frac{(1-\nu)\alpha_k}{1+M+\alpha_k}\right)^2 \|\mathcal{G}(x^k)\|^2 = 0.$$

Remembering the choice of $\alpha_k = \min\left\{\bar{\alpha}, c\|\mathcal{G}(x^k)\|^\rho\right\}$ with $\bar{\alpha}, c, \rho > 0$ ensures that

$$\lim_{k\to\infty} \|\mathcal{G}(x^k)\| = 0,$$

hence (23) holds. This readily verifies by the continuity of $\mathcal{G}(\cdot)$ that any limiting point of $\{x^k\}_{k\in K}$ provides a global minimum to (1). □

We conclude this section with a consequence of Theorem 3.1 giving an easily verifiable condition for the boundedness of the sequence of iterates in Algorithm 1. Recall that a function $\varphi\colon \mathbb{R}^n \to \bar{\mathbb{R}}$ is *coercive* if $\varphi(x) \to \infty$ provided that $\|x\| \to \infty$.

Corollary 3.1 *In addition to the standing assumptions imposed above, suppose that the cost function F in (1) is coercive. Then we have $\lim_{k\to\infty} F(x^k) = F^*$ for the sequence of iterates $\{x^k\}$ generated by Algorithm 1, and any limiting point of $\{x^k\}$ is a global minimizer of (1).*

Proof According to Steps 4 and 5 of Algorithm 1, the sequence $\{x^k\}$ generated by the algorithm satisfies the condition $F(x^k) \le C$ for all k. Then the coercivity of F implies that the sequence $\{x^k\}$ is bounded. Thus we deduce the conclusions of the corollary from Theorem 3.1. □

4 Fast local convergence under metric *q*-subregularity

This section is devoted to the local convergence of the proximal Newton-type Algorithm 1 under the imposed metric q-subregularity in both cases where $q \in (0, 1]$ and $q > 1$. In the first case, which is referred to as the Hölder metric subregularity, we do not consider any $q \in (0, 1]$, but precisely specify the lower bound of q and respectively modify some parameters of our algorithm. Namely, for the case where $q = 1$, i.e., the metric subregularity assumption of the subgradient mapping in (1) holds, the main result here establishes superlinear local convergence rates depending on the selected exponent $\rho \in (0, 1]$ in the algorithm, which gives us the *quadratic* convergence in the case where $\rho = 1$. Our analysis partly follows the scheme of [22] for a Newtonian algorithm to solve generalized equations with nonisolated solutions under certain Lipschitzian properties of perturbed solution sets. Note that the imposed metric subregularity allows us to avoid limitations of the line search procedure (needed for establishing the global convergence of our algorithm in Sect. 2 that is not addressed in [22]) to achieve now the fast local convergence. The imposed metric q-subregularity assumption is weaker for $q < 1$ than the metric subregularity, but allows us to achieve a local *superlinear* (while not quadratic) convergence of the algorithm. In the other case where $q > 1$, we achieve a *higher-than-quadratic* rate of the local convergence of the proposed algorithm.

Starting with the Hölder metric subregularity, we first provide the following norm estimate of directions d^k in the proposed Algorithm 1.

Lemma 4.1 *Let $\{x^k\}$ be the sequence generated by Algorithm 1, and let $\bar{x} \in \mathcal{X}^*$ be any limiting point of the sequence $\{x^k\}$. In addition to the standing assumptions, suppose that the subgradient mapping $\nabla f(x) + \partial g(x)$ is metrically q-subregular at $(\bar{x}, 0)$ for some $q \in (0, 1]$, that the Hessian $\nabla^2 f$ is locally Lipschitzian around $\bar{x}$, that $\alpha_k = \min\{\bar{\alpha}, c\|\mathcal{G}(x^k)\|^\rho\}$ with $\bar{\alpha}, c > 0$, $\rho \in (0, q]$, and $\varrho \ge \rho$ in Algorithm 1, and that the estimate $\|B_k - \nabla^2 f(x^k)\| \le C_1 \operatorname{dist}(x^k; \mathcal{X}^*)$ holds with some constant $C_1 > 0$. Then there exist positive numbers ϵ and c_1 such that for $d^k := \hat{x}^k - x^k$ we*

have the direction estimate

$$\|d^k\| \le c_1 \operatorname{dist}(x^k;\mathcal{X}^*) \ \ as \ \ x^k \in \mathbb{B}_\epsilon(\bar{x}). \tag{24}$$

Proof Remembering that $\hat{x}^k$ is an inexact solution to (3) satisfying conditions (14), we apply Lemma 3.1 and find a vector e^k such that the relationships in (16) hold. Denoting by π_x^k the (unique) projection of x^k onto the solution map $\mathcal{X}^*$, we get by basic convex analysis that $0 \in \nabla f(\pi_x^k) + \partial g(\pi_x^k)$ and thus

$$\nabla f(x^k) - \nabla f(\pi_x^k) + H_k(\pi_x^k - x^k) \in \nabla f(x^k) + H_k(\pi_x^k - x^k) + \partial g(\pi_x^k).$$

Since the mapping $\nabla f(x^k) + H_k(\cdot - x^k) + \partial g(\cdot)$ is strongly monotone on $\mathbb{R}^n$ with constant α_k, we have

$$\langle \nabla f(x^k) - \nabla f(\pi_x^k) + H_k(\pi_x^k - x^k) - e_k + H_k e_k, \pi_x^k - \hat{x}^k + e_k\rangle \ge \alpha_k \|\pi_x^k - \hat{x}^k + e_k\|^2.$$

Combining the above with the algorithm constructions gives us the estimates

$$\begin{aligned}
\|\pi_x^k - \hat{x}^k + e_k\| \le & \frac{1}{\alpha_k}\|\nabla f(x^k) - \nabla f(\pi_x^k) + H_k(\pi_x^k - x^k) - e_k + H_k e_k\| \\
\le & \frac{1}{\alpha_k}\Big(\|\nabla f(x^k) + \nabla^2 f(x^k)(\pi_x^k - x^k) - \nabla f(\pi_x^k)\| \\
& + \|(H_k - \nabla^2 f(x^k))(\pi_x^k - x^k)\| + \|e_k - H_k e_k\|\Big) \\
\le & \frac{1}{\alpha_k}\Big(\|\nabla f(x^k) + \nabla^2 f(x^k)(\pi_x^k - x^k) - \nabla f(\pi_x^k)\| \\
& + \|B_k - \nabla^2 f(x^k)\| \cdot \|x^k - \pi_x^k\| + \alpha_k\|x^k - \pi_x^k\| + (1+M)\|e_k\|\Big) \\
\le & \frac{1}{\alpha_k}\Big(\|\nabla f(x^k) + \nabla^2 f(x^k)(\pi_x^k - x^k) - \nabla f(\pi_x^k)\| + \|B_k \\
& - \nabla^2 f(x^k)\| \operatorname{dist}(x^k;\mathcal{X}^*) + \alpha_k \operatorname{dist}(x^k;\mathcal{X}^*) + (1+M)\nu\|\mathcal{G}(x^k)\|^{1+\varrho}\Big),
\end{aligned} \tag{25}$$

where the third inequality follows from the choice of $H_k = B_k + \alpha_k I$ while the fourth inequality is implied by $\|e_k\| \le \nu\|\mathcal{G}(x^k)\|^{1+\varrho}$. Since the Hessian mapping $\nabla^2 f$ is locally Lipschitzian around $\bar{x}$, there exist positive numbers ϵ_2 and L_2 such that for any $x, y \in \mathbb{B}_{\epsilon_2}(\bar{x})$ we get

$$\|\nabla f(x) + \nabla^2 f(x)(y - x) - \nabla f(y)\| \le \frac{L_2}{2}\|x - y\|^2.$$

Furthermore, the imposed assumption that $\|B_k - \nabla^2 f(x^k)\| \le C_1 \operatorname{dist}(x^k;\mathcal{X}^*)$ and the fact that $x^k \in \mathbb{B}_{\epsilon_2}(\bar{x})$ implying $\pi_x^k \in \mathbb{B}_{\epsilon_2}(\bar{x})$ give us the estimate

$$\|\pi_x^k - \hat{x}^k + e_k\| \le \frac{1}{\alpha_k}\Big(\Big(\frac{L_2}{2} + C_1\Big)\operatorname{dist}(x^k;\mathcal{X}^*)^2 + \alpha_k \operatorname{dist}(x^k;\mathcal{X}^*) + (1+M)\nu\|\mathcal{G}(x^k)\|^{1+\varrho}\Big)$$

provided $x^k \in \mathbb{B}_{\epsilon_2}(\bar{x})$. Next we employ the relationships

$$\begin{aligned}\|d^k\| = \|\hat{x}^k - x^k\| &\le \|\pi_x^k - \hat{x}^k + e_k\| \\ &\quad + \|\pi_x^k - x^k\| + \|e_k\| \ \text{ with } \ \|e_k\| \le \nu\|\mathcal{G}(x^k)\|^{1+\varrho}\end{aligned} \tag{26}$$

together with $\|\mathcal{G}(x^k)\| \le (2 + L_1)\operatorname{dist}(x^k; \mathcal{X}^*)$ by Proposition 2.3 to obtain that

$$\begin{aligned}\alpha_k\|d^k\| \le \left(\frac{L_2}{2} + C_1\right)\operatorname{dist}(x^k; \mathcal{X}^*)^2 + 2\alpha_k \operatorname{dist}(x^k; \mathcal{X}^*) \\ + (1 + M + \alpha_k)\nu\|\mathcal{G}(x^k)\|^\rho(2 + L_1)^{1+\varrho-\rho}\operatorname{dist}(x^k; \mathcal{X}^*)^{1+\varrho-\rho}\end{aligned} \tag{27}$$

provided that $x^k \in \mathbb{B}_{\epsilon_2}(\bar{x})$. The assumed metric q-subregularity of $\nabla f(x) + \partial g$ gives us by Proposition 2.4 numbers $\epsilon_1, \kappa_1 > 0$ with

$$\operatorname{dist}(x; \mathcal{X}^*) \le \kappa_1\|\mathcal{G}(x)\|^q \ \text{ for all } \ x \in \mathbb{B}_{\epsilon_1}(\bar{x}).$$

Supposing without loss of generality that $\epsilon_1 \le \min\{1, \epsilon_2\}$ and $\alpha_k = \min\left\{\bar{\alpha}, c\|\mathcal{G}(x^k)\|^\rho\right\}$ $= c\|\mathcal{G}(x^k)\|^\rho$ when $x^k \in \mathbb{B}_{\epsilon_1}(\bar{x})$ and remembering that $\rho \in (0, q]$ imply that

$$\alpha_k = c\|\mathcal{G}(x^k)\|^\rho \ge c\kappa_1^{-\frac{\rho}{q}}\operatorname{dist}(x; \mathcal{X}^*)^{\frac{\rho}{q}} \ge c\kappa_1^{-\frac{\rho}{q}}\operatorname{dist}(x; \mathcal{X}^*) \ \text{ as } \ x^k \in \mathbb{B}_{\epsilon_1}(\bar{x}). \tag{28}$$

Since $\rho \in (0, q]$ and $\varrho \ge \rho$, we deduce from (27) and (28) the existence of positive numbers ϵ and c_1 ensuring the fulfillment of estimate (24) claimed in the lemma. □

Next, under the imposed Hölder metric subregularity, we show that the set K_0 defined in (22) is infinite.

