An Inexact Regularized Proximal Newton Method without Line Search

Simeon vom Dahl[*] Christian Kanzow[†]

April 5, 2024

Abstract. In this paper, we introduce an inexact regularized proximal Newton method (IRPNM) that does not require any line search. The method is designed to minimize the sum of a twice continuously differentiable function f and a convex (possibly non-smooth and extended-valued) function φ. Instead of controlling a step size by a line search procedure, we update the regularization parameter in a suitable way, based on the success of the previous iteration. The global convergence of the sequence of iterations and its superlinear convergence rate under a local Hölderian error bound assumption are shown. Notably, these convergence results are obtained without requiring a global Lipschitz property for ∇f, which, to the best of the authors' knowledge, is a novel contribution for proximal Newton methods. To highlight the efficiency of our approach, we provide numerical comparisons with an IRPNM using a line search globalization and a modern FISTA-type method.

Keywords. nonsmooth and nonconvex optimization; regularized proximal Newton method; global and local convergence; Hölderian local error bound

AMS Subject Classifications. 49M15, 65K10, 90C26, 90C30, 90C55

1 Introduction

We are interested in solving the composite optimization problem

$$\min_{x\in\mathbb{R}^n} F(x) := f(x) + \varphi(x) \quad \text{with } f(x) := \psi(Ax-b), \tag{1}$$

where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$ represent some given data, and with $\psi\colon \mathbb{R}^m \to \overline{\mathbb{R}} := \mathbb{R}\cup\{\infty\}$ and $\varphi\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ being proper lower semicontinuous (lsc) functions satisfying the following conditions.

Assumption 1. (a) ψ is twice continuously differentiable on an open set containing $A(\Omega) - b$, where $\Omega \supseteq \operatorname{dom}\varphi$ is a closed subset of $\mathbb{R}^n$,

(b) φ is convex and continuous on its domain $\operatorname{dom}\varphi$,

(c) F is bounded from below, i.e., $F^* := \inf_{x\in\mathbb{R}^n} F(x) > -\infty$.

[*] University of Würzburg, Institute of Mathematics, Emil-Fischer-Str. 30, 97074 Würzburg, Germany; simeon.vomdahl@uni-wuerzburg.de

[†] University of Würzburg, Institute of Mathematics, Emil-Fischer-Str. 30, 97074 Würzburg, Germany; christian.kanzow@uni-wuerzburg.de

From this structure, it is clear that the objective function $F\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ is also proper and lower semicontinuous, but possibly nonsmooth and nonconvex. Assumption 1(a) and the chain rule guarantee that f is twice continuously differentiable on an open set containing Ω with

$$\nabla f(x) = A^\top \nabla\psi(Ax-b), \quad \nabla^2 f(x) = A^\top \nabla^2\psi(Ax-b)A \quad \text{for all } x \in \Omega. \tag{2}$$

Note that model (1) along with the above assumptions is almost the same as in [23]. The only difference lies in assumption 1(c), where [23] requires coerciveness of F instead of boundedness from below. Note that this coercivity is a much stronger condition. In particular, together with the assumed lower semicontinuity assumption, it implies that all sublevel sets are compact, so that (1) has a compact set of minimizers. Moreover, it guarantees global Lipschitz continuity of the gradient ∇f on all sublevel sets of F. The elimination of the coercivity requirement on F is therefore significant.

Problems of type (1) frequently arise in various fields, including statistics, machine learning, image processing, and many others. Notably, the well-known LASSO problem, as introduced by Tibshirani in [36], represents a special (convex) instance of (1). Applications to compressive sensing problems are discussed in detail in [10]. Machine learning applications like low rank approximations are extensively treated in the book [26], and dictionary learning algorithms are surveyed in the monograph [8]. Matrix completion problems, both convex and nonconvex, have been extensively explored in the past [25, 39]. Additionally, [3] serves as a representative example of the numerous applications of (1) in the field of image processing.

1.1 Related Work

Proximal methods have a long history, beginning with Martinet's proximal point algorithm [28, 27]. Later, Rockafellar generalized the theory and applied it to convex minimization problems [34, 33]. The first proximal method for nonconvex problems of the form (1) was the proximal gradient method introduced by Fukushima and Mine [13]. Subsequently, several proximal gradient methods emerged, including the well-known Iterative Shrinkage/Thresholding Algorithm (ISTA) and its accelerated version, FISTA, introduced by Beck and Teboulle [1]. New FISTA-type methods continue to be introduced, such as the recent example in [22] by Liang and Monteiro.

The idea of proximal Newton methods is to find in each step, for a current iterate x^k, an approximate minimizer y^k of the subproblem

$$\min_x \hat{q}_k(x) := f(x^k) + \nabla f(x^k)^\top (x-x^k) + \frac{1}{2}(x-x^k)^\top G_k (x-x^k) + \varphi(x), \tag{3}$$

where G_k is either the Hessian $\nabla^2 f(x^k)$ or a suitable approximation of the exact Hessian. The main difference to proximal gradient methods is the incorporation of second-order information, which leads to a faster convergence rate due to a better local approximation of the nonlinear function f. On the other hand, iterative methods for the solution of the subproblem (3) usually take longer due to the more complex nature of this subproblem. In fact, note that the proximal Newton method reduces to the proximal gradient method if G_k is a multiple of the identity matrix at each iteration, so that the proximal gradient subproblem is (usually) easier to solve, in several applications even analytically.

Stationary points of (1) are given by the solutions of the generalized equation $0 \in \nabla f(x) + \partial\varphi(x)$, where $\partial\varphi(x)$ denotes the (convex) subdifferential of φ at x, and this inclusion can be rewritten as

$$r(x) = 0$$

for a certain residual function, see (10) below for the precise definition. Similary, the stationary conditions of the subproblems (3) reduce to the solution of the partially linearized generalized equation at iterate x^k:

$$0 \in \nabla f(x^k) + G_k(x - x^k) + \partial\varphi(x). \tag{4}$$

Various results on the convergence of iterative methods for solving (4) can be found in the literature. Fischer [9] proposes a very general iterative framework for solving generalized equations and proves local superlinear and quadratic convergence of the resulting iterates under an upper Lipschitz continuity assumption of the solution set map of a perturbed generalized equation. Early proximal Newton methods were designed for special instances of (1), mostly with convex ψ and φ such as GLMNET [11, 12] and newGLMNET [40] for generalized linear models with elastic-net penalties, QUIC [14] for the l_1-regularized Gaussian maximum likelihood estimator and the Newton-Lasso method [32] for the sparse inverse covariance estimation problem.

Lee et al. [21] were the first to propose a generic version of the exact proximal Newton method for (1) with convex f. They assume that ∇f is Lipschitz continuous and show global convergence under the uniform positive definiteness of $\{G_k\}$ and local quadratic convergence under the strong convexity of f and the Lipschitz continuity of $\nabla^2 f$. Byrd et al. [6], considering (1) with the l_1-regularizer $\varphi(x) = \lambda\|x\|_1$, propose an implementable inexactness criterion for minimizing $\hat{q}_k$ while achieving global convergence, and local fast convergence results under similar assumptions to [21]. Their global convergence theory also works for nonconvex f. Yue et al. [41] used the inexactness criterion and the line search procedure of [6] to develop an inexact proximal Newton method with a regularized Hessian and proved its local superlinear and quadratic convergence under the Luo-Tseng error bound condition [24], which is significantly weaker than the strong convexity assumption on f. Mordukhovich et al. [31] further improve on [41] by eliminating an impractical assumption where the parameters of their method satisfy a condition involving a constant that is difficult to estimate. They also prove local superlinear convergence under the metric q-subregularity of ∂F for $q \in \left(\frac{1}{2}, 1\right)$, a condition even weaker than the Luo-Tseng error bound. Their entire analysis, however, concentrates on convex functions f.

While proximal Newton-type methods for problem (1) with convex f have been extensively explored in the past, there has been limited research to date on the case where f is nonconvex. In the previously referenced paper [6], global convergence was established with nonconvex f and the l_1-regularizer, albeit still requiring a strong convexity assumption on f for the local convergence theory. Lee and Wright [20] investigated an inexact proximal Newton method, presenting a sublinear global convergence rate result on the first-order optimality condition for general choices of G_k, with the sole assumption of ∇f being Lipschitz continuous. Combining the advantages of proximal Newton and proximal gradient methods, Kanzow and Lechner [16] introduced a globalized inexact proximal Newton method (GIPN). In this approach, a proximal gradient step is taken whenever the proximal Newton step fails to satisfy a specified sufficient decrease condition. They proved global convergence with a local superlinear convergence rate under the local strong convexity of F and uniformly bounded positive definiteness of G_k. Inspired by the work [38] for smooth nonconvex optimization problems, Liu et al. [23] extended the theory of [31] to the case of (1), where f is allowed to be nonconvex. Instead of the metric q-subregularity on ∂F, they assumed that accumulation points of the iterate sequence satisfy a Hölderian local error bound condition on the set of so-called strongly stationary points to show convergence of the iterates with a local superlinear convergence rate. They achieve a local superlinear convergence rate without F being locally strongly convex. However, they require that F is level-bounded.

All aforementioned works employed a proximal Newton-type method in conjunction with an appropriate line search strategy for global convergence. There has been minimal exploration of proximal Newton methods with alternative globalization strategies. Yamashita and Ueda [37] investigated regularized Newton methods for smooth unconstrained problems, achieving global convergence by adjusting the regularization parameter based on the success of the previous iteration, similar to a trust-region scheme. As of the authors' knowledge, the method described in the PhD thesis [19, Chapter 4] remains the only instance where this globalization strategy was applied within the framework of proximal Newton-type methods.

Historically, the global Lipschitz continuity of ∇f has been a standard assumption for the convergence analysis of proximal gradient and proximal Newton methods. While recent works have successfully eliminated this assumption for proximal gradient methods (see, for example, [4, 17, 15, 7]), there are no known comparable results for proximal Newton methods.

1.2 Our Contributions

In this work, we present a proximal Newton method without a line search for problem (1) under assumption 1. Building upon the selection in [23], we employ the following expression as the regularized Hessian at iteration x^k:

$$G_k = \nabla^2 f(x^k) + \Lambda_k A^\top A + \nu_k \overline{r}_k{}^\delta I \tag{5}$$

with

$$\Lambda_k := a\left[-\lambda_{min}\left(\nabla^2\psi\left(Ax^k - b\right)\right)\right]_+, \quad a \geq 1, \quad \text{and} \quad \delta \in (0,1]. \tag{6}$$

Recall from (2) that G_k can be rewritten as

$$G_k = A^\top\big(\nabla^2\psi(Ax^k - b) + \Lambda_k I\big)A + \nu_k \overline{r}_k{}^\delta I,$$

hence the definition of Λ_k immediately implies that the matrix G_k is positive definite (the first term is positive semidefinite). The only difference to [23] resides in the final term, where the sequence $\{\overline{r}_k\}_{k\in\mathbb{N}_0}$ is recursively given by

$$\overline{r}_0 := \|r(x^0)\| \text{ and } \overline{r}_{k+1} = \begin{cases} \|r(\hat{x}^k)\|, & \text{if } \|r(\hat{x}^k)\| \leq \eta\overline{r}_k \\ \overline{r}_k, & \text{otherwise} \end{cases} \text{ for } k \in \mathbb{N}_0, \tag{7}$$

with $\hat{x}^k$ being an approximate solution of subproblem (3), $\eta \in (0,1)$, and r is the residual function already mentioned before and formally defined in (10) below. Additionally, the regularization parameter ν_k follows an update strategy akin to [37] and [19], detailed in Section 3. Notably, the sequence $\{G_k\}$ is not uniformly positive definite, since $\{\overline{r}_k\}$ converges to 0, as clarified later.

We establish the global convergence of the iterate sequence and its convergence rate of $q(1+\delta) > 1$, assuming the existence of an accumulation point that satisfies a local Hölderian error bound of order $q > \max\left\{\frac{1}{1+\delta}, \delta\right\}$ on the set of strongly stationary points. In comparison with [23], our approach reproduces essentially the same convergence results, employing an update strategy for the regularization parameter instead of a line-search technique. Most notably, we eliminate the requirement for F to be coercive. The coerciveness guarantees a compact minimizer set for (1), as well as the Lipschitz continuity of ∇f on all sublevel sets of F. It is noteworthy that this global Lipschitz continuity of the gradient of f has been a standard assumption in order to prove convergence of the iterate sequence. To the best of the authors' knowledge, this work is the first to eliminate this assumption.

Utilizing the dual semismooth Newton augmented Lagrangian method (SNALM) developed in [23] as a subproblem solver, we compare the performance of our method (IRPNM-reg) with the line search based inexact regularized proximal Newton method (IRPNM-ls) from [23] and AC-FISTA [22] on five distinct test problems.

1.3 Notation

In this paper, $\mathbb{N} = \{1, 2, 3, ...\}$ denotes the set of positive integers and we write $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. The extended real numbers are given by $\overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$. For $a \in \mathbb{R}$ we write $a_+ := \max(0, a)$. For $x \in \mathbb{R}^n$, $\|x\|$ represents the Euclidean norm, $B_\varepsilon(x)$ stands for the closed ball around x with radius $\varepsilon > 0$, and $\operatorname{dist}(x, C)$ denotes the Euclidean distance from x to a closed set $C \subseteq \mathbb{R}^n$. The set $\mathbb{S}^n$ comprises all real symmetric matrices of dimension $n \times n$ and $\mathbb{S}^n_{++}$ is the set of all positive definite matrices in $\mathbb{S}^n$. For $M \in \mathbb{S}^n$, its spectral norm is denoted by $\|M\|$ and $M \succeq 0$ indicates that M is positive semidefinite. The smallest eigenvalue of M is denoted by $\lambda_{min}(M)$. The identity matrix is denoted by I, with its dimension being evident from the context. The domain of a function $g\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ is defined as $\operatorname{dom} g := \{x \in \mathbb{R}^n \mid g(x) < \infty\}$ and g is called proper if $\operatorname{dom} g \neq \emptyset$.

2 Preliminaries

This section summarizes some background material from variational analysis that will be important in our subsequent sections.

First of all, we denote by $\partial F(x)$ the *basic (or limiting or Mordukhovich) subdifferential* of F at x, see the standard references [35, 30] for more details. Its precise definition plays no role in our subsequent discussion since only some of its basic properties will be used. In particular, it is known that, for convex functions, this basic subdifferential simplifies to the well-known convex subdifferential. Furthermore, according to [35, Exercise 8.8(c)], for any $x \in \operatorname{dom} \varphi$, it holds that $\partial F(x) = \nabla f(x) + \partial\varphi(x)$.

Based on this notion, we now introduce two stationarity concepts, the first one being the standard stationarity condition for composite optimization problems, the second one being a stronger concept taken from [23].

