FULL LENGTH PAPER

A family of inexact SQA methods for non-smooth convex minimization with provable convergence guarantees based on the Luo–Tseng error bound property

Man-Chung Yue[1] · **Zirui Zhou[2]** ·
Anthony Man-Cho So[3]

Abstract We propose a new family of inexact sequential quadratic approximation (SQA) methods, which we call the *inexact regularized proximal Newton* (IRPN) method, for minimizing the sum of two closed proper convex functions, one of which is smooth and the other is possibly non-smooth. Our proposed method features strong convergence guarantees even when applied to problems with degenerate solutions while allowing the inner minimization to be solved inexactly. Specifically, we prove that when the problem possesses the so-called Luo–Tseng error bound (EB) property, IRPN converges globally to an optimal solution, and the local convergence rate of the sequence of iterates generated by IRPN is linear, superlinear, or even quadratic, depending on the choice of parameters of the algorithm. Prior to this work, such EB property has been extensively used to establish the linear convergence of various first-order methods. However, to the best of our knowledge, this work is the first to

✉ Zirui Zhou
zrzhou@math.hkbu.edu.hk

Man-Chung Yue
m.yue@imperial.ac.uk

Anthony Man-Cho So
manchoso@se.cuhk.edu.hk

[1] Imperial College Business School, Imperial College London, London, UK

[2] Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Kowloon, Hong Kong

[3] Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong

use the Luo–Tseng EB property to establish the superlinear convergence of SQA-type methods for non-smooth convex minimization. As a consequence of our result, IRPN is capable of solving regularized regression or classification problems under the high-dimensional setting with provable convergence guarantees. We compare our proposed IRPN with several empirically efficient algorithms by applying them to the ℓ_1-regularized logistic regression problem. Experiment results show the competitiveness of our proposed method.

Keywords Convex composite minimization · Sequential quadratic approximation · Proximal Newton method · Error bound · Superlinear convergence

Mathematics Subject Classification 49M15 · 65K10 · 90C55

1 Introduction

A wide range of tasks in machine learning and statistics can be formulated as a non-smooth convex minimization problem of the form[1]

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

where $f, g : \mathbb{R}^n \to (-\infty, +\infty]$ are closed proper convex functions with f being twice continuously differentiable on an open subset of $\mathbb{R}^n$ containing the effective domain $\mathrm{dom}(g)$ of g and g being non-smooth. A popular choice for solving problem (1) is the sequential quadratic approximation (SQA) method (also called the proximal Newton method). Roughly speaking, in iteration k of a generic SQA method, one computes an (approximate) minimizer $\hat{x}^{k+1}$ of a quadratic model q_k of the objective function F at x^k, where

$$q_k(x) := f(x^k) + \nabla f(x^k)^T (x - x^k) + \frac{1}{2}(x - x^k)^T H_k (x - x^k) + g(x) \tag{2}$$

and H_k is a positive definite matrix approximating the Hessian $\nabla^2 f(x^k)$ of f at x^k. A step size α_k is obtained by performing a backtracking line search along the direction $d_k := \hat{x}^{k+1} - x^k$, and then $x^{k+1} := x^k + \alpha_k d^k$ is returned as the next iterate. Since in most cases, the (approximate) minimizer $\hat{x}^{k+1}$ of (2) does not admit a closed-form expression, an iterative algorithm, which we shall refer to as the *inner solver* in the sequel, is invoked. Throughout the paper, we will call the problem of minimizing the quadratic model (2) the *inner problem* and its solution the *inner solution*. We will also refer to the task of minimizing (2) as *inner minimization*.

There are three important ingredients that, more or less, determine an SQA method: the approximate Hessian H_k, the inner solver for minimizing q_k, and the stopping criterion of the inner solver to control the inexactness of the approximate inner solution $\hat{x}^{k+1}$. Indeed, many existing SQA methods and their variants that are tailored for

---

[1] Some authors refer to this as a convex composite minimization problem.

special instances of problem (1) can be obtained by specifying the aforementioned ingredients. Friedman et al. [13] developed the GLMNET algorithm for solving the ℓ_1-regularized logistic regression, where H_k is set to be the exact Hessian $\nabla^2 f(x^k)$ and a coordinate minimization method is used as the inner solver. Yuan et al. [45] improved GLMNET by replacing H_k with $\nabla^2 f(x^k) + \nu I$ for some constant $\nu > 0$ and adding a heuristic adaptive stopping strategy for inner minimization. This algorithm, called newGLMNET, is now the workhorse of the well-known LIBLINEAR package [10] for large-scale linear classification. Hsieh et al. [15] proposed the QUIC algorithm for solving sparse inverse covariance matrix estimation, which makes use of a quasi-Newton model to form the quadratic model (2). Similar strategies have also been employed in Olsen et al. [29], where the inner problem is solved by the fast iterative soft-shrinkage algorithm (FISTA) [1]. Other SQA variants can be found, e.g., in [2, 36,48].