Lemma 4.2 *Let $\{x^k\}$ be the sequence generated by Algorithm 1, and let $\bar{x} \in \mathcal{X}^*$ be any limiting point of $\{x^k\}$. In addition to the standing assumptions, suppose that the mapping $\nabla f(x) + \partial g(x)$ is metrically q-subregular at $(\bar{x}, 0)$ for some $q \in (0, 1]$, that the Hessian $\nabla^2 f$ is locally Lipschitzian around $\bar{x}$, that $\alpha_k = \min\left\{\bar{\alpha}, c\|\mathcal{G}(x^k)\|^\rho\right\}$ with $\bar{\alpha}, c > 0$, $\rho \in (0, q]$, and $\varrho \ge \rho$ in Algorithm 1, and that the estimate $\|B_k - \nabla^2 f(x^k)\| \le C_1 \operatorname{dist}(x^k; \mathcal{X}^*)$ holds with some constant $C_1 > 0$. Then the set K_0 defined in (22) is infinite.*

Proof On the contrary, suppose that K_0 is finite and denote $\bar{k} := \max_{k \in K_0} k$. Arguing as in the proof of Lemma 3.1 tells us that

$$\lim_{k\to\infty} \|\mathcal{G}(x^k)\| = 0.$$

According to Lemma 4.1, there exist positive numbers ϵ and c_1 such that for $d^k := \hat{x}^k - x^k$ we have

$$\|d^k\| \le c_1 \operatorname{dist}(x^k; \mathcal{X}^*) \ \text{ as } \ x^k \in \mathbb{B}_\epsilon(\bar{x}),$$

which implies that $\liminf_{k\to\infty}\|d^k\|=0$. Combining this with the $(2+L_1)$-Lipschitz continuity of $\|\mathcal{G}(x)\|$ as follows from Proposition 2.1 gives us

$$\liminf_{k\to\infty}\|\mathcal{G}(\hat{x}^k)\|\le\lim_{k\to\infty}\|\mathcal{G}(x^k)\|+\liminf_{k\to\infty}(2+L_1)\|d^k\|=0.$$

Hence there exists $k>\bar{k}$ such that $k\in K_0$; a contradiction showing that the set K_0 is infinite. □

Having Lemmas 4.1, 4.2, and the previous estimates in hand, next we derive the following fast local convergence result for Algorithm 1 with a particular choice of parameters under the imposed Hölder metric subregularity of the subgradient mapping $\nabla f+\partial g$ with an appropriate factor q.

Theorem 4.1 *Let $\{x^k\}$ be the sequence generated by Algorithm 1, and let $\bar{x}\in\mathcal{X}^*$ be any limiting point of the sequence $\{x^k\}_{k\in K_0}$, where K_0 is defined in (22). In addition to the standing assumptions, suppose that the subgradient mapping $\nabla f(x)+\partial g(x)$ is metrically q-subregular at $(\bar{x},0)$ with $q\in(\frac{1}{2},1]$, that the Hessian mapping $\nabla^2 f$ is locally Lipschitzian around $\bar{x}$, that $\alpha_k=\min\{\bar{\alpha},c\|\mathcal{G}(x^k)\|^\rho\}$, $\bar{\alpha},c>0$, $\rho\in[\frac{1}{2},q]$, and $\varrho\ge\rho$ in Algorithm 1, and that $\|B_k-\nabla^2 f(x^k)\|=O(\|\mathcal{G}(x^k)\|)$. Then there exists a natural number k_0 such that*

$$t_k=1\ \text{ for all }\ k\ge k_0,$$

and the sequence $\{x^k\}$ converges to the point $\bar{x}$. Furthermore, this convergence is superlinear with the rate of $\rho+q>1$ in the sense that there exist a positive number C_0 and a natural number k_0 for which

$$\|\mathcal{G}(x^{k+1})\|\le C_0\|\mathcal{G}(x^k)\|^{\rho+q}\ \text{ whenever }\ k\ge k_0,\tag{29}$$

and $\operatorname{dist}(x^k;\mathcal{X}^*)$ converges R-superlinearly to 0 in the sense that $\lim_{k\to\infty}\sqrt[k]{\operatorname{dist}(x^k;\mathcal{X}^*)}=0$.

In particular, when the subgradient mapping $\nabla f(x)+\partial g(x)$ is metrically subregular at $(\bar{x},0)$, i.e., $q=1$, and when $\rho=1$, we have the quadratic convergence of $x^k\to\bar{x}$ with the exponent $\rho+q=2$ in (29), and there exists a positive constant $\tilde{C}_0$ such that

$$\operatorname{dist}(x^{k+1};\mathcal{X}^*)\le\tilde{C}_0\operatorname{dist}(x^k;\mathcal{X}^*)^2\ \text{ whenever }\ k\ge k_0.\tag{30}$$

Proof Observe first that the assumed metric q-subregularity of the mapping $\nabla f(x)+\partial g(x)$ at $(\bar{x},0)$ gives us positive numbers ϵ_1 and κ_1 such that for all p near $0\in\mathbb{R}^n$ we have the inclusion

$$\Sigma(p)\cap\mathbb{B}_{\epsilon_1}(\bar{x})\subset\mathcal{X}^*+\kappa_1\|p\|^q\mathbb{B}\ \text{ with }\ \Sigma(p):=\big\{x\in\mathbb{R}^n\ \big|\ p\in\nabla f(x)+\partial g(x)\big\}.\tag{31}$$

Employing Proposition 2.4 allows us to find $\kappa_2 > 1$ ensuring the estimate

$$\operatorname{dist}(x;\mathcal{X}^*) \le \kappa_2 \|\mathcal{G}(x)\|^q \ \text{ whenever } \ x \in \mathbb{B}_{\epsilon_1}(\bar{x}). \tag{32}$$

Since $\|B_k - \nabla^2 f(x^k)\| = O(\|\mathcal{G}(x^k)\|)$, we deduce from Proposition 2.3 the existence of $C_1 > 0$ with

$$\|B_k - \nabla^2 f(x^k)\| \le C_1 \operatorname{dist}(x^k;\mathcal{X}^*). \tag{33}$$

Recalling that $\hat{x}^k$ is an inexact solution of (3) satisfying (14) and using Lemma 3.1 give us

$$\begin{aligned} &e_k \in \nabla f(x^k) + H_k(\hat{x}^k - x^k) + \partial g(\hat{x}^k - e_k) \ \text{ with} \\ &\|e_k\| \le \nu \min\left\{\|\mathcal{G}(x^k)\|, \|\mathcal{G}(x^k)\|^{1+\varrho}\right\}. \end{aligned}$$

By setting $\tilde{x}^k := \hat{x}^k - e_k$, we have the inclusion

$$e_k - H_k e_k \in \nabla f(x^k) + H_k(\tilde{x}^k - x^k) + \partial g(\tilde{x}^k),$$

which implies therefore that

$$\mathcal{R}_k(\tilde{x}^k, x^k) := \nabla f(\tilde{x}^k) - \nabla f(x^k) - H_k(\tilde{x}^k - x^k) + e_k - H_k e_k \in \nabla f(\tilde{x}^k) + \partial g(\tilde{x}^k). \tag{34}$$

The latter reads, by the above definition of the perturbed solution map $\Sigma(p)$, that $\tilde{x}^k \in \Sigma(\mathcal{R}_k(\tilde{x}^k, x^k))$. Since $\nabla^2 f$ is locally Lipschitzian around $\bar{x}$, there exist numbers $L_2, \epsilon_2 > 0$ such that

$$\|\nabla f(x) + \nabla^2 f(x)(y - x) - \nabla f(y)\| \le \frac{L_2}{2}\|x - y\|^2 \ \text{ for any } \ x, y \in \mathbb{B}_{\epsilon_2}(\bar{x}). \tag{35}$$

Then choose by Lemma 4.1 a small number $0 < \varepsilon_1 < \min\{1, \epsilon_2\}$ such that

$$\|d^k\| \le c_1 \operatorname{dist}(x^k;\mathcal{X}^*) \ \text{ for all } \ x^k \in \mathbb{B}_{\epsilon_1}(\bar{x}) \tag{36}$$

with some $c_1 > 0$. Since $\|\tilde{x}^k - \bar{x}\| \le \|x^k - \bar{x}\| + \|d^k\| + \|e_k\|$ with $\|d^k\| \to 0$ and $\|e_k\| \to 0$ when $x^k \to \bar{x}$ as $k \to \infty$, we find $0 < \epsilon_3 \le \epsilon_1$ such that $\tilde{x}^k \in \mathbb{B}_{\epsilon_1}(\bar{x})$ whenever $x^k \in \mathbb{B}_{\epsilon_3}(\bar{x})$. We also assume that ϵ_3 is sufficiently small with $\|\mathcal{G}(x)\| < 1$ for all $x \in \mathbb{B}_{\epsilon_3}(\bar{x})$. This leads us to the relationships

$$\begin{aligned}\|\mathcal{R}_k(\tilde{x}^k, x^k)\| &= \|\nabla f(\tilde{x}^k) - \nabla f(x^k) - H_k(\tilde{x}^k - x^k) + e_k - H_k e_k\| \\ &= \|\nabla f(\tilde{x}^k) - \nabla f(x^k) - (B_k + \alpha_k I)(\tilde{x}^k - x^k) + e_k - H_k e_k\| \\ &\le \|\nabla f(\tilde{x}^k) - \nabla f(x^k) - \nabla^2 f(x^k)(\tilde{x}^k - x^k)\| \\ &\quad + \|B_k - \nabla^2 f(x^k)\| \cdot \|\tilde{x}^k - x^k\| + \alpha_k \|\tilde{x}^k - x^k\| \\ &\quad + (1+M)\|e_k\| \le \frac{L_2}{2}\|\tilde{x}^k - x^k\|^2 + C_1 \operatorname{dist}(x^k; \mathcal{X}^*)\|\tilde{x}^k - x^k\| \\ &\quad + \alpha_k \|\tilde{x}^k - x^k\| + (1+M)\nu\|\mathcal{G}(x^k)\|^{1+\varrho}\end{aligned} \tag{37}$$

if $x^k \in \mathbb{B}_{\epsilon_3}(\bar{x})$, where the second inequality follows from (35), $\|B_k - \nabla^2 f(x^k)\| \le C_1 \operatorname{dist}(x^k; \mathcal{X}^*)$, and $\|e_k\| \le \nu\|\mathcal{G}(x^k)\|^{1+\varrho}$. We have by Proposition 2.3 that $\|e_k\| \le \nu\|\mathcal{G}(x^k)\| \le \nu(2 + L_1)\operatorname{dist}(x^k; \mathcal{X}^*)$ for this choice of x^k. Thus it follows that

$$\begin{aligned}\|\tilde{x}^k - x^k\| \le \|\hat{x}^k - x^k\| + \|e^k\| = \|d^k\| \\ + \|e_k\| \le \|d^k\| + \nu(2 + L_1)\operatorname{dist}(x^k; \mathcal{X}^*),\end{aligned}$$

which being combined with (37), Proposition 2.3, and $\varrho \ge \rho$ gives us $c_2 > 0$ such that

$$\begin{aligned}\|\mathcal{R}_k(\tilde{x}^k, x^k)\| \le c_2\|d^k\|^2 + c_2 \operatorname{dist}(x^k; \mathcal{X}^*)^2 + c_2\alpha_k \operatorname{dist}(x^k; \mathcal{X}^*) \\ + \alpha_k\|d^k\| \ \text{ for all } \ x^k \in \mathbb{B}_{\epsilon_3}(\bar{x}).\end{aligned}$$