Definition 1. A point $x \in \operatorname{dom} \varphi$ is called a

(a) *stationary point* of problem (1) if $0 \in \partial F(x) \ \big(= \nabla f(x) + \partial\varphi(x)\big)$;

(b) *strongly stationary point* of problem (1) if it is a stationary point which, in addition, satisfies $\nabla^2\psi(Ax - b) \succeq 0$.

We denote by S^* and X^* the sets of all stationary and strongly stationary points, respectively.

Note that Assumption 1(a) together with the outer semicontinuity of $\partial\varphi$ implies that the $\mathcal{S}^*$ and $\mathcal{X}^*$ are closed. In contrast to the situation discussed in [23], we stress that both $\mathcal{S}^*$ and $\mathcal{X}^*$ might be empty in our setting due to the removal of the coerciveness assumption on F. We further note that there might be stationary points (i.e., $\mathcal{S}^* \neq \emptyset$), while $\mathcal{X}^*$ is still empty. A local minimizer is always a stationary point, but is not guaranteed to be strongly stationary, and the converse may not be true either.

Proximal Newton-type methods rely on the proximity operator. For a proper, lower semicontinuous and convex function $g\colon \mathbb{R}^n \to \overline{\mathbb{R}}$, the proximity operator $\operatorname{prox}_g\colon \mathbb{R}^n \to \mathbb{R}^n$ is defined by

$$\operatorname{prox}_g(x) := \underset{y}{\operatorname{argmin}} \left\{ g(y) + \frac{1}{2}\|y - x\|^2 \right\}.$$

The objective function $g(y) + \frac{1}{2}\|y - x\|^2$ is strongly convex on $\operatorname{dom} g$. This ensures a unique minimizer for every $x \in \mathbb{R}^n$, i.e., the proximity operator is well-defined. Moreover, the operator is nonexpansive, signifying Lipschitz continuity with constant one. It also satisfies the crucial relationship

$$y = \operatorname{prox}_g(x) \iff y \in x - \partial g(y), \tag{8}$$

which shows that

$$x \in \mathcal{S}^* \iff -\nabla f(x) \in \partial\varphi(x) \iff x = \operatorname{prox}_\varphi(x - \nabla f(x)). \tag{9}$$

Motivated by this, the *residual or prox-gradient mapping* is defined by

$$r(x) := x - \operatorname{prox}_\varphi(x - \nabla f(x)), \quad x \in \mathbb{R}^n. \tag{10}$$

Consequently, $x \in \mathbb{R}^n$ is a stationary point of F if and only if $r(x) = 0$. Hence, the norm of $r(x)$ can be used to measure the stationarity of x.

3 The Algorithm and its Basic Properties

Consider a fixed iteration $k \geq 0$ with a current iterate $x^k \in \mathbb{R}^n$. Then the core task of proximal Newton methods lies in solving the subproblem

$$\min_x q_k(x) := f(x^k) + \nabla f(x^k)^\top (x - x^k) + \frac{1}{2}(x - x^k)^\top \nabla^2 f(x^k)(x - x^k) + \varphi(x). \tag{11}$$

The first part of q_k provides a quadratic approximation of the smooth function f. However, since f is not necessarily convex, $\nabla^2 f(x^k)$ may not be positive semidefinite and, hence, q_k may not be convex. To address this difficulty, we consider the matrix

$$H_k := \nabla^2 f(x^k) + \Lambda_k A^\top A$$

with Λ_k defined in (6). Recall from the discussion following (6) that H_k, simply by definition of Λ_k, is positive semidefinite. Furthermore, given some regularization parameter $\mu_k > 0$, the corresponding matrix G_k from (5), which is given by

$$G_k = H_k + \mu_k I,$$

is then automatically positive definite. This implies that the resulting subproblem

$$\min_x \hat{q}_k(x) := f(x^k) + \nabla f(x^k)^\top (x - x^k) + \frac{1}{2}(x - x^k)^\top G_k (x - x^k) + \varphi(x), \tag{12}$$

has a strongly convex objective function and, thus, a unique solution. Throughout this paper, we write

$$\overline{x}^k := \operatorname{argmin}_x \hat{q}_k(x)$$

for this unique minimum. From a numerical point of view, computing this minimum exactly might be very demanding, and we therefore require an inexact solution only. We denote this inexact solution by $\hat{x}^k$. In order to prove suitable global and local convergence results, this inexact solution has to satisfy certain criteria which measure the quality of the inexact solution. Here, we assume that the inexact solution $\hat{x}^k$ is computed in such a way that the conditions

$$\|R_k(\hat{x}^k)\| \leq \theta \min\left\{\|r(x^k)\|, \|r(x^k)\|^{1+\tau}\right\} \text{ and } F(x^k) - \hat{q}_k(\hat{x}^k) \geq \frac{\alpha\mu_k}{2}\|\hat{x}^k - x^k\|^2 \tag{13}$$

hold, where $\alpha, \theta \in (0,1)$ and $\tau \geq \delta$ are certain constants, and the residual R_k is defined by

$$R_k(x) := x - \operatorname{prox}_\varphi\left(x - \nabla f(x^k) - (H_k + \mu_k I)(x - x^k)\right).$$

Note that R_k is the counterpart of the residual r from (10) for the subproblem (12). In particular, and similar to (9), a vector x is an optimal solution of (12) if and only if $R_k(x) = 0$. This explains why the first condition from (13) serves as an inexactness criterion. Regarding the second condition, we refer to Lemma 7 below for a justification.

Typically, see the recent papers [31] and [23], these (regularized) proximal Newton-type methods are combined with an appropriate line search strategy to achieve global convergence. In this work, our objective is to attain global convergence by controlling the regularization parameter itself, depending on the success of the previous iteration. This idea has already been used in [37] with a regularized Newton method for the minimization of a twice differentiable function. Recently, in the PhD thesis [19], it has been established for proximal Newton methods in the composite setting. To assess the success of a candidate $\hat{x}^k$, we consider the ratio

$$\rho_k := \frac{\text{ared}_k}{\text{pred}_k} \tag{14}$$

between the actual reduction

$$\text{ared}_k := F(x^k) - F(\hat{x}^k) \tag{15}$$

and the predicted reduction

$$\text{pred}_k := F(x^k) - q_k(\hat{x}^k). \tag{16}$$

It is important to note that for the predicted reduction, we use the unregularized approximation q_k instead of $\hat{q}_k$. From the second condition in (13) it follows that

$$\begin{aligned}\text{pred}_k &= F(x^k) - q_k(\hat{x}^k) = F(x^k) - \hat{q}_k(\hat{x}^k) + \frac{1}{2}(\hat{x}^k - x^k)^\top(\Lambda_k A^\top A + \mu_k I)(\hat{x}^k - x^k) \\ &\geq F(x^k) - \hat{q}_k(\hat{x}^k) + \frac{\mu_k}{2}\|\hat{x}^k - x^k\|^2 \geq \frac{\mu_k}{2}\|\hat{x}^k - x^k\|^2\end{aligned} \tag{17}$$

for all $k \geq 0$. In particular the predicted reduction is positive if x^k is not already a stationary point of (1). This follows from the following simple observation.

Remark 2. If $x^k = \hat{x}^k$, then x^k is already a stationary point of (1). Hence, $\text{pred}_k > 0$ at all iterations k such that x^k is not already a stationary point.

Proof. Let $x^k = \hat{x}^k$. Then the definitions of the corresponding residual functions yield $R_k(\hat{x}^k) = R_k(x^k) = r(x^k)$. Since $\theta \in (0,1)$, we then obtain $r(x^k) = 0$ from first inexactness test in (13). □

We are now ready to present our algorithm.

Algorithm 1 Regularized proximal Newton method

```
 1: Choose $x^0 \in \operatorname{dom}\varphi$ and parameters $c_1 \in (0,1)$; $c_2 \in (c_1,1)$; $\sigma_1 \in (0,1)$; $\sigma_2 > 1$; $\eta \in$
    $(0,1)$; $\theta \in (0,1)$; $\alpha \in (0,1)$; $a \geq 1$; $0 < \nu_{min} \leq \nu_0 \leq \overline{\nu}$; $0 < \delta \leq 1$; $\tau \geq \delta$; $p_{min} \in$
    $(0,1/2)$; $\kappa > 1+\delta$. Set $k := 0$; $\overline{r}_0 := \|r(x^0)\|$; $\mu_0 := \nu_0 \overline{r}_0^{\delta}$.
 2: for $k = 0,1,2,...$ do
 3:     Compute an inexact solution $\hat{x}^k$ of the proximal regularized Newton subproblem
        (12) satisfying the inexactness criterion (13).
 4:     Set $d^k := \hat{x}^k - x^k$.
 5:     Compute $\text{pred}_k$, $\text{ared}_k$ and $\rho_k$.
 6:     if $\text{pred}_k \leq p_{min}(1-\theta)\|d^k\| \min\{\|r(x^k)\|, \|r(x^k)\|^{\kappa}\}$ OR $\rho_k \leq c_1$ then
 7:         Set $x^{k+1} = x^k, \nu_{k+1} = \sigma_2 \nu_k$.                          ▷ unsuccessful iteration
 8:     else
 9:         Set $x^{k+1} = \hat{x}^k$.
10:         if $\rho_k \leq c_2$ then
11:             Set $\nu_{k+1} = \min\{\nu_k, \overline{\nu}\}$.                          ▷ successful iteration
12:         else
13:             Set $\nu_{k+1} = \min\{\max\{\sigma_1 \nu_k, \nu_{min}\}, \overline{\nu}\}$.      ▷ highly successful iteration
14:         end if
15:     end if
16:     if $\|r(x^{k+1})\| \leq \eta \overline{r}_k$ then                                  ▷ $k+1 \in \mathcal{K}$
17:         $\overline{r}_{k+1} = \|r(x^{k+1})\|$.
18:     else
19:         $\overline{r}_{k+1} = \overline{r}_k$.
20:     end if
21:     $\mu_{k+1} = \nu_{k+1} \overline{r}_{k+1}^{\delta}$.
22: end for
```

The basic idea of Algorithm 1 is to solve, iteratively, the proximal regularized Newton subproblem (12) and either to accept the inexact solution as the new iterate, provided that this makes a sufficient progress in the sense of the tests in line 6, or to stay at the current point and enlarge the regularization parameter. The steps between lines 10 and 20 are devoted to a very careful update of the parameter ν_k as well as $\overline{r}_k$, hence of the regularization parameter μ_k in line 21, since this update is essential especially for the local convergence analysis where we prove fast local convergence under fairly mild assumptions.

In the remaining part, we state a number of basic properties which might, partially, explain some of these careful updates.

Lemma 3. *(a) The sequence $\{\overline{r}_k\}$ is monotonically decreasing,*

(b) For all $k \geq 0$ it holds that $\|r(x^k)\| > \eta \overline{r}_k$.

(c) For all $k \geq 0$ it holds that $\overline{r}_k \geq \min\{\|r(x^j)\| \mid 0 \leq j \leq k\}$.

Proof. (a) We consider an iteration $k \geq 0$. If the condition $\|r(x^{k+1})\| \leq \eta \overline{r}_k$ in line 16 of Algorithm 1 is satisfied, we get $\overline{r}_{k+1} = \|r(x^{k+1})\| \leq \eta \overline{r}_k < \overline{r}_k$. Otherwise, the algorithm directly sets $\overline{r}_{k+1} = \overline{r}_k$. Combining these two cases shows that $\overline{r}_{k+1} \leq \overline{r}_k$. Hence, the sequence $\{\overline{r}_k\}$ is monotonically decreasing.

(b) For $k = 0$ this property obviously holds. Suppose now that this property holds for some $k \in \mathbb{N}_0$. If the condition in line 16 is satisfied at iteration k, the algorithm directly sets $\overline{r}_{k+1} = \|r(x^{k+1})\|$. If k is an unsuccessful iteration, then we get $\|r(x^{k+1})\| = \|r(x^k)\| > \eta\overline{r}_k \geq \eta\overline{r}_{k+1}$ by using the induction hypothesis together with $x^{k+1} = x^k$ and (a). In the remaining case it holds that $\|r(x^{k+1})\| > \eta\overline{r}_k$ and iteration k is successful or highly successful. Then we get

$$\|r(x^{k+1})\| > \eta\overline{r}_k = \eta\overline{r}_{k+1}.$$

The combination of the three cases yields the result.

(c) For $k = 0$ this property obviously holds. Suppose now that this property holds for some $k \in \mathbb{N}_0$. If the condition in line 16 is satisfied at iteration k we get

$$\overline{r}_{k+1} = \|r(x^{k+1})\| \geq \min\{\|r(x^j)\| \mid 0 \leq j \leq k+1\}.$$

Otherwise it holds that

$$\overline{r}_{k+1} = \overline{r}_k \geq \min\{\|r(x^j)\| \mid 0 \leq j \leq k\} \geq \min\{\|r(x^j)\| \mid 0 \leq j \leq k+1\},$$

where we used the induction hypothesis in the first inequality. □

The following result contains some estimates regarding the sequence $\{\nu_k\}$ and the corresponding sequence $\{\mu_k\}$ of regularization parameters.

Lemma 4. *For an iteration $k \geq 0$ it holds that*

(a) $\nu_k \geq \nu_{min}$,

(b) $x^{k+1} = x^k$, $\nu_{k+1} = \sigma_2\nu_k > \nu_k$ and $\mu_{k+1} = \sigma_2\mu_k > \mu_k$, if k is unsuccessful,

(c) $x^{k+1} = \hat{x}^k$, $\nu_{k+1} \leq \nu_k$ and $\mu_{k+1} \leq \mu_k$, if k is successful or highly successful.

Proof. (a) This statement follows recursively from $\nu_0 \geq \nu_{min}$ and the possible updates for ν_k in the algorithm.

(b) If k is an unsuccessful iteration, it follows by definition of the algorithm that $x^{k+1} = x^k$ and $\nu_{k+1} = \sigma_2\nu_k$. From Lemma 3(b) it immediately follows that $\|r(x^{k+1})\| = \|r(x^k)\| > \eta\overline{r}_k$, hence $\overline{r}_{k+1} = \overline{r}_k$ and eventually $\mu_{k+1} = \nu_{k+1}\overline{r}_{k+1}^\delta = \sigma_2\nu_k\overline{r}_k^\delta = \sigma_2\mu_k > \mu_k$.

(c) If k is a successful or highly successful iteration, it follows by definition of the algorithm and statement (a) that $x^{k+1} = \hat{x}^k$ and $\nu_{k+1} \leq \nu_k$. Using Lemma 3(a), we then get $\mu_{k+1} = \nu_{k+1}\overline{r}_{k+1}^\delta \leq \nu_k\overline{r}_k^\delta = \mu_k$. □

In the following we consider the set $\mathcal{K} \subset \mathbb{N}_0$ of iterations

$$\mathcal{K} := \{0\} \cup \{k \in \mathbb{N} \mid \text{The if-condition in line 16 was satisfied at iteration } k-1\}.$$

Several properties for the iterates k belonging to this set are summarized in the next result.