Although the numerical advantage of the algorithms mentioned above has been well documented, their global convergence and convergence rate guarantees require the inner problems to be solved exactly. Such requirement is rarely satisfied in practice. To address this issue, Lee et al. [17] and Byrd et al. [4] proposed several families of SQA methods along with stopping criteria for the inner problem and showed that they are globally convergent and have a local superlinear convergence rate. However, all these convergence guarantees require the strong convexity of f. In fact, the strong convexity property is needed not only in the analyses but also for the well-definedness of the proposed methods.[2] Unfortunately, such a property is absent in many applications of interest. For example, consider the ℓ_1-regularized least squares regression problem, whose objective function takes the form

$$ F(x) = \frac{1}{m} \sum_{i=1}^{m} \left(a_i^T x - b_i \right)^2 + \mu \|x\|_1, \quad \mu > 0. $$

Observe that the smooth part $x \mapsto \frac{1}{m} \sum_{i=1}^{m} \left(a_i^T x - b_i \right)^2$ is strongly convex if and only if the data matrix $A := [a_1 \cdots a_m] \in \mathbb{R}^{n \times m}$ has full row rank. The latter cannot be ensured in most applications and is even impossible in the high-dimensional (i.e., $n \gg m$) setting. In addition, the authors of [17] did not provide any global convergence result for the inexact version of their proposed method, while those of [4] considered only the case where $g(x) = c \cdot \|x\|_1$ for some constant $c > 0$. On another front, Scheinberg and Tang [35] proposed an inexact SQA-type method and analyzed its global complexity. Although their analysis does not require the strong convexity of f, it only yields a sublinear convergence rate for the proposed method. In view of the above discussion, the theoretical study of existing SQA methods is rather incomplete, and we are motivated to develop an algorithm for solving problem (1) that does not require an exact inner solver or the strong convexity of f but can still be shown to converge globally and possess a local superlinear convergence rate.

---

[2] For instance, the exact Hessian $H_k = \nabla^2 f(x^k)$ is used in [17]. If f is not strongly convex, then neither is the quadratic model (2). As such, the inner problem can have multiple minimizers and the next iterate x^{k+1} is not well defined.

In this paper, we propose a new family of inexact SQA methods, which we call the *inexact regularized proximal Newton* (IRPN) method, that are capable of solving non-strongly convex instances of problem (1) without using an exact inner solver. In iteration k, IRPN takes H_k to be the regularized Hessian $H_k = \nabla^2 f(x^k) + \mu_k I$, where $\mu_k = c \cdot r(x^k)^\rho > 0$ for some constants $c > 0$, $\rho \in [0, 1]$, and r is an easily computable residual function that measures the proximity of x^k to the optimal solution set (see Sect. 2.2 for details). The inner problem is solved up to an accuracy that is determined by an adaptive inexactness condition, which also depends on the residue $r(x^k)$ and the parameter ρ. It is worth noting that the idea of regularization is not new for second-order methods. Indeed, Li et al. [18] investigated the regularized Newton method for solving smooth convex minimization problems with degenerate solutions. The well-known Levenberg–Marquardt (LM) method for solving nonlinear equations is essentially a regularized Gauss–Newton method [25,42]. The algorithm that is closest in spirit to ours is newGLMNET [45], which aims at solving problems of the form (1) and also uses a regularized Hessian, albeit the regularization parameter remains constant throughout the entire execution; i.e., $\mu_k = \nu$ for all k with $\nu > 0$. Hence, it may be seen as a special case ($\rho = 0$) of IRPN. However, the constant ν is chosen empirically and a heuristic stopping rule is adopted for the inner solver.

Another motivation of this research comes from the seminal paper [24] of Luo and Tseng, in which a certain error bound (EB) property was introduced. The Luo–Tseng EB property is provably a less stringent requirement on the objective function F of problem (1) than strong convexity and has been shown to hold for a wide range of F's; see Fact 3. Since its introduction, the Luo–Tseng EB property has played an important role in the convergence rate analysis of various first-order methods. Specifically, it has been utilized to establish the linear convergence of the projected gradient descent, proximal gradient descent, coordinate descent, and block coordinate descent methods for solving possibly non-strongly convex instances of problem (1); see [14,24,38,39,47]. It is interesting to note that if the objective function F of problem (1) satisfies the Luo–Tseng EB property, then it also satisfies the so-called *Kurdyka–Łojasiewicz (KL)* property with exponent $1/2$ [19].[3] The latter can also be used to establish the linear convergence of various first-order methods; see, e.g., [19,21] and the references therein.