Then the direction estimate (36) together with (32) and the one of $\alpha_k = \min\left\{\bar{\alpha}, c\|\mathcal{G}(x^k)\|^\rho\right\}$ ensures the existence of a positive constant c_3 ensuring the estimates

$$\begin{aligned}\|\mathcal{R}_k(\tilde{x}^k, x^k)\| &\le c_2(c_1^2 + 1)\operatorname{dist}(x^k; \mathcal{X}^*)^2 + \alpha_k\,(c_1 + c_2)\operatorname{dist}(x^k; \mathcal{X}^*) \\ &\le c_2(c_1^2 + 1)\kappa_2^2\|\mathcal{G}(x^k)\|^{2q} + \alpha_k\,(c_1 + c_2)\,\kappa_2\|\mathcal{G}(x^k)\|^q \\ &\le c_3 \max\left\{\|\mathcal{G}(x^k)\|^{2q}, \|\mathcal{G}(x^k)\|^{\rho+q}\right\} \\ &\le c_3\|\mathcal{G}(x^k)\|^{\rho+q} \ \text{ for all } \ x^k \in \mathbb{B}_{\epsilon_3}(\bar{x}),\end{aligned} \tag{38}$$

where the last inequality follows from $\|\mathcal{G}(x^k)\| \le 1$ and $\rho \le q$. Recalling that $\mathcal{R}_k(\tilde{x}^k, x^k) \in \nabla f(x) + \partial g(x)$ and using $\|\mathcal{G}(x)\| \le \operatorname{dist}(0; \nabla f(x) + \partial g(x))$, which comes from Proposition 2.1, yield

$$\|\mathcal{G}(\tilde{x}^k)\| \le \|\mathcal{R}_k(\tilde{x}^k, x^k)\| \le c_3\|\mathcal{G}(x^k)\|^{\rho+q} \ \text{ for all } \ x^k \in \mathbb{B}_{\epsilon_3}(\bar{x}).$$

Combining the latter with the $(2 + L_1)$-Lipschitz continuity of $\|\mathcal{G}(x)\|$, which comes from Proposition 2.1, and with $\|e_k\| \le \nu\|\mathcal{G}(x^k)\|^{1+\varrho}$ as $\varrho \ge \rho$, gives us

$$\|\mathcal{G}(\hat{x}^k)\| \le (c_3 + (2 + L_1)\nu)\|\mathcal{G}(x^k)\|^{\rho+q}, \tag{39}$$

and thus we arrive at the estimate

$$\|\mathcal{G}(\hat{x}^k)\| \le (c_3 + (2 + L_1)\nu)\|\mathcal{G}(x^k)\|^{\rho+q-1} \cdot \|\mathcal{G}(x^k)\|$$

provided $x^k \in \mathbb{B}_{\epsilon_3}(\bar{x})$. Since $\rho + q - 1 > 0$ and $\|\mathcal{G}(x^k)\| \le (2 + L_1)\,\mathrm{dist}(x^k; \mathcal{X}^*)$, which comes from Proposition 2.3, this allows us to find $0 < \epsilon_0 < \epsilon_3$ such that

$$\|\mathcal{G}(\hat{x}^k)\| \le \sigma\,\|\mathcal{G}(x^k)\| \;\text{ for }\; x^k \in \mathbb{B}_{\epsilon_0}(\bar{x}). \tag{40}$$

Remembering that $C > F(x^0) \ge F_*$ and that F is continuous on the open domain $\mathrm{dom}\, F$, we select a positive number ϵ_0 to be so small that

$$\sup_{x\in\mathbb{B}_{\epsilon_0}(\bar{x})} F(x) \le C. \tag{41}$$

Let us next we introduce the positive constants

$$\tilde{\sigma} := \sigma^q < 1 \;\text{ and }\; \tilde{\epsilon} := \min\left\{\left(\frac{1-\tilde{\sigma}}{2c_1\kappa_2(2+L_1)^q}\epsilon_0\right)^{\frac{1}{q}},\ \frac{1}{2}\epsilon_0,\ \frac{1}{1+c_1}\epsilon_0\right\}$$

and show that if $x^{k_0} \in \mathbb{B}_{\tilde{\epsilon}}(\bar{x})$ with some $k_0 \in K_0$, then for any $k \ge k_0$ we have

$$k+1 \in K_0,\ t_k = 1,\ x^{k+1} = \hat{x}^k,\ \text{ and }\ x^{k+1} \in \mathbb{B}_{\epsilon_0}(\bar{x}). \tag{42}$$

To verify (42), set first $k := k_0$ and deduce from $x^k \in \mathbb{B}_{\tilde{\epsilon}}(\bar{x})$ that

$$\|\hat{x}^k - \bar{x}\| \le \|x^k - \bar{x}\| + \|d_k\| \le \|x^k - \bar{x}\| + c_1\,\mathrm{dist}(x^k; \mathcal{X}^*) \le (1 + c_1)\|x^k - \bar{x}\| \le \epsilon_0,$$

where the second inequality comes from (36). It follows from (40) and $k_0 \in K_0$ that

$$\|\mathcal{G}(\hat{x}^k)\| \le \sigma\,\|\mathcal{G}(x^k)\| = \sigma\vartheta_k.$$

Observe also that (41) obviously yields $F(\hat{x}^k) \le C$. Then by Step 4 of Algorithm 1 we get $k+1 \in K_0$, $t_k = 1$, $x^{k+1} = \hat{x}^k$, $\vartheta_{k+1} = \|\mathcal{G}(x^{k+1})\|$, and $x^{k+1} \in \mathbb{B}_{\epsilon_0}(\bar{x})$. To justify further (42) for any $k > k_0$, proceed by induction and suppose that for all $k - 1 \ge \ell \ge k_0$ we have

$$\begin{aligned}&\ell+1 \in K_0,\ t_\ell = 1,\ x^{\ell+1} = \hat{x}^\ell,\ x^{\ell+1} \in \mathbb{B}_{\epsilon_0}(\bar{x}),\ \text{and hence}\\ &\vartheta_\ell = \|\mathcal{G}(x^\ell)\|,\ \|\mathcal{G}(x^{\ell+1})\| \le \sigma\,\|\mathcal{G}(x^\ell)\|.\end{aligned}$$

This readily implies the estimates

$$\begin{aligned}\|\hat{x}^k - x^{k_0}\| &\le \sum_{\ell=k_0}^{k} \|d^\ell\| \le \sum_{\ell=k_0}^{k} c_1\kappa_2 \|\mathcal{G}(x^\ell)\|^q \le \sum_{\ell=k_0}^{k} c_1\kappa_2\tilde{\sigma}^{\ell-k_0} \|\mathcal{G}(x^{k_0})\|^q \\ &\le \frac{c_1\kappa_2}{1-\tilde{\sigma}} \|\mathcal{G}(x^{k_0})\|^q \le \frac{c_1\kappa_2(2+L_1)^q}{1-\tilde{\sigma}}\|x^{k_0} - \bar{x}\|^q,\end{aligned} \tag{43}$$

where the second inequality follows from (32) and (36), while the last inequality is a consequence of Proposition 2.3. Therefore, it gives us the conditions

$$\|\hat{x}^k - \bar{x}\| \le \|\hat{x}^k - x^{k_0}\| + \|x^{k_0} - \bar{x}\| \le \frac{c_1\kappa_2(2+L_1)^q}{1-\tilde{\sigma}}\tilde{\epsilon}^q + \tilde{\epsilon} \le \epsilon_0.$$

Arguing as above, we get that $\|\mathcal{G}(\hat{x}^k)\| \le \sigma\vartheta_k$ and $F(\hat{x}^k) \le C$, which ensures that (42) holds for any $k \ge k_0$ and thus verifies these conditions in the general case.