Lemma 5. *For all iterations $k \in \mathcal{K} \setminus \{0\} \subset \mathbb{N}$, the following properties hold:*

(a) $\|r(x^k)\| \leq \eta\overline{r}_{k-1}$,

(b) $\overline{r}_k = \|r(x^k)\|$,

(c) iteration $k-1$ was successful or highly successful,

(d) $\nu_k \leq \overline{\nu}$,

(e) $\mu_k \leq \overline{\nu}\|r(x^k)\|^\delta$.

Proof. Statements (a) and (b) follow directly from the if-condition in line 16 and the command in line 17. If iteration $k-1$ was unsuccessful, then it would follow from Lemma 3(b) that $\|r(x^k)\| = \|r(x^{k-1})\| > \eta\overline{r}_{k-1}$, a contradiction to $k \in \mathcal{K}$ according to (a). Hence (c) holds. Assertion (d) then follows from (c) and assertion (e) follows from (b) and (d). □

The index set $\mathcal{K}$ plays a central role in our convergence analysis. The following result indicates why this set is so important.

Lemma 6. *Let $\mathcal{K} = \{k_0, k_1, k_2, ...\}$. For all $i \in \mathbb{N}_0$ it then holds that $\overline{r}_{k_{i+1}} \leq \eta\overline{r}_{k_i}$ and the following three statements are equivalent:*

(i) *$\mathcal{K}$ is an infinite set.*

(ii) $\lim_{k\in\mathcal{K}} \|r(x^k)\| = 0$.

(iii) $\liminf_{k\to\infty} \|r(x^k)\| = 0$.

Proof. Consider $i \in \mathbb{N}$ and $k_i \in \mathcal{K}$. From Lemma 5(a), 5(b) and Lemma 3(a) it then follows immediately that $\overline{r}_{k_i} \leq \eta\overline{r}_{k_i-1} \leq \eta\overline{r}_{k_{i-1}}$. If $\mathcal{K}$ is an infinite set and using Lemma 3(a), this directly implies $\lim_{k\in\mathcal{K}} \overline{r}_k = \lim_{k\in\mathcal{K}} \|r(x^k)\| = 0$. From Lemma 3(c) it follows that $\liminf_{k\to\infty} \|r(x^k)\| = 0$. Suppose now that $\mathcal{K}$ is not an infinite set. Denote the last iteration in $\mathcal{K}$ by $\bar{k}$. Then it holds that $\overline{r}_k = \overline{r}_{\bar{k}}$ for all $k \geq \bar{k}$. It follows from Lemma 3(b) that $\|r(x^k)\| > \eta\overline{r}_k = \eta\overline{r}_{\bar{k}}$ for all $k \geq \bar{k}$. Hence, $\liminf_{k\to\infty} \|r(x^k)\| > 0$. □

4 Global Convergence Results

This section presents global convergence results which are in the same spirit as those known for trust-region-type methods.

The first result states that the inexactness criterion (13) is feasible, which implies that Algorithm 1 is well-defined.

Lemma 7. *For every $k \in \mathbb{N}_0$ such that x^k is not a stationary point of* (1)*, the inexactness criterion* (13) *is satisfied for any $x \in \operatorname{dom}\varphi$ sufficiently close to the exact solution $\overline{x}^k$ of* (12)*.*

Proof. Recall that there are two criteria in (13). We show that both of them hold for all x sufficiently close to the global minimum of the underlying subproblem. Hence, consider a fixed iteration index $k \in \mathbb{N}_0$ and assume that x^k is not already a stationary point of the given composite optimization problem (1). Since $\|R_k(\overline{x}^k)\| = 0$, it follows from the continuity of R_k relative to $\operatorname{dom}\varphi$ that

$$\|R_k(x)\| \leq \theta \min\left\{\|r(x^k)\|, \|r(x^k)\|^{1+\tau}\right\}$$

holds for $x \in \operatorname{dom}\varphi$ sufficiently close to $\overline{x}^k$, showing that the first test in (13) holds for these x. Furthermore, from [21, Proposition 2.4], it follows that the exact solution $\overline{x}^k$ of subproblem (12) satisfies

$$\nabla f(x^k)^\top(\overline{x}^k - x^k) + \varphi(\overline{x}^k) - \varphi(x^k) \leq -(\overline{x}^k - x^k)^\top G_k(\overline{x}^k - x^k). \tag{18}$$

Therefore we obtain

$$
\begin{aligned}
F(x^k) - \hat{q}_k(\overline{x}^k) &= \varphi(x^k) - \nabla f(x^k)^\top(\overline{x}^k - x^k) - \frac{1}{2}(\overline{x}^k - x^k)^\top G_k(\overline{x}^k - x^k) - \varphi(\overline{x}^k) \\
&= -\left(\nabla f(x^k)^\top(\overline{x}^k - x^k) + \varphi(\overline{x}^k) - \varphi(x^k)\right) - \frac{1}{2}(\overline{x}^k - x^k)^\top G_k(\overline{x}^k - x^k) \\
&\geq \frac{1}{2}(\overline{x}^k - x^k)^\top G_k(\overline{x}^k - x^k) \geq \frac{\mu_k}{2}\|\overline{x}^k - x^k\|^2 > \frac{\alpha\mu_k}{2}\|\overline{x}^k - x^k\|^2,
\end{aligned}
\tag{19}
$$

where the first inequality follows from (18) and the second from the positive semidefiniteness of H_k. From the continuity of $F(x^k) - \hat{q}_k(\cdot) - \frac{\alpha\mu_k}{2}\|\cdot - x^k\|^2$ relative to $\operatorname{dom}\varphi$, it follows that

$$
F(x^k) - \hat{q}_k(x) > \frac{\alpha\mu_k}{2}\|x - x^k\|^2
$$

holds for all $x \in \operatorname{dom}\varphi$ sufficiently close to $\overline{x}^k$.

□

The next result provides a lower and upper bound of the residual $r(x^k)$ in terms of the vector d^k.

Lemma 8. *For all $k \in \mathbb{N}_0$, it holds that*

$$
\frac{\mu_k}{(1 + \|G_k\|)(1+\theta)}\|d^k\| \leq \|r(x^k)\| \leq \frac{1 + \|G_k\|}{1-\theta}\|d^k\|.
$$

Proof. By $r(x^k) = x^k - \operatorname{prox}_\varphi(x^k - \nabla f(x^k))$, we get from (8) that $r(x^k) \in \nabla f(x^k) + \partial\varphi\left(x^k - r(x^k)\right)$. In the same way, $R_k(\hat{x}^k) \in \nabla f(x^k) + G_k d^k + \partial\varphi(\hat{x}^k - R_k(\hat{x}^k))$ follows from the definition of the proximal operator. The monotonicity of the subgradient mapping $\partial\varphi$ ensures that

$$
\left\langle R_k(\hat{x}^k) - r(x^k) - G_k d^k,\, d^k + r(x^k) - R_k(\hat{x}^k)\right\rangle \geq 0. \tag{20}
$$

Simply reordering the left-hand side yields

$$
0 \leq -\|r(x^k)\|^2 - \|R_k(\hat{x}^k)\|^2 + 2\langle R_k(\hat{x}^k), r(x^k)\rangle - (d^k)^\top G_k d^k + \langle R_k(\hat{x}^k) - r(x^k), d^k + G_k d^k\rangle.
$$

This implies

$$
\begin{aligned}
\|r(x^k) - R_k(\hat{x}^k)\|^2 &\leq \|r(x^k)\|^2 - 2\left\langle R_k(\hat{x}^k), r(x^k)\right\rangle + \|R_k(\hat{x}^k)\|^2 + (d^k)^\top G_k d^k \\
&\leq \left\langle R_k(\hat{x}^k) - r(x^k),\, d^k + G_k d^k\right\rangle \\
&\leq \|r(x^k) - R_k(\hat{x}^k)\| \cdot (1 + \|G_k\|)\|d^k\|.
\end{aligned}
$$

Together with the inexactness criterion $\|R_k(\hat{x}^k)\| \leq \theta\|r(x^k)\|$ and the Cauchy-Schwarz inequality, this results in

$$
\|r(x^k)\| \leq \|r(x^k) - R_k(\hat{x}^k)\| + \|R_k(\hat{x}^k)\| \leq (1 + \|G_k\|)\|d^k\| + \theta\|r(x^k)\|.
$$

Remembering $\theta \in (0,1)$, we get the upper estimate

$$
\|r(x^k)\| \leq \frac{1 + \|G_k\|}{1-\theta}\|d^k\|.
$$

Reordering (20) in a different way yields

$$\langle d^k, G_k d^k \rangle \leq \langle R_k(\hat{x}^k) - r(x^k), d^k - R_k(\hat{x}^k) + r(x^k) + G_k d^k \rangle \leq \langle (I + G_k) d^k, R_k(\hat{x}^k) - r(x^k) \rangle.$$

Using $G_k \succeq \mu_k I$ and the Cauchy-Schwarz inequality, we therefore get

$$\mu_k \|d^k\|^2 \leq \langle d^k, G_k d^k \rangle \leq (1 + \|G_k\|) \|d^k\| \|R_k(\hat{x}^k) - r(x^k)\| \leq (1 + \|G_k\|) \|d^k\| (1 + \theta) \|r(x^k)\|,$$

where the last inequality follows from (13). Hence, dividing by $\mu_k \|d^k\|$ (in the case of $\|d^k\| = 0$, the resulting inequality holds trivially) yields

$$\|d^k\| \leq \frac{(1 + \theta)(1 + \|G_k\|)}{\mu_k} \|r(x^k)\|.$$

This completes the proof. □

The following result provides (implicitly) a condition under which the quotient between the actual and the predicted reduction is greater than a suitable constant (note that, in the following, we often exploit the observation from Remark 2 that the predicted reduction is a positive number, without explicitly mentioning this fact).

Lemma 9. *Let $c \leq 1$. For every $k \geq 0$, there exists ξ^k on the line segment between x^k and $\hat{x}^k$ such that*

$$\mathit{ared}_k - c\, \mathit{pred}_k \geq \frac{1}{2} \left((1 - c)\mu_k - \|\nabla^2 f(\xi^k) - \nabla^2 f(x^k)\| \right) \|d^k\|^2. \tag{21}$$

Proof. It follows from Taylor's formula and the convexity of $\operatorname{dom} \varphi$ that, for every $k \geq 0$, there exists $\xi^k \in \operatorname{dom} \varphi$ on the line segment between x^k and $\hat{x}^k$ such that

$$f(\hat{x}^k) - f(x^k) - \nabla f(x^k)^\top d^k = \frac{1}{2} (d^k)^\top \nabla^2 f(\xi^k) d^k.$$

This yields

$$\begin{aligned} F(\hat{x}^k) - q_k(\hat{x}^k) &= f(\hat{x}^k) - f(x^k) - \nabla f(x^k)^\top d^k - \frac{1}{2} (d^k)^\top \nabla^2 f(x^k) d^k \\ &= \frac{1}{2} (d^k)^\top (\nabla^2 f(\xi^k) - \nabla^2 f(x^k)) d^k \\ &\leq \frac{1}{2} \|\nabla^2 f(\xi^k) - \nabla^2 f(x^k)\| \|d^k\|^2. \end{aligned}$$

Using this inequality together with (17), we get

$$\begin{aligned} \mathrm{ared}_k - c\, \mathrm{pred}_k &= (1 - c)\mathrm{pred}_k - \mathrm{pred}_k + \mathrm{ared}_k = (1 - c)\mathrm{pred}_k - \left(F(\hat{x}^k) - q_k(\hat{x}^k) \right) \\ &\geq \frac{1 - c}{2} \mu_k \|d^k\|^2 - \frac{1}{2} \|\nabla^2 f(\xi^k) - \nabla^2 f(x^k)\| \|d^k\|^2 \\ &= \frac{1}{2} \left((1 - c)\mu_k - \|\nabla^2 f(\xi^k) - \nabla^2 f(x^k)\| \right) \|d^k\|^2. \end{aligned}$$

This completes the proof. □

We next show that Algorithm 1 generates infinitely many successful or highly successful iterates.

Theorem 10. *Suppose* $\|r(x^k)\| \neq 0$ *for all* $k \geq 0$. *Then Algorithm 1 performs infinitely many successful or highly successful iterations.*

Proof. Suppose there exists $k_0 \geq 0$ such that all iterations $k \geq k_0$ are unsuccessful. Then at least one of the inequalities

$$\rho_k \leq c_1, \quad \text{pred}_k \leq p_{min}(1-\theta)\|d^k\| \min\{\|r(x^k)\|, \|r(x^k)\|^\kappa\} \tag{22}$$

has to hold for all $k \geq k_0$. We will derive a contradiction and show that both inequalities are eventually violated.

First note that Lemma 4(b) implies $x^k = x^{k_0}$ for all $k \geq k_0$ and $\{\mu_k\} \to \infty$ whereas both $\{\|r(x^k)\|\}$ and $\{\|H_k\|\}$ are bounded. Thus, remembering $\|G_k\| = \|H_k\| + \mu_k$, it follows from the first inequality in Lemma 8 that $\{\|d^k\|\}$ is bounded by some $\overline{d} > 0$. For all $k \geq k_0$ it then holds that ξ^k (from Lemma 9) belongs to the compact set $B_{\overline{d}}(x^{k_0}) \cap \Omega$ (recall that Ω was supposed to be a closed set). From the continuity of $\nabla^2 f(\cdot)$ on Ω it then follows that

$$\|\nabla^2 f(\xi^k) - \nabla^2 f(x^k)\| < (1-c_1)\mu_k \tag{23}$$

for sufficiently large $k \geq k_0$, which together with Lemma 9 guarantees

$$\text{ared}_k - c_1 \text{pred}_k > 0,$$

and therefore $\rho_k > c_1$, thus violating the first inequality in (22).

The second inequality in Lemma 8 ensures that $\|d^k\| > 0$ for all $k \geq 0$. Thus, from Lemma 8, we get

$$\frac{\|r(x^k)\|}{\|d^k\|\mu_k} \leq \frac{1 + \|G_k\|}{(1-\theta)\mu_k} \leq \frac{1 + \|H_k\| + \mu_k}{(1-\theta)\mu_k}$$

for all $k \geq k_0$. Taking $k \to \infty$, it follows that the expression on the right-hand side tends to $1/(1-\theta)$. Hence, for $k \geq k_0$ sufficiently large it holds that

$$\frac{\|r(x^k)\|}{\|d^k\|\mu_k} < \frac{1}{2p_{min}(1-\theta)}.$$

This inequality together with (17) then yields

$$\text{pred}_k \geq \frac{\mu_k}{2}\|d^k\|^2 > p_{min}(1-\theta)\|r(x^k)\|\|d^k\| \geq p_{min}(1-\theta)\|d^k\| \min\{\|r(x^k)\|, \|r(x^k)\|^\kappa\} \tag{24}$$

for sufficiently large $k \geq k_0$, which contradicts the second inequality in (22). □

We next present our first global convergence result for Algorithm 1.