Now we prove the claimed convergence $x^k \to \bar{x}$ as $k \to \infty$ with the convergence rate (29), where $\bar{x}$ is the designated limiting point $\bar{x}$ of the sequence $\{x^k\}_{k\in K_0}$. As shown above, for any $k \ge k_0$ we have $k+1 \in K_0$, $t_k = 1$, $x^{k+1} = \hat{x}^k$, and $x^{k+1} \in \mathbb{B}_{\epsilon_0}(\bar{x})$. Using the conditions in (42) and the arguments similarly to to the proof of (43), we are able to show that

$$\|x^k - \bar{x}\| \le \frac{c_1\kappa_2(2+L_1)^q}{1-\tilde{\sigma}}\|x^{\tilde{k}} - \bar{x}\|^q + \|x^{\tilde{k}} - \bar{x}\| \quad \text{for any} \quad k \ge \tilde{k}, \tag{44}$$

whenever $\tilde{k} \ge k_0$. This shows that the sequence $\{x^k\}$ is bounded. Picking any limiting point $\tilde{x}$ of $\{x^k\}$ and passing to the limit as $k \to \infty$ in (44) lead us to the estimate

$$\|\tilde{x} - \bar{x}\| \le \frac{c_1}{1-\tilde{\sigma}}\|x^{\tilde{k}} - \bar{x}\| + \|x^{\tilde{k}} - \bar{x}\|.$$

Recalling that $\bar{x}$ is a limiting point of $\{x^k\}_{k\in K_0}$, we pass to the limit as $\tilde{k} \to \infty$ in the estimate above and get $\|\tilde{x} - \bar{x}\| = 0$, which implies that $\{x^k\}$ converges to $\bar{x}$. Finally, employing (39) gives us numbers $C_0, k_0 > 0$ such that the claimed condition (29) holds. Then $\|\mathcal{G}(x^{k+1})\|^q \le C_0^q(\|\mathcal{G}(x^k)\|^q)^{\rho+q}$ for any $k \ge k_0$, which means that $\|\mathcal{G}(x^k)\|^q$ converges Q-superlinearly to 0. Combined the latter with (32) implies that $\operatorname{dist}(x^k; \mathcal{X}^*)$ converges R-superlinearly to 0. When the subgradient mapping $\nabla f(x) + \partial g(x)$ is metrically subregular at $(\bar{x}, 0)$, we have the following condition by setting $q = \rho = 1$ in (29):

$$\|\mathcal{G}(x^{k+1})\| \le C_0\|\mathcal{G}(x^k)\|^2 \quad \text{whenever} \quad k \ge k_0.$$

Employing (32) and $\|\mathcal{G}(x^k)\| \le (2+L_1)\operatorname{dist}(x^k; \mathcal{X}^*)$, which comes from Proposition 2.3, gives us a positive number $\tilde{C}_0$ such that the claimed condition (30) holds. This completes the proof of the theorem. □

The concluding result of this section concerns the other kind of metric q-subregularity of the subgradient mapping in (1) in the case where $q > 1$. As discussed in Sect. 2, this type of higher-order metric subregularity is rather new in the literature, and it has never been used in applications to numerical optimization. The following theorem shows that the higher-order subregularity assumption imposed on the subgradient mapping ∂F at the point in question allows us to derive a counterpart of Theorem 4.1 with establishing the *convergence rate*, which may be *higher than quadratic*. Indeed, Proposition 2.5 characterizes the metric q-subregularity of the subgradient mapping by an equivalent $\frac{1+q}{q}$ growth condition for each $q > 0$. Based on this equivalence for $q > 1$, the imposed metric q-subregularity implies a sharper growth behavior of F around the solution point in comparison with the quadratic growth. Consequently, the convergence rate faster than the quadratic rate can be achieved.

Theorem 4.2 *Let $\{x^k\}$ be the sequence generated by Algorithm 1 with $\alpha_k = \min\{\bar{\alpha}, c\|\mathcal{G}(x^k)\|^\rho\}$ as $\rho \in (0, 1]$, and let $\bar{x} \in \mathcal{X}^*$ be any limiting point of $\{x^k\}_{k\in K_0}$, where K_0 is taken from (22). In addition to the standing assumptions, suppose that the mapping $\nabla f(x) + \partial g(x)$ is metrically q-subregular at $(\bar{x}, 0)$ with $q > 1$, that the Hessian $\nabla^2 f$ is locally Lipschitzian around $\bar{x}$, that $\|B_k - \nabla^2 f(x^k)\| = O(\|\mathcal{G}(x^k)\|)$, and that $\varrho \geq q(1+\rho) - 1$ in (14). Then there exists k_0 such that $t_k = 1$ for all $k \geq k_0$ and that $\{x^k\}$ converges to the point $\bar{x}$ with the convergence rate $q(1+\rho)$. The latter means that for some $k_0, C_0 > 0$ we have*

$$\operatorname{dist}(x^{k+1}; \mathcal{X}^*) \leq C_0 \operatorname{dist}(x^k; \mathcal{X}^*)^{q(1+\rho)} \quad \text{whenever} \quad k \geq k_0. \tag{45}$$

Proof It follows from the imposed metric q-subregularity condition with a fixed degree $q > 1$ that

$$\Sigma(p) \cap \mathbb{B}_{\epsilon_1}(\bar{x}) \subset \mathcal{X}^* + \kappa_1 \|p\|^q \mathbb{B} \quad \text{for some} \quad \epsilon_1, \kappa_1 > 0 \tag{46}$$

whenever $p \in \mathbb{R}^n$ is sufficiently close to the origin. Following the proof of Theorem 4.1, we arrive at the estimate of $\|\mathcal{R}_k(\tilde{x}^k, x^k)\|$ in (38) with some constant $c_3 > 0$, where $\tilde{x}^k := \hat{x}^k - e_k$ while $\mathcal{R}_k(\tilde{x}^k, x^k)$, $\hat{x}^k$, and e_k are defined and analyzed similarly to the case of Theorem 4.1. Then there exists $\epsilon_3 \in (0, 1]$ such that $\tilde{x}^k \in \mathbb{B}_{\epsilon_1}(\bar{x})$ when $x^k \in \mathbb{B}_{\epsilon_3}(\bar{x})$. Since $\tilde{x}^k \in \Sigma(\mathcal{R}_k(\tilde{x}^k\, x^k))$, we combine this with (46) and get the estimates

$$\begin{aligned}
&\operatorname{dist}(\tilde{x}^k; \mathcal{X}^*) \leq \kappa_1 \|\mathcal{R}_k(\tilde{x}^k, x^k)\|^q \leq \kappa_1 c_3^q \|\mathcal{G}(x^k)\|^{q(1+\rho)} \\
&\quad \leq \kappa_1 c_3^q (2 + L_1)^q \operatorname{dist}(x^k; \mathcal{X}^*)^{q(1+\rho)} \quad \text{and} \\
&\operatorname{dist}(\hat{x}^k; \mathcal{X}^*) \leq \operatorname{dist}(\tilde{x}^k; \mathcal{X}^*) + \|e_k\| \\
&\quad \leq \kappa_1 c_3^q (2 + L_1)^q \operatorname{dist}(x^k; \mathcal{X}^*)^{q(1+\rho)} + \nu \|\mathcal{G}(x^k)\|^{1+\varrho} \\
&\quad \leq (\kappa_1 c_3^q (2 + L_1)^q + \nu (2 + L_1)^{1+\varrho}) \operatorname{dist}(x^k; \mathcal{X}^*)^{q(1+\rho)} \quad \text{whenever} \quad x^k \in \mathbb{B}_{\epsilon_3}(\bar{x}).
\end{aligned} \tag{47}$$

Employing the induction arguments as in the proof of Theorem 4.1 yields the existence of a natural number k_0 such that we have $k+1 \in K_0, t_k = 1, x^{k+1} = \hat{x}^k, x^{k+1} \in \mathbb{B}_{\epsilon_3}(\bar{x})$

when $k \geq k_0$, and that the sequence $\{x^k\}$ converges to $\bar{x}$ as $k \to \infty$. Hence the second estimate in (47) gives a positive number C_0 and a natural number k_0, which ensure the fulfillment the claimed convergence rate (45) and thus complete the proof. □

5 Superlinear local convergence with non-Lipschitzian Hessians

As seen in Sect. 4, the imposed local Lipschitz continuity of the Hessian mapping $\nabla^2 f$ plays a crucial role in the justifications of the local convergence results obtained therein. In this section we show that the latter assumption can be dropped with preserving a local superlinear convergence of Algorithm 1 for a rather broad class of loss functions f that naturally appear in many practical models arising in machine learning and statistics, which includes, e.g., linear regression, logistic regression, and Poisson regression.

The class of loss functions f of our consideration in this section satisfies the following structural properties.

Assumption 5.1 The loss function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ of (1) is represented in the form

$$f(x) := h(Ax) + \langle b, x \rangle, \tag{48}$$

where A is an $m \times n$ matrix, $b \in \mathbb{R}^n$, and $h : \mathbb{R}^m \to \bar{\mathbb{R}}$ is a proper, convex, and l.s.c. function such that:

(1) h is strictly convex on any compact and convex subset of the domain $\operatorname{dom} h$.
(2) h is continuously differentiable on the set $\operatorname{dom} h$, which is assumed to be open, and the gradient mapping ∇h is Lipschitz continuous on any compact subset $\Omega \subset \operatorname{dom} h$.

Due to the strict convexity of h, the linear mapping $x \to Ax$ in (48) is invariant over the solution set $\mathcal{X}^*$ to (1). This is the contents of the following result taken from [5, Lemma 2.1].

Lemma 5.1 *Under the fulfillment of Assumption 5.1 there exists $\bar{y} \in \mathbb{R}^m$ such that $Ax = \bar{y}$ for all $x \in \mathcal{X}^*$.*

The next lemma is a counterpart of Lemma 4.1 without imposition the local Lipschitz continuity of the Hessian $\nabla^2 f$. By furnishing a similar while somewhat different scheme in comparison with Lemma 4.1, we establish new direction estimates of Algorithm 1 used in what follows. Note that we do not exploit in the lemma the structural conditions on f listed in Assumption 5.1.

Lemma 5.2 *Let $\{x^k\}$ be the sequence generated by Algorithm 1 with $\alpha_k = \min\{\bar{\alpha}, c\|\mathcal{G}(x^k)\|^\rho\}$ and $\rho \in (0, 1)$, and let $\bar{x} \in \mathcal{X}^*$ be any limiting point of $\{x^k\}$. In addition to Assumption 1.1 and (10), suppose that the Hessian mapping $\nabla^2 f$ is continuous at $\bar{x} \in \mathcal{X}^*$, that $\|B_k - \nabla^2 f(x^k)\| \to 0$ as $k \to \infty$, and that the subgradient mapping $\nabla f(x) + \partial g(x)$ is metrically subregular at $(\bar{x}, 0)$. Then given an arbitrary quantity $\delta > 0$, there exist $\epsilon > 0$ and $k_0 \in \mathbb{N}$ such that for $d^k := \hat{x}^k - x^k$ we have the estimates*

$$\alpha_k\|d^k\| \le \delta\,\mathrm{dist}(x^k;\mathcal{X}^*) \;\text{ and }\; \|d^k\| \le \delta\,\mathrm{dist}(x^k;\mathcal{X}^*)^{1-\rho} \;\text{ when }\; x^k \in \mathbb{B}_\epsilon(\bar{x}) \;\text{ and }\; k > k_0. \tag{49}$$

Proof Since $\hat{x}^k$ is an inexact solution to subproblem (3) satisfying (14), we get by Lemma 3.1 that there exists e_k for which both conditions in (16) hold. Taking the projection π_x^k of x^k onto the solution set $\mathcal{X}^*$ and arguing as in the proof of Lemma 4.1 bring us to the inequality in (25), which together with the direction estimate in (26) ensures that

$$\begin{aligned}\alpha_k\|d^k\| \le \Big(&\|\nabla f(x^k) + \nabla^2 f(x^k)(\pi_x^k - x^k) - \nabla f(\pi_x^k)\| + \|B_k - \nabla^2 f(x^k)\|\mathrm{dist}(x^k;\mathcal{X}^*)\\ &+ 2\alpha_k\,\mathrm{dist}(x^k;\mathcal{X}^*) + (1 + M + \alpha_k)\nu\|\mathcal{G}(x^k)\|^{1+\varrho}\Big).\end{aligned} \tag{50}$$

It follows from the mean value theorem and the choice of π_x^k as the projection of x^k onto $\mathcal{X}^*$ that

$$\begin{aligned}&\|\nabla f(x^k) + \nabla^2 f(x^k)(\pi_x^k - x^k) - \nabla f(\pi_x^k)\| = \|(\nabla^2 f(x^k) - \nabla^2 f(\xi^k))(\pi_x^k - x^k)\|\\ &\le \|\nabla^2 f(x^k) - \nabla^2 f(\xi^k)\|\mathrm{dist}(x^k;\mathcal{X}^*),\end{aligned}$$

where $\xi^k := \lambda^k x^k + (1-\lambda^k)\pi_x^k$ for some $\lambda^k \in (0,1)$, and hence $\xi^k \to \bar{x}$ when $x^k \to \bar{x}$ as $k \to \infty$. Then passing to the limit as $k \to \infty$ and using the assumed continuity of $\nabla^2 f$ at $\bar{x}$ show that $\|\nabla^2 f(x^k) - \nabla^2 f(\xi^k)\| \to 0$. Since $\alpha_k = \min\big\{\bar{\alpha}, c\|\mathcal{G}(x^k)\|^\rho\big\} \to 0$ and $\|\mathcal{G}(x^k)\| \le (2 + L_1)\mathrm{dist}(x^k;\mathcal{X}^*)$ by Proposition 2.3, and since $\|B_k - \nabla^2 f(x^k)\| \to 0$ as $k \to \infty$, for any $\delta > 0$ we find $\epsilon > 0$ and $k_0 \in \mathbb{N}$ such that

$$\alpha_k\|d^k\| \le \delta\,\mathrm{dist}(x^k;\mathcal{X}^*) \;\text{ when }\; x^k \in \mathbb{B}_\epsilon(\bar{x}) \;\text{ and }\; k > k_0,$$

which justifies the first estimate in (49). To verify finally the second one in (49), employ Proposition 2.2 and the above expression of α_k to find positive numbers ϵ_1 and c_1 ensuring the inequality

$$\alpha_k \ge c_1\,\mathrm{dist}(x^k;\mathcal{X}^*)^\rho \;\text{ for all }\; x \in \mathbb{B}_{\epsilon_1}(\bar{x}).$$

Combining the latter with the first estimate in (49) tells us that for the fixed number $\delta > 0$ there exist $\varepsilon > 0$ and $k > k_0$ such that the second estimate in (49) is also satisfied, and thus the proof is complete. □

By the same arguments as in the proof of Lemma 4.2, we can also show that the set K_0 defined in (22) is infinite in the setting of Lemma 5.2. Now we are ready to derive the promised result showing that the sequence of iterates, which are generated by Algorithm 1 for the structured problem (1) considered in this section, converges superlinearly to a given optimal solution $\bar{x} \in \mathcal{X}^*$ without the local Lipschitz continuity assumption on the Hessian mapping $\nabla^2 f$.

Theorem 5.1 *Let $\{x^k\}$ be the sequence of iterates generated by Algorithm 1 with $\alpha_k = \min\{\bar{\alpha}, c\|\mathcal{G}(x^k)\|^\rho\}$ and $\rho \in (0,1)$, and let $\bar{x} \in \mathcal{X}^*$ be any limiting point of the sequence $\{x^k\}_{k\in K_0}$ with the set K_0 defined in (22). Suppose in addition to the assumptions of Lemma 5.2 that the loss function f is given in the structured form (48) under the fulfillment of Assumption 5.1, and that at each iteration step k the matrix B_k is represented in the form $B_k = A^T D_k A$, where A is taken from (48) while $D_k \in \mathbb{R}^{m\times m}$ is some positive semidefinite matrix. Then there exists a natural number k_0 such that $t_k = 1$ for all $k \geq k_0$, and that the sequence $\{x^k\}$ converges to $\bar{x}$ with the superlinear convergence rate, i.e., there is k_1 for which we have*

$$\operatorname{dist}(x^{k+1};\mathcal{X}^*) = o\big(\operatorname{dist}(x^k;\mathcal{X}^*)\big) \ \text{ whenever } \ k \geq k_1. \tag{51}$$

Proof Proceeding similarly to the proof of Theorem 4.1, at each iteration step k we have the vector $\mathcal{R}_k(\tilde{x}^k, x^k)$ defined in (34) with $\tilde{x}^k := \hat{x}^k - e_k$, where $\hat{x}^k$ is an inexact solution of (3) satisfying (14), and where e_k is taken from (16). These relationships and the mean value theorem applied to the gradient mapping ∇f on $[x^k, \tilde{x}^k]$ give us a vector $\tilde{\xi}^k := \tilde{\lambda}^k x^k + (1-\tilde{\lambda}^k)\tilde{x}^k$ with some $\tilde{\lambda}^k \in (0,1)$ such that

$$\begin{aligned}
\|\mathcal{R}_k(\tilde{x}^k, x^k)\| &= \|\nabla f(\tilde{x}^k) - \nabla f(x^k) - H_k(\tilde{x}^k - x^k) + e_k - H_k e_k\| \\
&= \|\nabla f(\tilde{x}^k) - \nabla f(x^k) - (B_k + \alpha_k I)(\tilde{x}^k - x^k) + e_k - H_k e_k\| \\
&\leq \|\nabla f(\tilde{x}^k) - \nabla f(x^k) - \nabla^2 f(x^k)(\tilde{x}^k - x^k)\| + \|B_k - \nabla^2 f(x^k)\| \cdot \|\tilde{x}^k - x^k\| \\
&\quad + \alpha_k\|\tilde{x}^k - x^k\| + (1+M)\|e_k\| \\
&\leq \|(\nabla^2 f(\tilde{\xi}^k) - \nabla^2 f(x^k))(\tilde{x}^k - x^k)\| + \|(B_k - \nabla^2 f(x^k))(\tilde{x}^k - x^k)\| \\
&\quad + \alpha_k\|d^k\| + (1+M)\nu\|\mathcal{G}(x^k)\|^{1+\varrho}.
\end{aligned}$$

Let $\tilde{\pi}_x^k$ and π_x^k be the projections of $\tilde{x}^k$ and x^k onto $\mathcal{X}^*$, respectively. Then it follows from Lemma 5.1 that $A\tilde{\pi}_x^k = A\pi_x^k$. By Assumption 5.1 we have $\nabla^2 f(x) = A^T\nabla^2 h(x)A$, and thus

$$\big(\nabla^2 f(\tilde{\xi}^k) - \nabla^2 f(x^k)\big)(\tilde{x}^k - x^k) = \big(\nabla^2 f(\tilde{\xi}^k) - \nabla^2 f(x^k)\big)(\tilde{x}^k - \tilde{\pi}_x^k + \pi_x^k - x^k).$$

Using the assumed representation $B_k = A^T D_k A$ of the matrix B_k, we similarly get that

$$\big(B_k - \nabla^2 f(x^k)\big)(\tilde{x}^k - x^k) = \big(B_k - \nabla^2 f(x^k)\big)(\tilde{x}^k - \tilde{\pi}_x^k + \pi_x^k - x^k).$$

Plugging the obtained expressions into the above estimate of $\|\mathcal{R}_k\|$ gives us

$$\begin{aligned}
\|\mathcal{R}_k(\tilde{x}^k, x^k)\| &\leq \|(\nabla^2 f(\tilde{\xi}^k) - \nabla^2 f(x^k))(\tilde{x}^k - \tilde{\pi}_x^k + \pi_x^k - x^k)\| \\
&\quad + \|(B_k - \nabla^2 f(x^k))(\tilde{x}^k - \tilde{\pi}_x^k + \pi_x^k - x^k)\| \\
&\quad + \alpha_k\|d^k\| + (1+M)\nu\|\mathcal{G}(x^k)\|^{1+\varrho} \\
&\leq \|\nabla^2 f(\tilde{\xi}^k) - \nabla^2 f(x^k)\|\big(\operatorname{dist}(\tilde{x}^k;\mathcal{X}^*) + \operatorname{dist}(x^k;\mathcal{X}^*)\big)
\end{aligned}$$

$$+\|B_k-\nabla^2 f(x^k)\|\big(\operatorname{dist}(\tilde{x}^k;\mathcal{X}^*)+\operatorname{dist}(x^k;\mathcal{X}^*)\big)$$
$$+\alpha_k\|d^k\|+(1+M)\nu(2+L_1)^{\varrho}\operatorname{dist}(x^k;\mathcal{X}^*)^{1+\varrho}.$$

It follows from the second estimate of Lemma 5.2 that $\|d^k\|\to 0$ as $k\to\infty$ and $x^k\to\bar{x}$. Since $x^k\to\bar{x}$ implies that $\tilde{x}^k\to\bar{x}$ as $k\to\infty$, the assumed continuity of $\nabla^2 f$ at $\bar{x}$ and the above construction of $\tilde{\xi}^k$ tell us that $\|\nabla^2 f(\tilde{\xi}^k)-\nabla^2 f(x^k)\|\to 0$ as $k\to\infty$ and $x^k\to\bar{x}$. Now the first estimate of Lemma 5.2 ensures that $\alpha_k\|d^k\|=o(\operatorname{dist}(x^k;\mathcal{X}^*))$ as $k\to\infty$ and $x^k\to\bar{x}$. Combining this with $\|B_k-\nabla^2 f(x^k)\|\to 0$ as $k\to\infty$ allows us to conclude that for any $\delta>0$ there exist a positive number ϵ and a natural number k_0 such that

$$\|\mathcal{R}_k(\tilde{x}^k,x^k)\|\le\delta\Big(\operatorname{dist}(\tilde{x}^k;\mathcal{X}^*)+\operatorname{dist}(x^k;\mathcal{X}^*)\Big)\ \text{ whenever }\ x^k\in\mathbb{B}_\epsilon(\bar{x})\ \text{ and }\ k>k_0. \tag{52}$$