Theorem 11. *The sequence* $\{x^k\}$ *generated by Algorithm 1 satisfies* $\liminf_{k\to\infty} \|r(x^k)\| = 0$.

Proof. Let $\mathcal{S} \subset \mathbb{N}$ be the set of successful or highly successful iterations, and recall that this set is infinite due to Theorem 10. Assume, by contradiction, that $\liminf_{k\to\infty} \|r(x^k)\| > 0$. Then there exists $\varepsilon > 0$ such that $\min\{\|r(x^k)\|, \|r(x^k)\|^\kappa\} \geq \varepsilon$ for all $k \geq 0$. Lemma 6 implies that the set $\mathcal{K}$ is finite, hence the set $\overline{\mathcal{S}} := \mathcal{S} \setminus \mathcal{K}$ is still infinite. By definition, it holds for all $k \in \overline{\mathcal{S}}$ that

$$\begin{aligned} F(x^k) - F(\hat{x}^k) = \text{ared}_k > c_1\text{pred}_k &> c_1 p_{min}(1-\theta)\|d^k\| \min\{\|r(x^k)\|, \|r(x^k)\|^\kappa\} \\ &\geq c_1 p_{min}(1-\theta)\|d^k\|\varepsilon, \end{aligned}$$

cf. Lemma 4. Since F is bounded from below, summation yields

$$\infty > \sum_{k=0}^{\infty}[F(x^k) - F(x^{k+1})] \geq \sum_{k\in\overline{\mathcal{S}}}[F(x^k) - F(\hat{x}^k)] \geq c_1 p_{min}(1-\theta)\varepsilon \sum_{k\in\overline{\mathcal{S}}} \|d^k\|$$

(where we used the fact that $F(x^k) - F(x^{k+1}) \geq 0$ for all k). Taking into account that x^k is not updated in unsuccessful steps, it follows that

$$\infty > \sum_{k\in\overline{\mathcal{S}}} \|d^k\| + \sum_{k\in\mathcal{K}} \|d^k\| = \sum_{k\in\mathcal{S}} \|d^k\| = \sum_{k\in\mathcal{S}} \|x^{k+1} - x^k\| = \sum_{k=0}^{\infty} \|x^{k+1} - x^k\|, \tag{25}$$

where we used the previous inequality and the finiteness of $\mathcal{K}$ in the first inequality. Hence, $\{x^k\}$ is a Cauchy sequence and therefore convergent to some $\overline{x} \in \mathbb{R}^n$. The mapping $x \mapsto \nabla^2 f(x) + a[-\lambda_{min}(\nabla^2\psi(Ax-b))]_+ A^\top A$ is continuous, i.e., the sequence $\{H_k\}$ is also convergent. Define $M := \sup\{\|H_k\| \,|\, k \geq 0\} < \infty$. Since $\|r(\cdot)\|$ is continuous, we have $\|r(\overline{x})\| = \lim_{k\to\infty} \|r(x^k)\| \geq \varepsilon$ and $\overline{x}$ is not a stationary point of (1). Using the boundedness of $\{H_k\}$ together with Lemma 8 yields

$$\|r(x^k)\| \leq \frac{1 + M + \mu_k}{1-\theta}\|d^k\|.$$

Note that (25) implies $\|d^k\| \to_{\mathcal{S}} 0$. If there were a subset $\mathcal{S}' \subseteq \mathcal{S}$ such that $\{\mu_k\}_{\mathcal{S}'}$ is bounded, then $\{\|r(x^k)\|\}_{\mathcal{S}'}$ would converge to zero, a contradiction. Hence, $\{\mu_k\} \to_{\mathcal{S}} \infty$. Since μ_k can not decrease during unsuccessful iterations, it follows that $\{\mu_k\} \to \infty$. This implies that Algorithm 1 also performs infinitely many unsuccessful iterations.

For every $k \geq 0$, Taylor's formula yields the existence of a vector ξ^k on the straight line between x^k and $\hat{x}^k$ such that $f(\hat{x}^k) - f(x^k) = \nabla f(\xi^k)^\top d^k$. Note that, similar to the proof of Theorem 10, $\{\|d^k\|\}$ is bounded. Hence, for some $\overline{d} > 0$ and k sufficiently large, ξ^k belongs to the compact set $B_{\overline{d}}(\overline{x}) \cap \Omega$. Note that ∇f is continuously differentiable and therefore also locally Lipschitz continuous, hence Lipschitz continuous on compact sets. In particular, there exists a constant $\overline{L} > 0$ such that

$$\|\nabla f(\xi^k) - \nabla f(x^k)\| \leq \overline{L}\|\xi^k - x^k\| \leq \overline{L}\|d^k\| \tag{26}$$

holds for k sufficiently large. By using (17) in the first, Taylor's formula in the second and (26) in the last inequality, we obtain

$$\begin{aligned}
|\rho_k - 1| = \left|\frac{\mathrm{ared}_k}{\mathrm{pred}_k} - 1\right| &= \left|\frac{\mathrm{ared}_k - \mathrm{pred}_k}{\mathrm{pred}_k}\right| = \left|\frac{F(\hat{x}^k) - q_k(\hat{x}^k)}{\mathrm{pred}_k}\right| \\
&\leq \frac{|f(\hat{x}^k) - f(x^k) - \nabla f(x^k)^\top d^k - \frac{1}{2}(d^k)^\top \nabla^2 f(x^k) d^k|}{\frac{1}{2}\mu_k \|d^k\|^2} \\
&\leq \frac{2\left|\nabla f(\xi^k)^\top d^k - \nabla f(x^k)^\top d^k\right| + \left|(d^k)^\top \nabla^2 f(x^k) d^k\right|}{\mu_k \|d^k\|^2} \\
&\leq \frac{2\|\nabla f(\xi^k) - \nabla f(x^k)\|\|d^k\| + \|\nabla^2 f(x^k)\|\|d^k\|^2}{\mu_k \|d^k\|^2} \\
&\leq \frac{2\overline{L} + \|\nabla^2 f(x^k)\|}{\mu_k} \to 0
\end{aligned}$$

for $k \to \infty$. Hence, $\{\rho_k\} \to 1$, i.e., eventually all steps are highly successful, which yields a contradiction and therefore $\liminf_{k\to\infty} \|r(x^k)\| = 0$. □

The following global convergence theorem is the same as [19, Theorem 5.7]. Its proof is only slightly adapted to our case.

Theorem 12. *Assume that ∇f is uniformly continuous on a set $\mathcal{X}$ satisfying $\{x^k\} \subset \mathcal{X}$. Then $\lim_{k\to\infty} \|r(x^k)\| = 0$ holds. In particular, every accumulation point of $\{x^k\}$ is a stationary point of F.*

Proof. Assume, by contradiction, that there exists $\varepsilon > 0$ and $\mathcal{L} \subset \mathbb{N}$ such that $\|r(x^k)\| \geq 2\varepsilon$ for all $k \in \mathcal{L}$. Set $\overline{\varepsilon} := \min\{\varepsilon, \varepsilon^\kappa\}$. By Theorem 11, for each $k \in \mathcal{L}$, there is an index $l_k > k$ such that $\|r(x^l)\| \geq \varepsilon$ for all $k \leq l < l_k$ and $\|r(x^{l_k})\| < \varepsilon$. If, for $k \in \mathcal{L}$, an iteration $k \leq l < l_k$ is successful or highly successful, we get

$$F(x^l) - F(x^{l+1}) \geq c_1 \mathrm{pred}_l > c_1(1-\theta)p_{min}\|d^l\|\|r(x^l)\| \geq c_1(1-\theta)p_{min}\overline{\varepsilon}\|x^{l+1} - x^l\|.$$

For unsuccessful iterations l, this estimate holds trivially. Thus,

$$\begin{aligned}(1-\theta)p_{min}c_1\overline{\varepsilon}\|x^{l_k} - x^k\| &\leq (1-\theta)p_{min}c_1\varepsilon \sum_{l=k}^{l_k-1} \|x^{l+1} - x^l\| \\ &\leq \sum_{l=k}^{l_k-1} F(x^l) - F(x^{l+1}) = F(x^k) - F(x^{l_k})\end{aligned}$$

holds for all $k \in \mathcal{L}$. By Assumption 1(c), F is bounded from below, and by construction, the sequence $\{F(x^k)\}$ is monotonically decreasing, hence convergent. This implies that the sequence $\{F(x^k) - F(x^{l_k})\}_{\mathcal{L}}$ converges to 0. Hence, we get $\{\|x^{l_k} - x^k\|\}_{\mathcal{L}} \to 0$. The uniform continuity of ∇f and of the proximity operator together with the fact that the composition of uniformly continuous functions is uniformly continuous, yields the uniform continuity of the residual funciton $r(\cdot)$. Thus, we get $\{\|r(x^{l_k}) - r(x^k)\|\}_{\mathcal{L}} \to 0$. On the other hand, by the choice of l_k, we have

$$\|r(x^k) - r(x^{l_k})\| \geq \|r(x^k)\| - \|r(x^{l_k})\| \geq 2\varepsilon - \varepsilon = \varepsilon$$

for all $k \in \mathcal{L}$, which yields the desired contradiction. □

5 Local Superlinear Convergence

The aim of this section is to prove local fast superlinear convergence of Algorithm 1 under the following (fairly mild) assumptions.

Assumption 2. (a) The set X^* of strongly stationary points of (1) is nonempty and there exists an accumulation point $x^* \in X^*$ of $\{x^k\}_{\mathcal{K}}$.

(b) $\nabla^2\psi$ is locally Lipschitz continuous at $Ax^* - b$ relative to $A(\operatorname{dom}\varphi) - b$, i.e., there exists $\varepsilon > 0$ and $L_\psi > 0$ such that

$$\|\nabla^2\psi(Ax - b) - \nabla^2\psi(Ay - b)\| \leq L_\psi\|Ax - Ay\|, \quad \forall x, y \in B_\varepsilon(x^*) \cap \operatorname{dom}\varphi.$$

(c) $\|r(x)\|$ provides a local Hölderian error bound for problem (1) on $B_\varepsilon(x^*) \cap \operatorname{dom}\varphi$, i.e., there exist constants $\beta > 0$ and $q > \max\{\delta, 1-\delta\}$ such that

$$\beta \operatorname{dist}(x, X^*) \leq \|r(x)\|^q, \quad \forall x \in B_\varepsilon(x^*) \cap \operatorname{dom}\varphi, \tag{27}$$

where $\delta > 0$ denotes the constant from Algorithm 1.

Note that Lemma 6 and Theorem 11 ensure that $\mathcal{K}$ is an infinite set. Hence, the subsequence $\{x^k\}_{\mathcal{K}}$ in Assumption 2(a) is well-defined. Define

$$\varepsilon_0 := \min\left\{\varepsilon, \varepsilon/\|A\|\right\} \leq \varepsilon,$$

where $\varepsilon > 0$ denotes the radius from Assumption 2(b). For $x, y \in B_{\varepsilon_0}(x^*) \cap \operatorname{dom}\varphi$, it then follows from (2) and Assumption 2(b)¸ that $\nabla^2 f$ is locally Lipschitz continuous at x^* relative to $\operatorname{dom}\varphi$ with Lipschitz constant $L := \|A\|^3 L_\psi$, i.e.

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L\|x - y\|, \quad \forall x, y \in B_{\varepsilon_0}(x^*). \tag{28}$$

This, in turn, implies that

$$\|\nabla f(x) - \nabla f(y) - \nabla^2 f(x)(x - y)\| \leq \frac{L}{2}\|x - y\|^2, \quad \forall x, y \in B_{\varepsilon_0}(x^*). \tag{29}$$

Furthermore, since f is twice continuously differentiable, ∇f is continuously differentiable and, therefore, locally Lipschitz continuous. Consequently, there exists a constant $L_g > 0$ such that

$$\|\nabla f(x) - \nabla f(y)\| \leq L_g\|x - y\|, \quad \forall x, y \in B_{\varepsilon_0}(x^*). \tag{30}$$

In particular, we therefore have

$$\|\nabla^2 f(x)\| \leq L_g, \quad \forall x \in B_{\varepsilon_0}(x^*). \tag{31}$$

In the following, for each $k \geq 0$, we denote by $\widetilde{x}^k$ a point satisfying the properties

$$\|x^k - \widetilde{x}^k\| = \operatorname{dist}(x^k, X^*), \quad \widetilde{x}^k \in X^*, \tag{32}$$

i.e., $\widetilde{x}^k$ is a (not necessarily unique) projection of x^k onto the nonempty and closed (not necessarily convex) set X^*.

Lemma 13. *Suppose that Assumptions 2 hold. Then, for every iteration $k \geq 0$ with $x^k \in B_{\varepsilon_0/2}(x^*)$, it holds that*

$$\|r(x^k)\| \leq (2 + L_g)\operatorname{dist}(x^k, X^*).$$

Proof. First observe that

$$\|\widetilde{x}^k - x^*\| \leq \|x^k - x^*\| + \|\widetilde{x}^k - x^k\| \leq 2\|x^k - x^*\|, \tag{33}$$

i.e., for $x^k \in B_{\varepsilon_0/2}(x^*)$, it holds that $\widetilde{x}^k \in B_{\varepsilon_0}(x^*)$. Remembering the definition of $\widetilde{x}^k$, we obtain

$$\begin{aligned}
\|r(x^k)\| &= \|r(x^k) - r(\widetilde{x}^k)\| \\
&= \|\operatorname{prox}_\varphi(x^k - \nabla f(x^k)) - x^k - \operatorname{prox}_\varphi(\widetilde{x}^k - \nabla f(\widetilde{x}^k)) + \widetilde{x}^k\| \\
&\leq \|\operatorname{prox}_\varphi(x^k - \nabla f(x^k)) - \operatorname{prox}_\varphi(\widetilde{x}^k - \nabla f(\widetilde{x}^k))\| + \|x^k - \widetilde{x}^k\| \\
&\leq \|x^k - \widetilde{x}^k - \nabla f(x^k) + \nabla f(\widetilde{x}^k)\| + \|x^k - \widetilde{x}^k\| \\
&\leq \|\nabla f(x^k) - \nabla f(\widetilde{x}^k)\| + 2\|x^k - \widetilde{x}^k\| \\
&\leq (2 + L_g)\operatorname{dist}(x^k, X^*),
\end{aligned}$$

where the second inequality follows from the non-expansiveness of the proximity operator and the last inequality follows from (30), taking into account that $x^k, \widetilde{x}^k \in B_{\varepsilon_0}(x^*)$. □

The following lemma is almost identical to [23, Lemma 4.2]. For the convenience of our readers, its proof is provided here, with slight adaptations to our case.

Lemma 14. *For each $k \in \mathcal{K}$, it holds that $\|\hat{x}^k - \overline{x}^k\| \leq \nu_{min}^{-1}\theta(1 + \|G_k\|)\|r(x^k)\|^{1+\tau-\delta}$.*

Proof. Consider a fixed index $k \in \mathcal{K}$. From the definition of $R_k(\hat{x}^k)$ and relation (8), it follows that

$$\begin{aligned} &\hat{x}^k - R_k(\hat{x}^k) \in \hat{x}^k - \nabla f(x^k) - G_k d^k - \partial\varphi(\hat{x}^k - R_k(\hat{x}^k)) \\ \iff &R_k(\hat{x}^k) - \nabla f(x^k) - G_k d^k \in \partial\varphi(\hat{x}^k - R_k(\hat{x}^k)). \end{aligned}$$

Since $\overline{x}^k$ is the exact solution of (12) it holds by Fermat's theorem that

$$-\nabla f(x^k) - G_k(\overline{x}^k - x^k) \in \partial\varphi(\overline{x}^k).$$

By the monotonicity of $\partial\varphi$, we have

$$\langle R_k(\hat{x}^k) - G_k(\hat{x}^k - \overline{x}^k), \hat{x}^k - R_k(\hat{x}^k) - \overline{x}^k\rangle \geq 0.$$

Reordering yields

$$\begin{aligned} \langle \hat{x}^k - \overline{x}^k, G_k(\hat{x}^k - \overline{x}^k)\rangle &\leq \langle R_k(\hat{x}^k), \hat{x}^k - \overline{x}^k - R_k(\hat{x}^k) + G_k(\hat{x}^k - \overline{x}^k)\rangle \\ &\leq \langle R_k(\hat{x}^k), (I + G_k)(\hat{x}^k - \overline{x}^k)\rangle. \end{aligned}$$

Combining this inequality with $G_k \succeq \mu_k I$ and using (13) yields

$$\mu_k\|\hat{x}^k - \overline{x}^k\|^2 \leq (1 + \|G_k\|)\|R_k(\hat{x}^k)\|\|\hat{x}^k - \overline{x}^k\| \leq \theta(1 + \|G_k\|)\|r(x^k)\|^{1+\tau}\|\hat{x}^k - \overline{x}^k\|.$$

Dividing by $\mu_k\|\hat{x}^k - \overline{x}^k\|$ (the case $\|\hat{x}^k - \overline{x}^k\| = 0$ is trivial) and using Lemma 5(b) along with $\nu_k \geq \nu_{min}$ demonstrates that the desired result holds. □

The following lemma is identical to [23, Lemma 4.4]. Again, its proof is presented here, only adapting the notation to our case.

Lemma 15. *Suppose that Assumptions 2 hold. Then for every $k \geq 0$ with $x^k \in B_{\varepsilon_0/2}(x^*)$ it holds that*

$$\Lambda_k \leq aL_\psi\|A\|\operatorname{dist}(x^k, X^*).$$

Proof. Let $x^k \in B_{\varepsilon_0/2}(x^*)$ be fixed. By definition of Λ_k, it suffices to consider the case where $\lambda_{min}(\nabla^2\psi(Ax^k - b)) < 0$. In view of (33), we obtain $\|\widetilde{x}^k - x^*\| \leq \varepsilon_0$, and consequently $\widetilde{x}^k \in B_\varepsilon(x^*) \cap \operatorname{dom}\varphi$. From $\widetilde{x}^k \in X^*$, we have $\nabla^2\psi(A\widetilde{x}^k - b) \succeq 0$. When $\lambda_{min}(\nabla^2\psi(A\widetilde{x}^k - b)) = 0$, then

$$\begin{aligned} \Lambda_k &= -a\lambda_{min}(\nabla^2\psi(Ax^k - b)) = a[\lambda_{min}(\nabla^2\psi(A\widetilde{x}^k - b)) - \lambda_{min}(\nabla^2\psi(Ax^k - b))] \\ &\leq a\|\nabla^2\psi(A\widetilde{x}^k - b) - \nabla^2\psi(Ax^k - b)\| \leq aL_\psi\|A\|\|x^k - \widetilde{x}^k\|, \end{aligned}$$

where the first inequality is by the Lipschitz continuity of the function $\mathbb{S}^n \ni Z \mapsto \lambda_{min}(Z)$ with modulus 1 (follows from Weyl's inequality), and the second one is using Assumption 2(b). So we only need to consider the case $\lambda_{min}(\nabla^2\psi(A\widetilde{x}^k - b)) > 0$. For this purpose, let $\phi_k(t) := \lambda_{min}[\nabla^2\psi(Ax^k - b + tA(\widetilde{x}^k - x^k))]$ for $t \geq 0$. Clearly, ϕ_k is continuous on any open interval containing $[0, 1]$. Note that $\phi_k(0) < 0$ and $\phi_k(1) > 0$. Hence, there exists $\overline{t}_k \in (0, 1)$ such that $\phi_k(\overline{t}_k) = 0$. Consequently,

$$\begin{aligned} \Lambda_k &= -a\lambda_{min}(\nabla^2\psi(Ax^k - b)) \\ &= a[\lambda_{min}(\nabla^2\psi(Ax^k - b + \overline{t}_k A(\widetilde{x}^k - x^k))) - \lambda_{min}(\nabla^2\psi(Ax^k - b))] \\ &\leq a\|\nabla^2\psi(Ax^k - b + \overline{t}_k A(\widetilde{x}^k - x^k)) - \nabla^2\psi(Ax^k - b)\| \leq aL_\psi\|A\|\|\widetilde{x}^k - x^k\|. \end{aligned}$$

This shows that the desired result holds. □

Lemma 16. *Suppose that Assumption 2 holds. Define* $\varepsilon_1 := \min\left\{\frac{1}{2+L_g}, \frac{\varepsilon_0}{2}\right\}$. *Then, for* $k \in \mathcal{K}$ *with* $x^k \in B_{\varepsilon_1}(x^*)$, *it holds that*

$$\|d^k\| \leq c\,\mathrm{dist}(x^k, X^*),$$

where $c := \nu_{min}^{-1}\theta(2+L_g)^{1+\tau-\delta}(1+L_g+aL+\overline{\nu}(2+L_g)^{\delta}) + \frac{L+2aL}{2\nu_{min}\beta} + 2$.

Proof. Let $k \in \mathcal{K}$ and $x^k \in B_{\varepsilon_1}(x^*)$ be fixed. From the definition of $\widetilde{x}^k$ it follows that $0 \in \nabla f(\widetilde{x}^k) + \partial\varphi(\widetilde{x}^k)$ and thus

$$\nabla f(x^k) - \nabla f(\widetilde{x}^k) + (H_k + \mu_k I)(\widetilde{x}^k - x^k) \in \nabla f(x^k) + (H_k + \mu_k I)(\widetilde{x}^k - x^k) + \partial\varphi(\widetilde{x}^k). \quad (34)$$

Together with

$$0 \in \nabla f(x^k) + (H_k + \mu_k I)(\overline{x}^k - x^k) + \partial\varphi(\overline{x}^k) \quad (35)$$

it follows from the strong monotonicity of the mapping $\nabla f(x^k) + (H_k + \mu_k I)(\cdot - x^k) + \partial\varphi(\cdot)$ on $\mathbb{R}^n$ that

$$\left\langle \nabla f(x^k) - \nabla f(\widetilde{x}^k) + (H_k + \mu_k I)(\widetilde{x}^k - x^k),\, \widetilde{x}^k - \overline{x}^k \right\rangle \geq \mu_k \|\widetilde{x}^k - \overline{x}^k\|^2. \quad (36)$$

As in (33) it holds that $\widetilde{x}^k \in B_{\varepsilon_0}(x^*)$ and from Lemma 13 it follows that $\|r(x^k)\| \leq 1$. We now get

$$\begin{aligned}
\|\overline{x}^k - x^k\| &= \|\overline{x}^k - \widetilde{x}^k + \widetilde{x}^k - x^k\| \leq \|\overline{x}^k - \widetilde{x}^k\| + \|\widetilde{x}^k - x^k\| \\
&\leq \frac{1}{\mu_k}\|\nabla f(x^k) - \nabla f(\widetilde{x}^k) + (H_k + \mu_k I)(\widetilde{x}^k - x^k)\| + \|\widetilde{x}^k - x^k\| \\
&\leq \frac{1}{\mu_k}\left(\|\nabla f(x^k) - \nabla f(\widetilde{x}^k) + H_k(\widetilde{x}^k - x^k)\|\right) + 2\|\widetilde{x}^k - x^k\| \\
&\leq \frac{1}{\mu_k}\left(\frac{L}{2}\|\widetilde{x}^k - x^k\|^2 + \Lambda_k\|A^2\|\|\widetilde{x}^k - x^k\|\right) + 2\|\widetilde{x}^k - x^k\| \\
&\leq \frac{L + 2aL_{\psi}\|A\|^3}{2\mu_k}\,\mathrm{dist}(x^k, X^*)^2 + 2\,\mathrm{dist}(x^k, X^*) \\
&= \frac{L + 2aL}{2\nu_k\|r(x^k)\|^{\delta}}\,\mathrm{dist}(x^k, X^*)^2 + 2\,\mathrm{dist}(x^k, X^*) \qquad (37)\\
&\leq \frac{L + 2aL}{2\nu_k\|r(x^k)\|^{q}}\,\mathrm{dist}(x^k, X^*)^2 + 2\,\mathrm{dist}(x^k, X^*) \\
&\leq \frac{L + 2aL}{2\nu_{min}\beta\,\mathrm{dist}(x^k, X^*)}\,\mathrm{dist}(x^k, X^*)^2 + 2\,\mathrm{dist}(x^k, X^*) \\
&= \left(\frac{L + 2aL}{2\nu_{min}\beta} + 2\right)\mathrm{dist}(x^k, X^*),
\end{aligned}$$

where we used (36) together with the Cauchy-Schwarz inequality in the second, the triangle inequality and (29) in the fourth, Lemma 15 and the definition of $\widetilde{x}^k$ in the fifth, $q \geq \delta$ together with $\|r(x^k)\| \leq 1$ in the sixth, and Assumption 2(c) in the seventh inequality. In the second equality we used Lemma 5(b). Since $k \in \mathcal{K}$, it holds that

$$\begin{aligned}
\|G_k\| &\leq \|\nabla^2 f(x^k)\| + \Lambda_k\|A^\top A\| + \mu_k \leq L_g + aL\,\mathrm{dist}(x^k, X^*) + \overline{\nu}\|r(x^k)\|^{\delta} \\
&\leq L_g + aL\,\mathrm{dist}(x^k, X^*) + \overline{\nu}(2 + L_g)^{\delta}\,\mathrm{dist}(x^k, X^*)^{\delta} \\
&\leq L_g + aL + \overline{\nu}(2 + L_g)^{\delta},
\end{aligned}$$

where we used the triangle inequality in the first, (31), Lemma 15 and Lemma 5(e) in the second, Lemma 13 in the third, and $\operatorname{dist}(x^k, X^*) \le 1$ (simply because $x^k \in B_{\varepsilon_1}(x^*)$ and $\varepsilon_1 < 1$) in the last inequality. We now obtain

$$\begin{aligned}
\|d^k\| &= \|\hat{x}^k - \overline{x}^k + \overline{x}^k - x^k\| \le \|\hat{x}^k - \overline{x}^k\| + \|\overline{x}^k - x^k\| \\
&\le \nu_{min}^{-1}\theta(1 + \|G_k\|)\|r(x^k)\|^{1+\tau-\delta} + \|\overline{x}^k - x^k\| \\
&\le \nu_{min}^{-1}\theta(1 + L_g + aL + \overline{\nu}(2 + L_g)^{\delta})(2 + L_g)^{1+\tau-\delta} \operatorname{dist}(x^k, X^*) + \|\overline{x}^k - x^k\| \\
&\le c \operatorname{dist}(x^k, X^*),
\end{aligned}$$

where we used Lemma 14 in the second, Lemma 13, $\operatorname{dist}(x^k, X^*) \le 1$, $\tau \ge \delta$ and the previous inequality in the third, and (37) in the last inequality. □

Lemma 17. *Suppose that Assumption 2 holds. Define* $\varepsilon_2 := \min\left\{\frac{1}{2+L_g}, \frac{1}{aL_\psi\|A\|}, \frac{\varepsilon_0}{1+c}\right\}$, *where $c > 0$ is the constant from Lemma 16. For $k \in \mathcal{K}$ with $x^k \in B_{\varepsilon_2}(x^*)$, it then holds that*

$$\|r(\hat{x}^k)\| \le \hat{c}\|r(x^k)\|^{\min\{\delta+q,1+\tau\}}, \tag{38}$$

$$\operatorname{dist}(\hat{x}^k, X^*) \le \widetilde{c} \operatorname{dist}(x^k, X^*)^{(1+\delta)q}, \tag{39}$$

with constants $\hat{c}$ and $\widetilde{c}$ defined by

$$\begin{aligned}
\hat{c} &:= \frac{c^2L + 2acL_\psi\|A\|^3 + 2\beta c\overline{\nu}}{2\beta^2} + \theta, \\
\widetilde{c} &:= \frac{1}{\beta}\left(\frac{c^2L}{2} + acL_\psi\|A\|^3 + c\overline{\nu}(2 + L_g)^{\delta} + \theta(2 + L_g)^{1+\tau}\right)^q.
\end{aligned}$$