It follows from the metric subregularity assumption that we have inclusion (31) with $q=1$ and the perturbed solution map $\Sigma(p)$ therein. Since $\tilde{x}^k\in\Sigma(\mathcal{R}_k(\tilde{x}^k,x^k))$ as shown above, there is $\kappa_1>0$ with

$$\begin{aligned}&\operatorname{dist}(\tilde{x}^k;\mathcal{X}^*)\\&\quad\le\kappa_1\|\mathcal{R}_k(\tilde{x}^k,x^k)\|\le\kappa_1\delta\Big(\operatorname{dist}(\tilde{x}^k;\mathcal{X}^*)+\operatorname{dist}(x^k;\mathcal{X}^*)\Big)\\&\quad\text{for all }\ x^k\in\mathbb{B}_\epsilon(\bar{x})\ \text{ and }\ k>k_0,\end{aligned}$$

which implies that $\operatorname{dist}(\tilde{x}^k;\mathcal{X}^*)=o(\operatorname{dist}(x^k;\mathcal{X}^*))$ as $k\to\infty$. Recalling the estimates

$$\begin{aligned}\operatorname{dist}(\hat{x}^k;\mathcal{X}^*)&\le\operatorname{dist}(\tilde{x}^k;\mathcal{X}^*)+\|e_k\|\ \text{ and }\ \|e_k\|\le\nu\|\mathcal{G}(x^k)\|^{1+\varrho}\\&\le\nu(2+L_1)^{1+\varrho}\operatorname{dist}(x^k;\mathcal{X}^*)^{1+\varrho}=o\big(\operatorname{dist}(x^k;\mathcal{X}^*)\big),\end{aligned}$$

we readily get, for all the numbers δ,ε,k_0,k taken from (52), the conditions

$$\operatorname{dist}(\hat{x}^k;\mathcal{X}^*)=o\big(\operatorname{dist}(x^k;\mathcal{X}^*)\big)\ \text{ and }\ \operatorname{dist}(\hat{x}^k;\mathcal{X}^*)\le\delta\operatorname{dist}(x^k;\mathcal{X}^*), \tag{53}$$

which ensure therefore the existence of a positive number ε_0 and a natural number k_0 such that

$$\operatorname{dist}(\hat{x}^k;\mathcal{X}^*)\le\frac{\sigma}{(2+L_1)\kappa_2}\operatorname{dist}(x^k;\mathcal{X}^*)\ \text{ whenever }\ x^k\in\mathbb{B}_{\epsilon_0}(\bar{x})\ \text{ and }\ k>k_0. \tag{54}$$

Employing Lemma 5.2, suppose without loss of generality that there exists $c_1>0$ with

$$\|d^k\|\le c_1\operatorname{dist}(x^k;\mathcal{X}^*)^{1-\rho}\ \text{ for all }\ x^k\in\mathbb{B}_{\epsilon_0}(\bar{x})\ \text{ and }\ k>k_0. \tag{55}$$

Since $C > F(x^0) \geq F^*$ in our algorithm, and since F is continuous on dom F, let us pick $\epsilon_0 > 0$ to be so small that condition (41) holds. Defining the positive numbers

$$\tilde{\sigma} := \frac{\sigma}{(2+L_1)\kappa_2} < 1 \ \text{ and } \ \tilde{\epsilon} := \min\left\{\frac{\epsilon_0}{2}, \Big(\frac{1-\tilde{\sigma}^{1-\rho}}{2c_1}\epsilon_0\Big)^{\frac{1}{1-\rho}}\right\} \tag{56}$$

and invoking the set K_0 from (22), we intend to show that if $x^{k_1} \in \mathbb{B}_{\tilde{\epsilon}}(\bar{x})$ with $k_1 > k_0$ and $k_1 \in K_0$, then

$$k+1 \in K_0, \ t_k = 1, \ x^{k+1} = \hat{x}^k, \ \text{ and } \ x^{k+1} \in \mathbb{B}_{\epsilon_0}(\bar{x}) \ \text{ whenever } \ k \geq k_1. \tag{57}$$

To prove it by induction, observe first that for $k := k_1$ all the conditions in (57) can be verified similarly to the proof of (42) in Theorem 4.1 with the replacement of k_0 by k_1. Considering now the general case where $k > k_1$ in (57), suppose that the latter holds for any $k-1 \geq \ell \geq k_1$, which clearly yields $\operatorname{dist}(x^{\ell+1}; \mathcal{X}^*) \leq \tilde{\sigma}\operatorname{dist}(x^\ell; \mathcal{X}^*)$. Then the above estimates and the parameter choice in (56) ensure that

$$\begin{aligned}
\|\hat{x}^k - x^{k_1}\| &\leq \sum_{\ell=k_1}^{k} \|d^\ell\| \leq \sum_{\ell=k_1}^{k} c_1 \operatorname{dist}(x^\ell; \mathcal{X}^*)^{1-\rho} \\
&\leq \sum_{\ell=k_1}^{k} c_1 \tilde{\sigma}^{(1-\rho)(\ell-k_1)} \operatorname{dist}(x^{k_1}; \mathcal{X}^*)^{1-\rho} \leq \frac{c_1}{1-\tilde{\sigma}^{1-\rho}} \operatorname{dist}(x^{k_1}; \mathcal{X}^*)^{1-\rho} \\
&\leq \frac{c_1}{1-\tilde{\sigma}^{1-\rho}} \|x^{k_1} - \bar{x}\|^{1-\rho},
\end{aligned} \tag{58}$$

where the second inequality follows from (55). Thus by (56) and (58) we have

$$\|\hat{x}^k - \bar{x}\| \leq \|\hat{x}^k - x^{k_1}\| + \|x^{k_1} - \bar{x}\| \leq \frac{c_1}{1-\tilde{\sigma}^{1-\rho}} \|x^{k_1} - \bar{x}\|^{1-\rho} + \|x^{k_1} - \bar{x}\| \leq \epsilon_0,$$

which readily implies, similarly to the case where $k = k_1$, the fulfillment of (57) for any $k \geq k_1$. Moreover, remembering that $\bar{x}$ is a limiting point of $\{x^k\}_{k\in K_0}$ and using (57) together with (58) allow us to check that for any $\tilde{k} \geq k_1$ we have

$$\|x^k - \bar{x}\| \leq \frac{c_1}{1-\tilde{\sigma}^{1-\rho}} \|x^{\tilde{k}} - \bar{x}\|^{1-\rho} + \|x^{\tilde{k}} - \bar{x}\| \ \text{ whenever } \ k > \tilde{k}.$$

Further, let $\tilde{x}$ be any limiting point of the original iterative sequence $\{x^k\}$. Then the passage to the limit in the above inequality as $k \to \infty$ gives us

$$\|\tilde{x} - \bar{x}\| \leq \frac{c_1}{1-\tilde{\sigma}^{1-\rho}} \|x^{\tilde{k}} - \bar{x}\|^{1-\rho} + \|x^{\tilde{k}} - \bar{x}\| \ \text{ for all } \ \tilde{k} \geq k_1.$$

Passing finally the limit as $\tilde{k} \to \infty$ in the latter inequality and recalling that $\bar{x}$ is a limiting point of $\{x^k\}_{k\in K_0}$ tell us that $\|\tilde{x} - \bar{x}\| = 0$, which verifies therefore that $\{x^k\}$ converges to $\bar{x}$ as $k \to \infty$. The claimed estimate (51) of the convergence rate follows now from (53), and this completes the proof of the theorem. □

To conclude this section, observe that the standard choice of $B_k = \nabla^2 f(x^k)$ in Algorithm 1 clearly implies that the assumed representation $B_k = A^T D_k A$ and the condition $\|B_k - \nabla^2 f(x^k)\| \to 0$ as $k \to \infty$ *hold automatically* due to $\nabla^2 f(x^k) = A^T \nabla^2 h(Ax^k) A$ and the positive semidefiniteness of the Hessian $\nabla^2 h(Ax^k)$ under Assumption 5.1 on the loss function f imposed here.

6 Numerical experiments for regularized logistic regression

In the last section of the paper we conduct numerical experiments on solving the l_1 regularized logistic regression problem to support our theoretical results and compare them with the numerical algorithm from [26] applicable to this problem. All of the numerical experiments are implemented on a laptop with Intel(R) Core(TM) i7-9750H CPU@ 2.60GHz and 32.00 GB memory. All the codes are written in MATLAB 2021a.

Supposing that we have some given training data pairs $(a_i, b_i) \in \mathbb{R}^n \times \{-1, 1\}$ as $i = 1, \ldots, N$, the optimization problem for l_1 regularized logistic regression is defined by

$$\min_x \frac{1}{N} \sum_{i=1}^{N} \log(1 + \exp(-b_i x^T a_i)) + \lambda \|x\|_1, \tag{59}$$

where the regularization term $\|x\|_1$ promotes sparse structures on solutions, and where $\lambda > 0$ is the regularization parameter balancing sparsity and fitting error. Problem (59) is a special case of (1) with $f(x) := \frac{1}{N} \sum_{i=1}^{N} \log(1 + \exp(-b_i x^T a_i))$ and $g(x) := \lambda \|x\|_1$. In all the experiments, the matrix B_k in our proximal Newton-type Algorithm 1 is chosen as the Hessian matrix of f at the iterate x^k, i.e., $B_k := \nabla^2 f(x^k)$. We set $\nu := 0.9$ and $\varrho := \rho$ in the inexact conditions (14) for determining an inexact solution $\hat{x}^k$ to subproblem (3). We also set $\theta := 0.1$, $\sigma := 0.5$, $\gamma := 0.5$, $C := 2F(x^0)$, $\bar{\alpha} := 10^{-4}$ and $c := 10^{-8}$, and then test the three values $\{0.1, 0.5, 1\}$ of parameter ρ in Algorithm 1. As shown in [8, Theorem 8], the subgradient mapping $\nabla f(x) + \partial g(x)$ is metrically subregular at $(\bar{x}, 0)$ for any $\bar{x} \in \mathcal{X}^*$. It can be easily verified that all the assumptions required in Theorem 4.1 are satisfied, and hence the sequence of iterates generated by the proposed algorithm for the tested problem (59) locally converges to the prescribed optimal solution with a superlinear/quadratic convergence rate.

Here we test four real datasets "colon-cancer", "duke breast-cancer", "leukemia" and "rcv1_train.binary" downloaded from the SVMLib repository [43][1], and their sizes are given in Table 1, where nnz(A) denotes the number of nonzero elements of the feature matrix A. All of these tested real datasets have more columns than rows, and hence the loss function f in the corresponding problem (59) is not strongly convex. We use `normr` in Matlab to normalize the rows of each dataset to make them have unit norm.

Since the IRPN algorithm proposed in [26] does not require f in (59) to be strongly convex and problem (59) satisfies all the assumptions required by IRPN, we are going

[1] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/.