Proof. Using the definition of ε_2 as well as Lemmas 13 and 15, it follows that $\operatorname{dist}(x^k, X^*) \le 1$, $\|r(x^k)\| \le 1$ and $\Lambda_k \le 1$ whenever $x^k \in B_{\varepsilon_2}(x^*)$. Additionally, for $x^k \in B_{\varepsilon_2}(x^*) \subseteq B_{\varepsilon_1}(x^*)$, it follows from Lemma 16 that

$$\|\hat{x}^k - x^*\| \le \|x^k - x^*\| + \|d^k\| \le (1 + c)\|x^k - x^*\| \le \varepsilon_0,$$

i.e., $\hat{x}^k \in B_{\varepsilon_0}(x^*)$. We now get

$$\begin{aligned}
\|r(\hat{x}^k)\| &= \|\operatorname{prox}_\varphi(\hat{x}^k - \nabla f(\hat{x}^k)) - \hat{x}^k\| \\
&= \|\operatorname{prox}_\varphi(\hat{x}^k - \nabla f(\hat{x}^k)) - \operatorname{prox}_\varphi(\hat{x}^k - \nabla f(x^k) - (H_k + \mu_k I)d^k) - R_k(\hat{x}^k)\| \\
&\le \|\operatorname{prox}_\varphi(\hat{x}^k - \nabla f(\hat{x}^k)) - \operatorname{prox}_\varphi(\hat{x}^k - \nabla f(x^k) - (H_k + \mu_k I)d^k)\| + \|R_k(\hat{x}^k)\| \\
&\le \|\nabla f(\hat{x}^k) - \nabla f(x^k) - (H_k + \mu_k I)d^k\| + \|R_k(\hat{x}^k)\| \\
&\le \|\nabla f(\hat{x}^k) - \nabla f(x^k) - \nabla^2 f(x^k)d^k\| + \Lambda_k\|A^\top A d^k\| + \mu_k\|d^k\| + \|R_k(\hat{x}^k)\| \\
&\le \frac{L}{2}\|d^k\|^2 + \Lambda_k\|A\|^2\|d^k\| + \mu_k\|d^k\| + \|R_k(\hat{x}^k)\| \\
&\le \frac{c^2L}{2} \operatorname{dist}(x^k, X^*)^2 + acL_\psi\|A\|^3 \operatorname{dist}(x^k, X^*)^2 + c\mu_k \operatorname{dist}(x^k, X^*) + \|R_k(\hat{x}^k)\| \\
&\le \frac{c^2L}{2\beta^2}\|r(x^k)\|^{2q} + \frac{acL_\psi\|A\|^3}{\beta^2}\|r(x^k)\|^{2q} + \frac{c\overline{\nu}}{\beta}\|r(x^k)\|^{\delta+q} + \theta\|r(x^k)\|^{1+\tau} \\
&\le \left(\frac{c^2L + 2acL_\psi\|A\|^3 + 2\beta c\overline{\nu}}{2\beta^2} + \theta\right)\|r(x^k)\|^{\min\{\delta+q,1+\tau\}},
\end{aligned}$$

where we used the nonexpansiveness in the second, (29) and $\Lambda_k \le 1$ in the fourth, Lemma 15 and Lemma 16 in the fifth, Assumption 2(c), Lemma 5(e) and the inexactness criterion (13) in the sixth, and $q \geq \delta$ together with $\|r(x^k)\| \leq 1$ in the last inequality. Reusing the fifth inequality from above we also get

$$\begin{aligned}\|r(\hat{x}^k)\| &\leq \frac{c^2 L}{2} \operatorname{dist}(x^k, X^*)^2 + acL_\psi \|A\|^3 \operatorname{dist}(x^k, X^*)^2 + c\mu_k \operatorname{dist}(x^k, X^*) + \|R_k(\hat{x}^k)\| \\ &\leq \left(\frac{c^2 L}{2} + acL_\psi \|A\|^3 + c\overline{\nu}(2 + L_g)^\delta\right) \operatorname{dist}(x^k, X^*)^{1+\delta} + \theta \|r(x^k)\|^{1+\tau} \\ &\leq \left(\frac{c^2 L}{2} + acL_\psi \|A\|^3 + c\overline{\nu}(2 + L_g)^\delta + \theta(2 + L_g)^{1+\tau}\right) \operatorname{dist}(x^k, X^*)^{1+\delta},\end{aligned}$$

where we used Lemma 5(e), Lemma 13, $\operatorname{dist}(x^k, X^*) \leq 1$ and the inexactness criterion (13) in the second, and Lemma 13 as well as $\tau \geq \delta$ in the third inequality. From Assumption 2(c) and the previous inequality, we then obtain

$$\operatorname{dist}(\hat{x}^k, X^*) \leq \frac{1}{\beta} \|r(\hat{x}^k)\|^q \leq \widetilde{c} \operatorname{dist}(x^k, X^*)^{(1+\delta)q},$$

and this completes the proof. □

We finally present the main local rate-of-convergence result.

Theorem 18. *Suppose that Assumption 2 holds. Then $\{x^k\}$ converges to x^* and $\{\|r(x^k)\|\}$ converges to 0 at the rate of $\rho := \min\{1 + \tau, \delta + q\} > 1$.*

Proof. We define the constants

$$\varepsilon_3 := \frac{1}{2 + L_g} \left(\frac{\eta}{\hat{c}}\right)^{\frac{1}{\rho - 1}}, \ \varepsilon_4 := \left(\frac{(1 - c_1)\nu_{min}\beta}{cL(2 + L_g)^{q-\delta}}\right)^{\frac{1}{q-\delta}}$$

$$\varepsilon_5 := \frac{1}{2 + L_g} \left(\frac{1 + L_g + \|A\|^2 + \overline{\nu}}{\nu_{min}}\right)^{-\frac{1}{\kappa - 1 - \delta}}$$

$$\varepsilon_6 := \min\{\varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5\}, \ \varepsilon_7 := \left(\varepsilon_6 / \left(1 + \frac{c(2 + L_g)^q}{\beta(1 - \eta^q)}\right)\right)^{\frac{1}{\min(1,q)}}.$$

Assumption 2(a) ensures the existence of a subset $\mathcal{L} \subset \mathcal{K}$ with $\{x^k\}_\mathcal{L} \to x^*$. Consider some $k_0 \in \mathcal{L}$ with $x^{k_0} \in B_{\varepsilon_7}(x^*) \subset B_{\varepsilon_6}(x^*)$. We want to show that for all $k \geq k_0$, it holds that

$$k \in \mathcal{K}, \tag{40a}$$

$$x^k \in B_{\varepsilon_6}(x^*). \tag{40b}$$

For k_0 the above properties hold. Suppose now that (40) is satisfied for $k_0, ..., k$ with some $k \geq k_0$. Using Lemma 9, we then get

$$\begin{aligned}\operatorname{ared}_k - c_3 \operatorname{pred}_k &\geq \frac{1}{2} \left((1 - c_3)\mu_k - \|\nabla^2 f(\xi^k) - \nabla^2 f(x^k)\|\right) \|d^k\|^2 \\ &\geq \frac{1}{2} \left((1 - c_3)\nu_{min} \|r(x^k)\|^{\delta - q} \|r(x^k)\|^q - cL \operatorname{dist}(x^k, X^*)\right) \|d^k\|^2 \\ &\geq \frac{1}{2} \left((1 - c_3)\nu_{min}\beta \|r(x^k)\|^{\delta - q} - cL\right) \operatorname{dist}(x^k, X^*) \|d^k\|^2 \\ &\geq \frac{1}{2} \left((1 - c_3)\nu_{min}\beta(2 + L_g)^{\delta - q} \varepsilon_6^{\delta - q} - cL\right) \operatorname{dist}(x^k, X^*) \|d^k\|^2 \\ &\geq 0\end{aligned}$$

for some $c_3 \in (c_1, 1)$, where the second inequality follows from Lemma 5(b), (28) and Lemma 16, the third from Assumption 2(c), the fourth from Lemma 13 and (40b) together with $q > \delta$ and the fifth from the definition of $\varepsilon_6 \le \varepsilon_4$. It follows that $\rho_k > c_1$. We also get

$$\begin{aligned}\frac{\|r(x^k)\|^\kappa}{\mu_k\|d^k\|} &= \|r(x^k)\|^{\kappa-1}\frac{\|r(x^k)\|}{\mu_k\|d^k\|} \le \|r(x^k)\|^{\kappa-1}\frac{1+\|G_k\|}{(1-\theta)\mu_k} \le \|r(x^k)\|^{\kappa-1-\delta}\frac{1+\|H_k\|+\mu_k}{(1-\theta)\nu_{min}}\\ &\le (2+L_g)^{\kappa-1-\delta}\frac{1+L_g+\|A\|^2+\overline{\nu}}{(1-\theta)\nu_{min}}\varepsilon_6^{\kappa-1-\delta}\\ &\le \frac{1}{1-\theta},\end{aligned}$$

where the first inequality follows from Lemma 8, the second from Lemma 5(b) and $\nu_k \ge \nu_{min}$, the third from Lemma 13, (31), Lemma 5(e), $\Lambda_k \le 1$ and $\|r(x^k)\| \le 1$, and the fourth from the definition of $\varepsilon_6 \le \varepsilon_5$. Together with (17) it then follows that

$$\text{pred}_k \ge \frac{\mu_k}{2}\|d^k\|^2 \ge \frac{1-\theta}{2}\|d^k\|\|r(x^k)\|^\kappa > p_{min}(1-\theta)\|d^k\|\|r(x^k)\|^\kappa.$$

Therefore, iteration k is successful or highly successful. Furthermore it holds that

$$\begin{aligned}\|r(x^{k+1})\| = \|r(\hat{x}^k)\| &\le \hat{c}\|r(x^k)\|^\rho = \hat{c}\|r(x^k)\|^{\rho-1}\|r(x^k)\|\\ &\le \hat{c}\left((2+L_g)\operatorname{dist}(x^k, X^*)\right)^{\rho-1}\|r(x^k)\|\\ &\le \hat{c}(2+L_g)^{\rho-1}\left(\frac{1}{2+L_g}\left(\frac{\eta}{\hat{c}}\right)^{\frac{1}{\rho-1}}\right)^{\rho-1}\|r(x^k)\| = \eta\|r(x^k)\| = \eta\overline{r}_k,\end{aligned}$$

where we used (38) in the first, Lemma 13 and $\rho > 1$ in the second, and the definition of $\varepsilon_6 \le \varepsilon_3$ in the third inequality. In the last equality we used Lemma 5(b). It follows that $k+1 \in \mathcal{K}$. For all $j = k_0, ..., k+1$ it holds that $j \in \mathcal{K}$ and thus

$$\|r(x^j)\| \le \eta\overline{r}_{j-1} = \eta\|r(x^{j-1})\| \le ... \le \eta^{j-k_0}\overline{r}_{k_0} = \eta^{j-k_0}\|r(x^{k_0})\|, \tag{41}$$

by using Lemma 5(a) and Lemma 5(b) repeatedly. Moreover, it holds that $x^j = \hat{x}^{j-1} = x^{j-1} + d^{j-1}$ as all iterations $k_0, ..., k$ are successful or highly successful by definition of $\mathcal{K}$. Thus we get

$$\begin{aligned}\|x^{k+1} - x^{k_0}\| &= \sum_{j=k_0}^{k}\|d^j\| \le c\sum_{j=k_0}^{k}\operatorname{dist}(x^j, X^*) \le \frac{c}{\beta}\sum_{j=k_0}^{k}\|r(x^j)\|^q \le \frac{c}{\beta}\|r(x^{k_0})\|^q\sum_{j=k_0}^{k}(\eta^q)^{j-k_0}\\ &\le \frac{c}{\beta}\|r(x^{k_0})\|^q\sum_{j=0}^{\infty}(\eta^q)^j = \frac{c}{\beta(1-\eta^q)}\|r(x^{k_0})\|^q \le \frac{c(2+L_g)^q}{\beta(1-\eta^q)}\|x^{k_0} - x^*\|^q,\end{aligned} \tag{42}$$

where we used Lemma 16 in the first, Assumption 2(c) in the second, (41) in the third and Lemma 13 in the last inequality. This implies

$$\begin{aligned}\|x^{k+1} - x^*\| \le \|x^{k+1} - x^{k_0}\| + \|x^{k_0} - x^*\| &\le \frac{c(2+L_g)^q}{\beta(1-\eta^q)}{\varepsilon_7}^q + \varepsilon_7\\ &\le \left(\frac{c(2+L_g)^q}{\beta(1-\eta^q)} + 1\right)\varepsilon_7^{\min(1,q)} = \varepsilon_6.\end{aligned}$$

By induction, it follows that (40) holds for all $k \geq k_0$. For an iteration $k \geq k_0$ define l_k as the iteration which satisfies the following three properties:

$$l_k \leq k, \, l_k \in \mathcal{L} \text{ and } j \notin \mathcal{L} \text{ for } l_k < j \leq k. \tag{43}$$

In words, l_k is the last iteration belonging to $\mathcal{L}$ before iteration k. By construction it follows that $l_k \to \infty$ if $k \to \infty$ and therefore $\{x^{l_k}\} \to x^*$. Similar to (42) it follows for $k \geq l_k \geq k_0$ that

$$\|x^k - x^*\| \leq \|x^k - x^{l_k}\| + \|x^{l_k} - x^*\| \leq \frac{c(2+L_g)^q}{\beta(1-\eta^q)}\|x^{l_k} - x^*\|^q + \|x^{l_k} - x^*\|.$$

Hence, $\{x^k\}$ converges to x^*. Now it immediately follows from (38) that $\{\|r(x^k)\|\}$ converges to 0 at the rate of $\rho > 1$. □

Corollary 19. *Suppose that Assumption 2 holds with $q > \frac{1}{1+\delta}$. Then $\{x^k\}$ converges to x^*, $\{\|r(x^k)\|\}$ converges to 0 at the rate of $\rho > 1$ and $\{\text{dist}(x^k, X^*)\}$ converges to 0 at the rate of $(1+\delta)q > 1$.*

Proof. It holds that $\frac{1}{1+\delta} > \frac{1-\delta^2}{1+\delta} = \frac{(1+\delta)(1-\delta)}{1+\delta} = 1 - \delta$, i.e. the assumption here is stronger than in Assumption 2(c). The result follows directly from Theorem 18 and (39). □

6 Numerical Results

In this section, we present the numerical results of Algorithm 1 (denoted as IRPNM-reg) for various instances of Problem 1. We compare these results with the outcomes of the inexact regularized proximal Newton method using line-search (IRPNM-ls) proposed in [23], as well as a modern FISTA-type method (AC-FISTA) from [22].

We start by considering the convex logistic regression problem with l_1-regularizer (Section 6.1) and group regularizer (Section 6.2). Subsequently, we investigate three non-convex problem classes introduced in [23]: l_1-regularized Student's t-regression (Section 6.3), Group regularized Student's t-regression (Section 6.4), and Restoration of a blurred image (Section 6.5).

For all tests, we fix the parameters for IRPNM-reg as follows: $c_1 = 10^{-4}$, $c_2 = 0.9$, $\sigma_1 = 0.5$, $\sigma_2 = 4$, $\eta = 0.9999$, $\theta = 0.9999$, $\alpha = 0.99$, $a = 1$, $\nu_{\min} = 10^{-8}$, $\nu_0 = \min\left(\frac{10^{-2}}{\max(1, \|r(x^0)\|)}, 10^{-4}\right)$, $\overline{\nu} = 100$, $\delta = 0.45$, $\tau \geq \delta$, $p_{\min} = 10^{-8}$, and $\kappa = 2$. For IRPNM-ls and AC-FISTA, we adopt the recommended parameters from their respective papers.

The tests are conducted using Matlab R2022b on a 64-bit Linux system with an Intel(R) Core(TM) i5-3470 CPU @ 3.20GHz and 16 GB RAM.

Since IRPNM-reg solves exactly the same subproblems as IRPNM-ls, we employ the efficient strategy developed in [23]. This strategy solves the dual of an equivalent reformulation of (12) using an augmented Lagrangian method. The semismooth system of equations arising from the augmented Lagrangian method is solved using the semismooth Newton method. Notably, this strategy is tailored to address problems where ψ is a separable function, a characteristic shared by many applications, including those under consideration here. For more comprehensive details on the subproblem solver, please refer to [23, Section 5.1].

We terminate each of the tested methods once the current iterate x^k satisfies $\|r(x^k)\| \leq$ `tol`. Here, `tol` is chosen independently for each problem class and further distinguished between the two second order methods and AC-FISTA.

6.1 l_1-regularized Logistic Regression

First we explore the logistic regression problem defined as

$$\min_{y,v} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 + \exp\left(-b_i(a_i^\top y + v)\right)\right) + \lambda\|y\|_1. \tag{44}$$

In this context, $a_i \in \mathbb{R}^n$ denotes feature vectors, $b_i \in \{-1, 1\}$ represents corresponding labels for $i = 1, ..., m$, and we have $\lambda > 0$, $y \in \mathbb{R}^n$, and $v \in \mathbb{R}$. In standard instances of this problem, $m \gg n$. The logistic regression problem aligns with the general form of (1), where $\psi\colon \mathbb{R}^{n+1} \to \mathbb{R}$ is defined as

$$\psi(u) := \frac{1}{m} \sum_{i=0}^{m} \log(1 + \exp(-u_i)), \quad u := (y^\top, v)^\top,$$

where the i-th row of the matrix $A \in \mathbb{R}^{m\times(n+1)}$ takes the form of $(b_i a_i^\top, b_i)$, and $b = 0 \in \mathbb{R}^m$. The regularization function $\varphi\colon \mathbb{R}^{n+1} \to \mathbb{R}$ is given by $\varphi(u) := \lambda\|y\|_1$.

Following the methodology outlined in [5] and , we create test problems using $n = 10^4$ feature vectors and $m = 10^6$ training sets. Each a_i has approximately $s \in \{10, 100\}$ nonzero entries, independently sampled from a standard normal distribution. We choose $y^{\text{true}} \in \mathbb{R}^n$ with $10s$ non-zero entries and $v^{\text{true}} \in \mathbb{R}$, independently sampled from a standard normal distribution. Labels b_i are determined by

$$b_i = \text{sign}\left(a_i^\top y^{\text{true}} + v^{\text{true}} + v_i\right),$$

where $v_i \in \mathbb{R}$, $i = 1, ..., m$, are generated independently from a normal distribution with variance 0.1. Similar to [16], the regularization parameter λ takes the form $c_\lambda \lambda_{\max}$, with $c_\lambda \in \{1, 0.1, 0.01\}$, and

$$\lambda_{\max} = \frac{1}{m} \left\| \frac{m_-}{m} \sum_{i\colon b_i=1} a_i + \frac{m_+}{m} \sum_{i\colon b_i=-1} a_i \right\|,$$

representing the smallest value such that $y^* = (0, v^*)$ is a solution of (44). Here, m_+ and m_- represent the counts of indices where b_i is equal to $+1$ or -1, respectively. The selection of this value is motivated in [18]. For each method, we select the starting point as $x^0 = 0$ and carry out 10 independent trials - that is, with ten sets of randomly generated data - for every combination of parameters s and c_λ. Tables 1 and 2 present the averaged number of (outer) iterations, objective values, residuals and running times for the two second order methods and AC-FISTA, respectively.

		IRPNM-reg				IRPNM-ls			
c_λ	s	iter	$F(x)$	$\|r(x)\|$	time	iter	$F(x)$	$\|r(x)\|$	time
1	10	63.0	0.0904	7.72e-06	39.8	63.0	0.0904	7.73e-06	34.8
	100	4.4	0.4518	3.48e-06	18.1	4.4	0.4518	3.49e-06	17.8
0.1	10	49.6	0.0785	9.98e-06	149.2	32.0	0.0785	9.99e-06	116.5
	100	7.9	0.2434	9.62e-06	252.6	8.2	0.2434	9.63e-06	262.2
0.01	10	117.3	0.0727	1.00e-05	227.4	87.3	0.0727	1.00e-05	193.2
	100	13.6	0.0844	9.92e-06	734.7	16.6	0.0844	9.98e-06	1038.2

Table 1: Averaged results of IRPNM-reg and IRPNM-ls for 10 independent trials with tolerance `tol` $= 10^{-5}$.

We observe that IRPNM-reg and IRPNM-ls produce identical objective values. Both methods exhibit improved performance for larger values of c_λ. Additionally, the algorithms perform better with sparser data ($s = 10$) for $c_\lambda \in \{0.1, 0.01\}$, but worse for $c_\lambda = 1$. The performance of the methods is comparable, with IRPNM-ls demonstrating slightly superior results in the case of $s = 10$, while IRPNM-reg performs better when $s = 100$ and $c_\lambda = 0.01$.

		AC-FISTA			
c_λ	s	iter	$F(x)$	$\|r(x)\|$	time
1	10	21.2	0.0904	6.36e-06	13.4
	100	10.0	0.4518	6.14e-06	50.9
0.1	10	210.2	0.0784	9.73e-06	129.5
	100	100.9	0.2434	8.02e-06	449.0
0.01	10	272.4	0.0727	9.84e-06	162.9
	100	179.8	0.0844	8.80e-06	759.7

Table 2: Averaged results of AC-FISTA for 10 independent trials with tolerance `tol` $= 10^{-5}$.

AC-FISTA produces nearly identical objective values as the second-order methods. It generally outperforms the second-order methods for $s = 10$ but performs worse for $s = 100$, with some exceptions. Notably, in the case of $s = 10$ and $c_\lambda = 0.1$, IRPNM-ls is slightly faster than AC-FISTA. Conversely, for $s = 100$ and $c_\lambda = 0.01$, AC-FISTA significantly outperforms IRPNM-ls, nearly matching the runtime of IRPNM-reg.

6.2 Group regularized Logistic Regression

We consider the group regularized logistic regression problem, given by

$$\min_{y,v} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 + \exp\left(-b_i(a_i^\top y + v)\right)\right) + \lambda \sum_{i=1}^{l} \|x_{J_i}\|_2,$$

where the data $a_i \in \mathbb{R}^n$, $b_i \in \{-1, 1\}$ for $i = 1, ..., m$ and $v \in \mathbb{R}$ follows the same generation process as in section 6.2 (with $s = 10$). The index sets $J_1, ..., J_l$ form a partition of $\{1, ..., n\}$, i.e. they satisfy $J_i \cap J_j = \emptyset$ for $i \neq j$ and $\cup_{i=1}^{l} J_i = \{1, ..., n\}$. We organize the $n = 10^4$ in two different configurations: $l = 1000$ groups of 100 variables and $l = 100$ groups of 1000 variables, while consistently preserving a sequential group structure. The regularization parameter λ mirrors the one in section 6.2 with $c_\lambda \in \{1, 0.1, 0.01\}$, and the initial value is set as $x^0 = 0$. Similar to the previous test problem, we conduct 10 independent trials for each value of c_λ. Tables 3 and 4 present the averaged number of (outer) iterations, objective values, residuals and running times for the two second order methods and AC-FISTA, respectively.

		IRPNM-reg				IRPNM-ls			
l	c_λ	iter	$F(x)$	$\|r(x)\|$	time	iter	$F(x)$	$\|r(x)\|$	time
1000	1	7.3	0.3049	8.22e-06	14.7	7.4	0.3049	8.57e-06	15.3
	0.1	11.6	0.2725	9.93e-06	98.8	9.6	0.2725	9.96e-06	79.1
	0.01	23.2	0.2574	9.98e-06	184.8	22.2	0.2574	1.00e-05	177.0
100	1	13.8	0.3039	9.74e-06	63.7	7.1	0.3039	9.80e-06	36.8
	0.1	10.0	0.2690	9.92e-06	98.6	9.9	0.2690	9.98e-06	92.1
	0.01	25.0	0.2560	9.99e-06	255.9	24.3	0.2560	1.00e-05	194.7

Table 3: Averaged results of IRPNM-reg and IRPNM-ls for 10 independent trials with tolerance $\texttt{tol} = 10^{-5}$.

Both methods yield the same objective values in essentially the same run times. IRPNM-ls is slightly faster than IRPNM-reg across all test instances.

		AC-FISTA			
l	c_λ	iter	$F(x)$	$\|r(x)\|$	time
1000	1	44.3	0.3049	8.23e-05	28.1
	0.1	134.3	0.2726	9.23e-05	81.4
	0.01	209.8	0.2582	9.84e-05	122.2
100	1	151.5	0.3039	9.44e-06	92.6
	0.1	307.9	0.2690	9.76e-06	251.8
	0.01	607.6	0.2560	9.90e-06	343.9

Table 4: Averaged results of AC-FISTA for 10 independent trials with tolerance $\texttt{tol} = 10^{-5}$.

AC-FISTA achieves the same objective values as the second-order methods. When $l = 100$, AC-FISTA underperforms compared to the second-order methods. For $l = 1000$, AC-FISTA exhibits inferior performance for large c_λ values but superior performance for smaller c_λ values.

6.3 l_1-regularized Student's t-regression

We consider the Student's t-regression problem with l_1-regularizer, given by

$$\min_x \sum_{i=1}^m \log(1 + (Ax - b)_i/\nu) + \lambda\|x\|_1,$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $\nu > 0$ and $\lambda > 0$. The test examples are randomly generated following the same procedure as in [2, 23, 29]. The matrix A is formed by taking $m = n/8$ random cosine measurements, i.e. $Ax = (\texttt{dct}(x))_J$, where $\texttt{dct}$ denotes the discrete cosine transform, and $J \subseteq \{1, ..., n\}$ is an index set selected at random with $|J| = m$. A true sparse signal x^{true} of length $n = 512^2$ is created, featuring $s = \lfloor \frac{n}{40} \rfloor$ randomly selected non-zero entries, calculated as $x_i^{\text{true}} = \eta_1(i)10^{\frac{d\eta_2(i)}{20}}$, where $\eta_1(i) \in \{-1, 1\}$ denotes a random sign and $\eta_2(i)$ is uniformly distributed in the interval $[0, 1]$. The signal possesses a dynamic range of d dB with $d \in \{20, 40, 60, 80\}$. The vector b is then obtained by summing Ax^{true} and Student's t-noise with a degree of freedom of 4, rescaled by 0.1.
The regularization parameter is expressed as $\lambda = c_\lambda\|\nabla f(0)\|_\infty$, where $c_\lambda \in \{0.1, 0.01\}$. For each combination of values d and c_λ we run the three solvers with $\nu = 0.25$ and

$x^{init} = A^\top b$ over 10 independent trials. Tables 5 and 6 present the averaged number of (outer) iterations, objective values, residuals and running times for IRPNM-reg and IRPNM-ls with `tol` $= 10^{-5}$, and AC-FISTA with `tol` $= 10^{-4}$, respectively.

		IRPNM-reg				IRPNM-ls			
c_λ	d	iter	$F(x)$	$\lVert r(x)\rVert$	time	iter	$F(x)$	$\lVert r(x)\rVert$	time
0.1	20	28.4	9532.5413	8.78e-06	13.5	24.2	9532.5413	8.92e-06	13.1
	40	19.5	23812.8786	6.00e-06	32.0	17.2	23812.8749	6.75e-06	33.3
	60	24.7	54228.0069	8.07e-06	88.1	23.8	54228.0069	6.85e-06	84.2
	80	80.3	134779.2596	8.54e-06	281.5	109.7	134779.2596	8.03e-06	323.2
0.01	20	11.8	1020.4271	7.09e-06	37.2	8.9	1020.4271	7.08e-06	37.0
	40	15.5	2395.0693	7.90e-06	129.4	14.1	2395.0693	7.63e-06	122.4
	60	12.4	5424.4039	7.33e-06	170.5	17.7	5424.4039	7.65e-06	261.9
	80	16.3	13478.1029	6.17e-06	314.2	116.3	13478.1029	7.50e-06	1103.9

Table 5: Averaged results of IRPNM-reg and IRPNM-ls for 10 independent trials with tolerance `tol` $= 10^{-5}$.

Both methods yield the same objective values except for the case $c_\lambda = 0.1$, $d = 40$, where IRPNM-ls yields (in average) a slightly smaller objective value. In most cases, the runtimes for the two methods are comparable. However, for $c_\lambda = 0.01$ and $d \in \{60, 80\}$ IRPNM-reg performs better, requiring only a third of the runtime of IRPNM-ls for $d = 80$.

		AC-FISTA			
c_λ	d	iter	$F(x)$	$\lVert r(x)\rVert$	time
0.1	20	507.5	9532.5413	9.71e-05	31.4
	40	1041.5	23812.8749	9.84e-05	95.2
	60	2238.9	54228.0069	9.90e-05	134.6
	80	7240.7.5	134779.2596	9.95e-05	434.2
0.01	20	1488.8	1020.4271	9.97e-05	98.0
	40	2531.7	2395.0693	9.98e-05	162.8
	60	5391.4	5424.4039	9.94e-05	327.9
	80	20694.0	13478.1029	9.93e-05	1243.0

Table 6: Averaged results of AC-FISTA for 10 independent trials with tolerance `tol` $= 10^{-4}$.

Note that here we chose `tol` $= 10^{-4}$ for AC-FISTA instead of 10^{-5}. It solves all the problems and returns the same objective values. In all cases it takes longer to solve the problems with tolerance 10^{-5} than both second order methods with tolerance 10^{-6}. However, in some cases (e.g. $c_\lambda = 0.01$ and $d = 80$), it does not perform much worse than IRPNM-ls.

6.4 Group penalized Student's t-regression

We consider the Student's t-regression problem with group regularizer, given by

$$\min_x \sum_{i=1}^{m} \log(1 + (Ax - b)_i/\nu) + \lambda \sum_{i=1}^{l} \|x_{J_i}\|_2.$$

This test problem is taken from [23, Section 5.3]. A true group sparse signal $x^{true} \in \mathbb{R}^n$ of length $n = 512^2$ with s nonzero groups is generated, whose indices are chosen randomly. Each nonzero entry of x^{true} is calculated using the same formula as in section 6.3. The matrix $A \in \mathbb{R}^{m\times n}$ and the vector $b \in \mathbb{R}^m$ are also obtained in the same way as in section 6.3, with the only difference being the choice of degree of freedom 5 for the Student's t-noise.
The regularization parameter is set as $\lambda = 0.1\|\nabla f(0)\|$. For each combination of values $d \in \{60, 80\}$ dB and non-zero groups $s = \{16, 64, 128\}$ we run the three solvers with $\nu = 0.2$ and $x^{init} = A^\top b$ over 10 independent trials. Tables 7 and 8 present the averaged number of (outer) iterations, objective values, residuals and running times for IRPNM-reg and IRPNM-ls with `tol` $= 10^{-5}$, and AC-FISTA with `tol` $= 10^{-3}$, respectively.