Table 1 Tested datasets

Dataset	Data points (N)	Features (n)	nnz (A)	Density(A) (%)
colon-cancer	62	2000	124000	100
duke breast-cancer	44	7129	313676	100
leukemia	38	7129	270902	100
rcv1_train.binary	20,242	47,236	1498952	0.16

to compare our proposed proximal Newton-type Algorithm 1 with IRPN. Note that the IRPN code is collected from https://github.com/ZiruiZhou/IRPN with some modifications to match the objective function in (59). We set $\theta = \beta := 0.25$, $\zeta := 0.4$, $\eta := 0.5$ and $c := 10^{-6}$ in IRPN as suggested in [26]. Also we set $\rho := 0$ and 0.5, since these two specifications of IRPN perform best as shown in [26]. It should be noticed that in such setting both our Algorithm 1 and IRPN require solving subproblem (3) at each iteration. This subproblem can be solved by the coordinate gradient descent method, which is implemented in MATLAB as a C source MEX-file.[2]

We also compare our proposed proximal Newton-type Algorithm 1 with the proximal-Newton method PNOPT (Proximal Newton OPTimizer) proposed in [9], although the theoretical result of PNOPT in [9] requires the strong convexity of f in (59). Note that the PNOPT code is collected from http://web.stanford.edu/group/SOL/software/pnopt/ with replacing the subproblem solver by the coordinate gradient descent method mentioned above for a fair comparison.

The initial points in all experiments are set to be a zero vector. Each method is terminated at the iterate x^k if the accuracy TOL is reached by $\|\mathcal{G}(x^k)\| \leq \texttt{TOL}$ with the residual $\|\mathcal{G}(x^k)\|$ defined via the prox-gradient mapping (8). The maximum number of iterations for the coordinate gradient descent method to solve the corresponding subproblems is 10000.

The achieved numerical results are presented in Tables 2, 3, 4, and 5.

We employ the two values $\{10^{-4}, 10^{-6}\}$ of penalty parameter λ for each dataset and the six levels $\{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}\}$ of accuracy TOL in the algorithms, and report the number of outer and inner iterations along with CPU time, where the inner iterations denote the total number of coordinate descent cycles of the coordinate gradient descent method during implementation. In all of the tests, line search procedures of each tested method provide the unit step size. This may be because the zero point is a good initial point for all tested problems. It can be observed from the numerical results that our proposed proximal Newton-type Algorithm 1 with $\rho = 0.1$ achieves the desired accuracy with the least total iteration number and time in many tested problems. Though in some tested problems, PNOPT achieves the desired accuracy with the least total iteration number and time, Algorithm 1 with $\rho = 0.1$ is still comparable with PNOPT. Although there is no theoretical guarantee for PNOPT in problems (59) in the absence of the strong convexity assumption, this algorithm happens to be efficient in practice. It can also be seen that Algorithm 1 with $\rho = 0.5$ and 1 always achieves the desired accuracy with the least outer iteration number. This

[2] The code is downloaded from https://github.com/ZiruiZhou/IRPN.

Table 2 Numerical comparison on colon-cancer dataset with $\lambda = 10^{-4}$ and $\lambda = 10^{-6}$

		$\lambda = 10^{-4}$						$\lambda = 10^{-6}$					
TOL	Solver	PNOPT	IRPN $\rho = 0$	IRPN $\rho = 0.5$	Algorithm 1 $\rho = 0.1$	Algorithm 1 $\rho = 0.5$	Algorithm 1 $\rho = 1$	PNOPT	IRPN $\rho = 0$	IRPN $\rho = 0.5$	Algorithm 1 $\rho = 0.1$	Algorithm 1 $\rho = 0.5$	Algorithm 1 $\rho = 1$
10^{-3}	Outer iterations	4	4	4	4	4	4	5	5	5	5	5	5
	Inner iterations	22	8	61	8	48	210	10	11	13	10	10	889
	Time(s)	0.01	0.01	0.02	0.01	0.02	0.08	0.00	0.01	0.01	0.01	0.01	0.32
10^{-4}	Outer iterations	5	7	5	7	5	5	7	7	7	7	7	7
	Inner iterations	54	29	127	42	107	319	15	17	231	14	136	1894
	Time(s)	0.02	0.01	0.04	0.03	0.05	0.12	0.01	0.01	0.08	0.01	0.05	0.66
10^{-5}	Outer iterations	7	12	6	9	6	6	9	14	9	9	9	9
	Inner iterations	103	74	188	89	150	442	67	31	464	23	357	3164
	Time(s)	0.04	0.03	0.06	0.05	0.06	0.16	0.02	0.02	0.15	0.01	0.12	1.07
10^{-6}	Outer iterations	8	24	7	11	7	7	10	156	10	11	10	10
	Inner iterations	134	155	264	118	200	603	113	315	731	88	580	3680
	Time(s)	0.05	0.06	0.09	0.06	0.08	0.22	0.04	0.16	0.23	0.03	0.19	1.25
10^{-7}	Outer iterations	9	43	7	12	7	7	12	878	11	13	11	11
	Inner iterations	176	264	264	135	200	603	258	1759	1025	168	813	4226
	Time(s)	0.07	0.10	0.09	0.06	0.08	0.22	0.08	0.92	0.33	0.06	0.27	1.42
10^{-8}	Outer iterations	10	66	8	13	8	8	13	3137	13	18	12	12
	Inner iterations	229	379	410	153	334	895	294	6277	1526	401	1049	4770
	Time(s)	0.09	0.15	0.14	0.07	0.14	0.33	0.09	3.22	0.49	0.13	0.34	1.59

Table 3 Numerical comparison on leukemia dataset with $\lambda = 10^{-4}$ and $\lambda = 10^{-6}$

		$\lambda = 10^{-4}$						$\lambda = 10^{-6}$					
TOL	Solver	PNOPT	IRPN $\rho = 0$	IRPN $\rho = 0.5$	Algorithm 1 $\rho = 0.1$	Algorithm 1 $\rho = 0.5$	Algorithm 1 $\rho = 1$	PNOPT	IRPN $\rho = 0$	IRPN $\rho = 0.5$	Algorithm 1 $\rho = 0.1$	Algorithm 1 $\rho = 0.5$	Algorithm 1 $\rho = 1$
10^{-3}	Outer iterations	3	3	3	3	3	3	5	5	5	5	5	5
	Inner iterations	8	8	22	6	14	89	10	12	10	10	10	561
	Time(s)	0.01	0.01	0.02	0.01	0.01	0.07	0.01	0.02	0.02	0.01	0.01	0.40
10^{-4}	Outer iterations	5	5	5	6	5	5	7	7	7	7	7	7
	Inner iterations	56	22	123	31	116	286	15	19	207	14	122	1298
	Time(s)	0.04	0.02	0.09	0.03	0.09	0.21	0.02	0.03	0.18	0.02	0.09	0.93
10^{-5}	Outer iterations	6	12	6	8	6	6	8	11	8	8	8	8
	Inner iterations	85	55	231	67	219	511	34	28	351	16	219	1833
	Time(s)	0.06	0.05	0.17	0.05	0.16	0.37	0.03	0.04	0.29	0.02	0.16	1.30
10^{-6}	Outer iterations	8	59	7	10	7	7	10	181	10	12	9	9
	Inner iterations	164	243	382	118	344	794	124	368	584	51	316	2088
	Time(s)	0.12	0.24	0.28	0.09	0.24	0.56	0.10	0.47	0.48	0.05	0.23	1.48
10^{-7}	Outer iterations	8	100	7	11	7	7	12	1353	11	16	11	11
	Inner iterations	164	406	382	162	344	794	226	2712	864	152	812	3236
	Time(s)	0.12	0.40	0.28	0.12	0.24	0.56	0.19	3.36	0.70	0.13	0.57	2.28
10^{-8}	Outer iterations	9	138	8	13	8	8	13	6869	14	27	12	12
	Inner iterations	225	558	572	242	523	1177	316	13744	1519	403	1090	3783
	Time(s)	0.16	0.56	0.41	0.18	0.37	0.82	0.26	18.93	1.21	0.32	0.76	2.66

Table 4 Numerical comparison on duke breast-cancer dataset with $\lambda = 10^{-4}$ and $\lambda = 10^{-6}$

		$\lambda = 10^{-4}$						$\lambda = 10^{-6}$					
TOL	Solver	PNOPT	IRPN $\rho = 0$	IRPN $\rho = 0.5$	Algorithm 1 $\rho = 0.1$	Algorithm 1 $\rho = 0.5$	Algorithm 1 $\rho = 1$	PNOPT	IRPN $\rho = 0$	IRPN $\rho = 0.5$	Algorithm 1 $\rho = 0.1$	Algorithm 1 $\rho = 0.5$	Algorithm 1 $\rho = 1$
10^{-3}	Outer iterations	3	3	3	3	3	3	5	5	5	5	5	5
	Inner iterations	10	7	25	6	15	86	10	12	12	10	10	545
	Time(s)	0.01	0.01	0.03	0.01	0.02	0.08	0.02	0.02	0.02	0.01	0.02	0.50
10^{-4}	Outer iterations	5	6	5	6	5	5	7	7	7	7	7	7
	Inner iterations	46	20	106	21	85	199	17	20	240	14	133	1234
	Time(s)	0.05	0.02	0.11	0.03	0.08	0.18	0.02	0.03	0.23	0.02	0.13	1.10
10^{-5}	Outer iterations	6	11	6	8	6	6	8	13	8	8	8	8
	Inner iterations	68	41	136	45	110	279	38	32	289	19	217	1868
	Time(s)	0.07	0.05	0.14	0.05	0.11	0.25	0.04	0.05	0.27	0.02	0.21	1.66
10^{-6}	Outer iterations	7	40	7	10	7	7	10	163	10	10	9	9
	Inner iterations	87	148	202	86	162	411	71	332	439	51	307	2507
	Time(s)	0.08	0.19	0.20	0.09	0.16	0.37	0.08	0.53	0.41	0.05	0.29	2.22
10^{-7}	Outer iterations	8	94	7	11	7	7	12	1098	11	12	11	11
	Inner iterations	113	362	202	115	162	411	158	2202	532	85	649	3468
	Time(s)	0.11	0.47	0.20	0.12	0.16	0.37	0.16	3.51	0.49	0.09	0.59	3.06
10^{-8}	Outer iterations	9	118	8	13	8	8	13	4738	12	22	11	11
	Inner iterations	138	441	300	149	252	587	234	9482	755	284	649	3468
	Time(s)	0.13	0.58	0.29	0.15	0.24	0.52	0.24	15.13	0.69	0.28	0.59	3.06

Table 5 Numerical comparison on rcv1_train.binary dataset with $\lambda = 10^{-4}$ and $\lambda = 10^{-6}$