		IRPNM-reg				IRPNM-ls			
d	s	iter	$F(x)$	$\|r(x)\|$	time	iter	$F(x)$	$\|r(x)\|$	time
60	16	6.1	12711.8673	6.54e-06	16.79	9.0	12711.8673	8.44e-06	16.69
	64	6.7	17852.9902	8.65e-06	19.53	12.0	17852.9902	8.08e-06	26.06
	128	7.0	21670.1861	8.64e-06	20.16	14.9	21670.1861	9.13e-06	34.48
80	16	9.0	37037.7136	9.26e-06	38.30	54.8	37037.7137	9.67e-06	133.25
	64	11.0	52741.5880	7.40e-06	49.86	91.7	52741.5881	9.77e-06	245.62
	128	13.3	63451.7421	7.07e-06	61.90	128.2	63451.7421	9.40e-06	372.52

Table 7: Averaged results of IRPNM-reg and IRPNM-ls for 10 independent trials with tolerance `tol` $= 10^{-5}$.

Both methods produce - essentially - the same objective values. IRPNM-reg shows better performance than IRPNM-ls for $d = 60$ and significantly better for $d = 80$.

		AC-FISTA			
d	s	iter	$F(x)$	$\|r(x)\|$	time
60	16	4204.3	12711.8731	9.91e-04	318.76
	64	6282.8	17853.0157	1.00e-03	438.99
	128	8936.6	21670.2322	1.00e-03	592.16
80	16	20954.6	37037.9708	9.97e-04	1493.72
	64	30273.6	52742.3811	9.93e-04	1926.70
	128	31849.3	63452.8920	9.91e-04	1925.60

Table 8: Averaged results of AC-FISTA for 10 independent trials with tolerance `tol` $= 10^{-3}$.

In this example, we had to select `tol` $= 10^{-3}$ for AC-FISTA. It is evident that this reduced accuracy leads to higher average objective values. For this problem class, AC-FISTA is clearly outperformed by both second order methods.

6.5 Nonconvex Image Restoration

In this section we apply the algorithms to image restoration using real-world data. The problem is the same as in [19, 23]. The goal is to find an approximation $x \in \mathbb{R}^n$ of the original image $x^{true} \in \mathbb{R}^n$ from a noisy blurred image $b \in \mathbb{R}^n$ and a blur operator $A \in \mathbb{R}^{n\times n}$, i.e., we seek x with $Ax \approx b$. The objective function incorporates a regularization term $\lambda\|Bx\|_1$ to ensure smooth gradations and antialiasing in the final image, where $B\colon \mathbb{R}^n \to \mathbb{R}^n$ is a two-dimensional discrete Haar wavelet transform. The problem can be expressed as

$$\min_x \sum_{i=1}^m \log(1 + (Ax - b)_i) + \lambda\|Bx\|_1,$$

with $\lambda > 0$. Making use of the orthogonality of B, the problem can be reformulated equivalently as

$$\min_y \sum_{i=1}^m \log(1 + (AB^\top y - b)_i) + \lambda\|y\|_1,$$

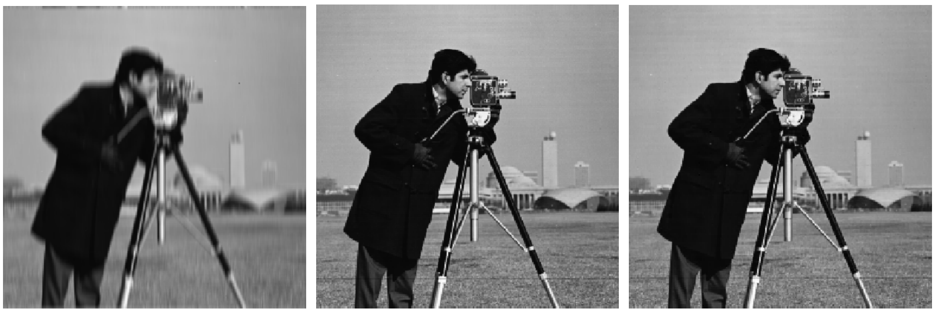
which clearly is an instance of the problem class considered in section 6.3.

The test setup being identical to [19, 23], we select the 256 × 256 grayscale image `cameraman.tif` as the test image $x^{true} \in \mathbb{R}^n$ with $n = 256^2$. The blur operator A is a 9×9 Gaussian filter with a standard deviation of 4, and B is a two-dimensional discrete Haar wavelet of level 4. The noisy image b is created by applying A to the original cameraman test image x^{true} and adding Student's t-noise with degree of freedom 1 and rescaled by 10^{-3}. For each $\lambda \in \{10^{-2}, 10^{-3}, 10^{-4}\}$, we run the three solvers with $y^{init} = Bb$ and $\texttt{tol} = 10^{-5}$ for 10 independent trials. Here we decided to use $\nu_{\min} = 10^{-4}$ instead of 10^{-8}. The reason for this change is that in this test scenario, instances where the subproblem couldn't be solved within the desired maximum number of iterations were much more frequent. Consequently, a significantly higher number of unsuccessful iterations occurred. It is noteworthy that these unsuccessful iterations tend to negatively affect IRPNM-reg more than IRPNM-ls. This is because the line search enables the algorithm to still make some progress, whereas IRPNM-reg simply repeats solving the same subproblem with a larger regularization parameter. Given that subproblems become more challenging to solve with smaller regularization parameters, selecting $\nu_{\min} = 10^{-4}$ instead of 10^{-8} notably reduced the number of unsuccessful iterations and consequently enhanced the performance of IRPNM-reg. Additionally, for this particular example, we experimented with a hybrid approach, IRPNM-reg-ls, which combines both methods. In IRPNM-reg-ls, a line search is conducted whenever an unsuccessful iteration occurs. Table 9 presents the averaged number of (outer) iterations, objective values, residuals and running times for the three second order methods and AC-FISTA.

	IRPNM-reg				IRPNM-ls			
λ	iter	$F(x)$	$\|r(x)\|$	time	iter	$F(x)$	$\|r(x)\|$	time
1e-2	97.4	11245.2731	9.84e-05	200.15	99.3	11245.2731	9.77e-05	201.37
1e-3	115.1	1199.4475	9.38e-05	465.71	113.1	1199.4475	8.77e-05	427.08
1e-4	122.7	146.9925	9.10e-05	709.44	121.3	146.9927	9.55e-05	667.09
	IRPNM-reg-ls				AC-FISTA			
1e-2	97.4	11245.2731	9.84e-05	199.33	2086.7	11245.2731	9.88e-05	279.63
1e-3	113.3	1199.4475	9.66e-05	445.34	3486.9	1199.3795	9.86e-05	494.03
1e-4	118.7	146.9926	9.20e-05	647.46	6825.9	146.7919	9.85e-05	908.78

Table 9: Averaged results of IRPNM-reg, IRPNM-ls, IRPNM-reg-ls and AC-FISTA for 10 independent trials with tolerance $\texttt{tol} = 10^{-4}$.

All three second order methods produce similar objective values for all instances, with IRPNM-ls showing slightly better performance than IRPNM-reg across all different choices of the regularization parameter λ. The hybrid method IRPNM-reg-ls yields similar results as IRPNM-ls, performing slighty worse for $\lambda = 10^{-3}$ and slightly better for $\lambda = 10^{-4}$. We can see that AC-FISTA converges (on average) to slightly better stationary points than the second order methods for $\lambda = 10^{-3}$ and $\lambda = 10^{-4}$. Additionally, it demonstrates good runtime performance.



(a) Noisy blurred image (b) Original image (c) IRPNM-reg

Figure 1: Nonconvex image restoration with IRPNM-reg for $\lambda = 10^{-2}$ and $\texttt{tol} = 10^{-4}$ (reconstructed images with IRPNM-ls and AC-FISTA are omitted since they are indistinguishable from those obtained with IRPNM-reg).

7 Final Remarks

In this work, we introduced an inexact proximal Newton method without line search, ensuring global convergence through a careful update strategy for the regularization parameter based on the previous iteration. A superlinear convergence rate of the iterate sequence was shown under a local Hölderian error bound condition and confirmed in numerical tests across various problem classes. Our findings suggest several avenues for future research. Similar convergence results, i.e. without requiring a global Lipschitz assumption on ∇f, may be achievable for an inexact proximal Newton method using line search. Exploring analogous outcomes for a proximal Quasi-Newton method is another potential research direction. Additionally, a convergence analysis for $\delta = 0$ could be pursued under the assumption that F is a KL (Kurdyka-Łojawiewicz) function, following the approach in [23].

Acknowledgements The authors express their gratitude to Jiaming Liang from Yale University, Renato D. C. Monteiro from the Georgia Institute of Technology, Ruyu Liu and Shaohua Pan from the South China University of Technology, as well as Yuqia Wu and Xiaoqi Yang from the Hong Kong Polytechnic University for generously sharing their code.

Data availability The test problems in Section 6 are based on randomly generated data, except for the restoration of a blurred image. The Matlab code employed for both data generation and numerical tests is accessible upon request from Simeon vom Dahl (simeon.vomdahl@uni-wuerzburg.de)

References

[1] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[2] Stephen Becker, Jérôme Bobin, and Emmanuel J. Candès. Nesta: A fast and accurate first-order method for sparse recovery. *SIAM Journal on Imaging Sciences*, 4(1):1–39, 2011.

[3] Wei Bian and Xiaojun Chen. Linearly constrained non-Lipschitz optimization for image restoration. *SIAM Journal on Imaging Sciences*, 8(4):2294–2322, 2015.

[4] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2):459–494, 2014.

[5] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3:1–122, 2011.

[6] Richard H. Byrd, Jorge Nocedal, and Figen Oztoprak. An inexact successive quadratic approximation method for ℓ_1 regularized optimization. *Mathematical Programming*, 157(2):375–396, 2016.

[7] Alberto De Marchi. Proximal gradient methods beyond monotony. *Journal of Nonsmooth Analysis and Optimization*, 4(Original research articles), 2023.

[8] Bogdan Dumitrescu and Paul Irofti. *Dictionary Learning Algorithms and Applications*. Springer, 2018.

[9] Andreas Fischer. Local behavior of an iterative framework for generalized equations with nonisolated solutions. *Mathematical Programming*, 94:91–124, 2002.

[10] Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Springer, New York, 2013.

[11] Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.

[12] Jerome H. Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[13] Masao Fukushima and Hisashi Mine. A generalized proximal point algorithm for certain non-convex minimization problems. *International Journal of Systems Science*, 12(8):989–1000, 1981.

[14] Cho-jui Hsieh, Inderjit Dhillon, Pradeep Ravikumar, and Mátyás Sustik. Sparse inverse covariance matrix estimation using quadratic approximation. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[15] Xiaoxi Jia, Christian Kanzow, and Patrick Mehlitz. Convergence analysis of the proximal gradient method in the presence of the Kurdyka–Łojasiewicz property without global Lipschitz assumptions. *SIAM Journal on Optimization*, 33(4):3038–3056, 2023.

[16] Christian Kanzow and Theresa Lechner. Globalized inexact proximal Newton-type methods for nonconvex composite functions. *Computational Optimization and Applications*, 78:1–34, 2021.

[17] Christian Kanzow and Patrick Mehlitz. Convergence properties of monotone and nonmonotone proximal gradient methods revisited. *Journal of Optimization Theory and Applications*, 195:1–23, 2022.

[18] Kwangmoo Koh, Seung-Jean Kim, and Stephen Boyd. An interior-point method for large-scale l1-regularized logistic regression. *Journal of Machine Learning Research*, 8:1519–1555, 2007.

[19] Theresa Lechner. *Proximal Methods for Nonconvex Composite Optimization Problems.* PhD Thesis, Institute of Mathematics, University of Würzburg, 2022.

[20] Ching-Pei Lee and Stephen J. Wright. Inexact successive quadratic approximation for regularized optimization. *Computational Optimization and Applications*, 72(3):641–674, 2019.

[21] Jason D. Lee, Yuekai Sun, and Michael A. Saunders. Proximal Newton-type methods for minimizing composite functions. *SIAM Journal on Optimization*, 24:1420–1443, 2014.

[22] Jiaming Liang and Renato D. C. Monteiro. Average curvature fista for nonconvex smooth composite optimization problems. *Computational Optimization Appications*, 86(1):275–302, 2023.

[23] Ruyu Liu, Shaohua Pan, Yuqia Wu, and Xiaoqi Yang. An inexact regularized proximal Newton method for nonconvex and nonsmooth optimization. *Computational Optimization and Applications*, pages 1–39, 2024.

[24] Zhi-Quan Luo and Paul Tseng. On the linear convergence of descent methods for convex essentially smooth minimization. *SIAM Journal on Control and Optimization*, 30(2):408–425, 1992.

[25] Goran Marjanovic and Victor Solo. On ℓ_q optimization and matrix completion. *IEEE Transactions on Signal Processing*, 60:5714–5724, 2012.

[26] Ivan Markovsky. *Low Rank Approximation: Algorithms, Implementation, Applications.* Springer, second edition, 2019.

[27] Bernard Martinet. Détermination approchée d'un point fixe d'une application pseudo-contractante. Cas de l'application prox. *C. R. Acad. Sci. Paris Sér. A-B*, 274:A163–A165, 1972.

[28] P. Martinet. Régularisation d'inéquations variationnelles par approximations successives. 1970.

[29] Andre Milzarek and Michael Ulbrich. A semismooth Newton method with multidimensional filter globalization for l_1-optimization. *SIAM Journal on Optimization*, 24:298–333, 2014.

[30] Boris Mordukhovich. *Variational Analysis and Applications*, volume 30. Springer, 2018.

[31] Boris Mordukhovich, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. A globally convergent proximal Newton-type method in nonsmooth convex optimization. *Mathematical Programming*, 198:899–936, 2023.

[32] Figen Oztoprak, Jorge Nocedal, Steven Rennie, and Peder A Olsen. Newton-like methods for sparse inverse covariance estimation. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[33] R. T. Rockafellar. Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of Operations Research*, 1(2):97–116, 1976.

[34] R. Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

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

[36] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[37] Kenji Ueda. A regularized Newton method without line search for unconstrained optimization. *Computational Optimization and Applications*, 59:321–351, 2014.

[38] Kenji Ueda and Nobuo Yamashita. Convergence properties of the regularized Newton method for the unconstrained nonconvex optimization. *Applied Mathematics and Optimization*, 62:27–46, 2010.

[39] Yongchao Yu, Jigen Peng, and Shigang Yue. A new nonconvex approach to low-rank matrix completion with application to image inpainting. *Multidimensional Systems and Signal Processing*, 30:145–174, 2019.

[40] Guo-Xun Yuan, Chia-Hua Ho, and Chih-Jen Lin. An improved glmnet for l1-regularized logistic regression. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, page 33–41, New York, NY, USA, 2011.

[41] Man-Chung Yue, Zirui Zhou, and Anthony Man-Cho So. A family of inexact SQA methods for non-smooth convex minimization with provable convergence guarantees based on the Luo–Tseng error bound property. *Mathematical Programming*, 174:327–358, 2019.