		$\lambda = 10^{-4}$						$\lambda = 10^{-6}$					
TOL	Solver	PNOPT	IRPN $\rho = 0$	IRPN $\rho = 0.5$	Algorithm 1 $\rho = 0.1$	Algorithm 1 $\rho = 0.5$	Algorithm 1 $\rho = 1$	PNOPT	IRPN $\rho = 0$	IRPN $\rho = 0.5$	Algorithm 1 $\rho = 0.1$	Algorithm 1 $\rho = 0.5$	Algorithm 1 $\rho = 1$
10^{-3}	Outer iterations	3	3	3	3	3	3	4	4	4	4	4	4
	Inner iterations	13	10	30	7	23	96	13	12	69	8	45	511
	Time(s)	0.12	0.12	0.24	0.08	0.19	0.68	0.13	0.16	0.55	0.09	0.37	3.77
10^{-4}	Outer iterations	5	5	4	5	4	4	6	6	6	8	6	6
	Inner iterations	52	22	43	21	42	190	36	16	199	20	208	1851
	Time(s)	0.39	0.23	0.35	0.19	0.33	1.33	0.31	0.21	1.50	0.22	1.52	13.09
10^{-5}	Outer iterations	6	8	5	7	5	5	9	24	9	11	9	9
	Inner iterations	95	81	84	64	96	326	167	83	1146	107	1183	5515
	Time(s)	0.69	0.74	0.64	0.51	0.71	2.30	1.24	0.90	8.13	0.85	8.31	38.39
10^{-6}	Outer iterations	7	10	6	9	6	6	11	125	10	13	10	10
	Inner iterations	139	121	132	121	180	506	501	362	2003	326	1917	13041
	Time(s)	1.01	1.07	0.99	0.91	1.30	3.58	3.58	4.05	14.04	2.38	13.43	90.34
10^{-7}	Outer iterations	8	12	7	11	7	7	12	386	12	15	12	11
	Inner iterations	192	161	256	191	303	765	806	1115	4154	908	5777	23041
	Time(s)	1.38	1.42	1.89	1.42	2.15	5.35	5.69	12.44	28.93	6.78	40.11	159.03
10^{-8}	Outer iterations	9	14	7	12	7	7	13	998	12	16	12	12
	Inner iterations	236	201	256	223	303	765	1259	2503	4154	1338	5777	33041
	Time(s)	1.69	1.76	1.89	1.66	2.15	5.35	8.86	29.48	28.93	9.82	40.11	228.39

supports the convergence rate result of Theorem 30 telling us that with larger values of ρ Algorithm 1 achieves a higher order of the convergence rate. However, larger ρ causes η_k in the inexact condition (14) to decrease faster, which makes the inexact condition (14) more restrictive, and it will take more inner iterations for solving the subproblem at each outer iteration. Hence, in practice, for the total computation time we need to trade off between the outer iteration number and the inner iteration number for solving each subproblem. For the l_1 regularized logistic regression problem (59) tested in this section, and when we use the coordinate gradient descent method for solving the subproblem at each iteration of Algorithm 1, the value $\rho = 0.1$ seems to be a good choice for achieving an overall efficient performance.

Acknowledgements The authors are very grateful to two anonymous referees for their helpful suggestions and remarks that allowed us to significantly improve the original presentation. The alphabetical order of the authors indicates their equal contributions to the paper.

References

1. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**, 83–202 (2009)
2. Nesterov, Yu.: Lectures on Convex Optimization, 2nd edn. Springer, Cham, Switzerland (2018)
3. Schmidt, M., Roux, N., Bach, F.: Convergence rates of inexact proximal-gradient methods for convex optimization. In: Shawe-Taylor, J., et al. (eds.) Advances in Neural Information Processing Systems 24, pp. 1458–1466. Curran Associates, New York (2011)
4. Li, G., Pong, T.K.: Calculus of the exponent of Kurdyka-Łojasiewicz inequality and its applications to linear convergence of first-order methods. Found. Comput. Math. **18**, 1199–1232 (2018)
5. Luo, Z.-Q., Tseng, P.: On the linear convergence of descent methods for convex essentially smooth minimization. SIAM J. Control. Optim. **30**, 408–425 (1992)
6. Necoara, I., Nesterov, Yu., Glineur, F.: Linear convergence of first order methods for non-strongly convex optimization. Math. Progr. **175**, 69–107 (2019)
7. Tseng, P.: Approximation accuracy, gradient methods, and error bound for structured convex optimization. Math. Progr. **125**, 263–295 (2010)
8. Ye, J.J., Yuan, X., Zeng, S., Zhang, J.: Variational analysis perspective on linear convergence of some first order methods for nonsmooth convex optimization problems. Set-Valued Var. Anal. (2021). https://doi.org/10.1007/s11228-021-00591-3
9. Lee, J.D., Sun, Y., Saunders, M.A.: Proximal Newton-type methods for minimizing composite functions. SIAM J. Optim. **24**, 1420–1443 (2014)
10. Fukushima, M., Mine, H.: A generalized proximal point algorithm for certain non-convex minimization problems. Int. J. Syst. Sci. **12**, 989–1000 (1981)
11. Rockafellar, R.T.: Monotone operators and the proximal point algorithm. SIAM J. Control. Optim. **14**, 877–898 (1976)
12. Facchinei, F., Pang, J.-S.: Finite-Dimesional Variational Inequalities and Complementarity Problems. Springer, New York (2003)
13. Izmailov, A.F., Solodov, M.V.: Newton-Type Methods for Optimization and Variational Problems. Springer, New York (2014)
14. Friedman, J., Hastie, T., Höfling, H., Tibshirani, R.: Pathwise coordinate optimization. Ann. Appl. Stat. **1**, 302–332 (2007)
15. Yuan, G.-X., Ho, C.-H., Lin, C.-J.: An improved GLMNET for L1-regularized logistic regression. J. Mach. Learn. Res. **13**, 1999–2030 (2012)
16. Hsieh, C., Dhillon, I.S., Ravikumar, P.K., Sustik, M.A.: Sparse inverse covariance matrix estimation using quadratic approximation. In: Shawe-Taylor, J., et al. (eds.) Advances in Neural Information Processing Systems 24, pp. 2330–2338. Curran Associates, New York (2011)

17. Oztoprak, F., Nocedal, J., Rennie, S., Olsen, P.A.: Newton-like methods for sparse inverse covariance estimation. In: Pereira, F., et al. (eds.) Advances in Neural Information Processing Systems 25, pp. 755–763. Curran Associates, New York (2012)
18. Sra, S., Nowozin, S., Wright, S.J. (eds.): Optimization for Machine Learning. MIT Press, Cambridge (2011)
19. Robinson, S.M.: Generalized equations and their solutions, Part II: applications to nonlinear programming. Math. Program. Stud. **19**, 200–221 (1982)
20. Mordukhovich, B.S.: Variational Analysis and Applications. Springer, Cham, Switzerland (2018)
21. Dontchev, A.L., Rockafellar, R.T.: Implicit Functions and Solution Mappings: A View from Variational Analysis, 2nd edn. Springer, New York (2014)
22. Fischer, A.: Local behavior of an iterative framework for generalized equations with nonisolated solutions. Math. Progr. **94**, 91–124 (2002)
23. Byrd, R.H., Nocedal, J., Oztoprak, F.: An inexact successive quadratic approximation method for L-1 regularized optimization. Math. Progr. **157**, 375–396 (2016)
24. Lee, C., Wright, S.J.: Inexact successive quadratic approximation for regularized optimization. Comput. Optim. Appl. **72**, 641–0674 (2019)
25. Scheinberg, K., Tang, X.: Practical inexact proximal quasi-Newton method with global complexity analysis. Math. Progr. **160**, 495–529 (2016)
26. Yue, M.-C., Zhou, Z., So, A.M.-C.: A family of inexact SQA methods for non-smooth convex minimization with provable convergence guarantees based on the Luo-Tseng error bound property. Math. Progr. **174**, 327–358 (2019)
27. Themelis, A., Stella, L., Patrinos, P.: Forward-backward envelope for the sum of two nonconvex functions: further properties and nonmonotone linesearch algorithms. SIAM J. Optim. **28**, 2274–2303 (2018)
28. Mordukhovich, B.S., Nam, M.N.: An Easy Path to Convex Analysis. Morgan & Claypool Publishers, San Rafael, CA (2014)
29. Rockafellar, R.T., Wets, R.J.-B.: Variational Analysis. Springer, Berlin (1998)
30. Aragón Artacho, F.J., Geoffroy, M.H.: Characterization of metric regularity of subdifferentials. J. Convex Anal. **15**, 365–380 (2008)
31. Aragón Artacho, F.J., Geoffroy, M.H.: Metric subregularity of the convex subdifferential in Banach spaces. J. Nonlin. Convex Anal. **15**, 35–47 (2015)
32. Drusvyatskiy, D., Mordukhovich, B.S., Nghia, T.T.A.: Second-order growth, tilt stability, and metric regularity of the subdifferential. J. Convex Anal. **21**, 1165–1192 (2014)
33. Gaydu, M., Geoffroy, M.H., Jean-Alexis, C.: Metric subregularity of order q and the solving of inclusions. Cent. Eur. J. Math. **9**, 147–161 (2011)
34. Li, G., Mordukhovich, B.S.: Hölder metric subregularity with applications to proximal point method. SIAM J. Optim. **22**, 1655–1684 (2012)
35. Zheng, X.Y., Ng, K.F.: Hölder stable minimizers, tilt stability sand Hölder metric regularity of subdifferentials. SIAM J. Optim. **120**, 186–201 (2015)
36. Mordukhovich, B.S., Ouyang, W.: Higher-order metric subregularity and its applications. J. Global Optim. **63**, 777–795 (2015)
37. Khanh, P.D., Mordukhovich, B.S., Phat, V.T.: A generalized Newton method for subgradient systems. arXiv:2009.10551v1 (2020)
38. Mordukhovich, B.S., Sarabi, M.E.: Generalized Newton algorithms for tilt-stable minimizers in nonsmooth optimization. arXiv:2004.02345 (2020)
39. Drusvyatskiy, D., Lewis, A.S.: Error bounds, quadratic growth, and linear convergence of proximal methods. Math. Oper. Res. **43**, 919–948 (2018)
40. Beck, A.: First-Order Methods in Optimization. SIAM, Philadelphia (2017)
41. Chen, X., Fukushima, M.: Proximal quasi-Newton methods for nondifferentiable convex optimization. Math. Progr. **85**, 313–334 (1999)
42. Dan, H., Yamashita, N., Fukushima, M.: Convergence properties of the inexact Levenberg-Marquardt method under local error bound conditions. Optim. Meth. Softw. **17**, 605–626 (2002)
43. Chang, C.C., Lin, C.J.: LIBSVM?: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**, 27:1–27:27 (2011)

Authors and Affiliations

Boris S. Mordukhovich[1] · Xiaoming Yuan[2] · Shangzhi Zeng[3] · Jin Zhang[4]

Xiaoming Yuan
xmyuan@hku.hk

Shangzhi Zeng
zengshangzhi@uvic.ca

1 Department of Mathematics, Wayne State University, Detroit MI,, USA

2 Department of Mathematics, The University of Hong Kong, Hong Kong, China

3 Department of Mathematics and Statistics, University of Victoria, Victoria, Canada

4 Department of Mathematics, Southern University of Science and Technology, National Center for Applied Mathematics Shenzhen, Shenzhen 518055, China