Following the great success of EB-based analysis of first-order methods, it is natural to ask whether one can develop an EB-based analysis of second-order methods for solving problem (1). Such an analysis is highly desirable, as it would advance our understanding of the strength and limitation of SQA-type methods and more accurately capture the interplay between the algorithmic and geometric properties of problem (1). It turns out that some partial answers in this direction are available. Li et al. [18] proposed a regularized Newton method along with inexactness conditions for solving *smooth* convex minimization problems and proved its global convergence and local quadratic convergence based on an EB property. Unfortunately, their analysis relies heavily on the eigenvalue perturbation properties of the Hessian of the smooth

---

[3] In [16] the authors considered global versions of the Luo–Tseng EB and KL properties and showed that they are equivalent. However, none of the scenarios listed in Fact 3 except (S1) are known to possess the global Luo–Tseng EB property stated in [16].

objective function and hence cannot be directly extended to second-order methods for solving the *non-smooth* minimization problem (1). Tseng and Yun [39] proposed the (block) coordinate gradient descent (CGD) method, which includes the SQA method as a special case, for solving problem (1) with non-convex f and non-smooth convex g that has a separable structure. Although their analysis utilizes the Luo–Tseng EB property, their proposed CGD method requires the sequence of approximate Hessians $\{H_k\}_{k \geq 0}$ to be uniformly lower bounded—i.e., $\lambda_{\min}(H_k) \geq \underline{\lambda}$ for all k, where $\underline{\lambda} > 0$ is some constant—and the provable local convergence rate is only linear. Another important contribution along this direction is the work [11] by Fischer, in which he studied an abstract iterative framework for solving a class of generalized equations and proved that it has a local superlinear convergence rate. Besides requiring an EB-type property and an inexactness condition, the result requires the distance between two consecutive iterates to be linearly bounded by the distance between the current iterate and the set of solutions $\mathcal{X}$ to the generalized equation; i.e., $\|x^{k+1} - x^k\| = O(\text{dist}(x^k, \mathcal{X}))$. However, for a concrete algorithm, establishing such an estimate is usually a non-trivial task. Moreover, the iterative framework in [11] does not involve regularization (see displayed equations (3) and (4) of [11]). Hence, it is not obvious how our proposed method can be put into that framework. Recently, the authors of [7,8,12] studied an inexact Newton-type method for solving constrained systems of equations and proved its global convergence and local quadratic convergence based on an EB property. As the set of optimal solutions to problem (1) can be characterized as the set of solutions to a certain system of equations (see Fact 2 and the discussion following it), their algorithm can also be applied to solve problem (1). However, their method is fundamentally different from ours, as the inner problem of their method involves solving a linear program, while that of ours involves solving an unconstrained strongly convex minimization problem (see Sect. 3 for details). Interestingly, in the context of problem (1), the EB property considered by them also coincides with ours. Nevertheless, the convergence analysis of their proposed method assumes that the linear program defining the inner problem satisfies a so-called uniform inf-boundedness property in a neighborhood of an optimal solution to problem (1); see [8, Assumption 3]. Although several sufficient conditions for such assumption to hold are developed in [8, Section 3], they do not apply to some of the instances of problem (1) that we are interested in.

Our contributions in this paper can be summarized as follows. First, we propose a new family of inexact SQA methods called IRPN for solving the non-smooth convex minimization problem (1). The proposed method has a parameter $\rho \in [0, 1]$ and comes with a stopping criterion for inner minimization that allows us to solve the inner problem inexactly. Second, we establish the global convergence and local convergence rate of IRPN. More precisely, we prove that each accumulation point of the sequence of iterates generated by IRPN is optimal for problem (1). Furthermore, when problem (1) possesses the Luo–Tseng EB property, the sequence converges to an optimal solution, and the local convergence rate is at least *R-linear* if $\rho = 0$, *Q-superlinear* if $\rho \in (0, 1)$, and *Q-quadratic* if $\rho = 1$ (see [27, Appendix A.2] for the definitions of these different types of convergence). Our analysis is novel and establish, for the first time, the superlinear convergence of inexact SQA-type methods when applied to instances of problem (1) that possess the Luo–Tseng EB property.

The paper is organized as follows. In Sect. 2, we set the stage for our theoretical development by stating the basic assumptions and reviewing the relevant existing results. We also discuss the Luo–Tseng EB property in the context of problem (1) and some other related regularity conditions. In Sect. 3, we detail our proposed IRPN, including the specification of the regularized Hessian and inexactness condition for inner minimization. In Sect. 4, we establish the global convergence and local convergence rate of IRPN. We then present some numerical results in Sect. 5. Finally, we conclude our paper in Sect. 6.

Notation. We denote the optimal value of and the set of optimal solutions to problem (1) by F^* and $\mathcal{X}$, respectively. We use $\|\cdot\|$ to denote the Euclidean norm for vectors and the operator norm (see [3, Exercise I.2.4]) for matrices. Given a set $C \subseteq \mathbb{R}^n$, we denote by $\text{dist}(x, C)$ the distance from x to C; i.e., $\text{dist}(x, C) = \inf\{\|x - u\| \mid u \in C\}$. For a closed convex function h, we denote by ∂h the subdifferential of h. If h is continuously differentiable (resp. twice continuously differentiable), we denote by ∇h (resp. $\nabla^2 h$) the gradient (resp. Hessian) of h.

2 Preliminaries

Consider the convex minimization problem (1). We assume that f, g are closed proper convex functions with f being twice continuously differentiable on an open subset of $\mathbb{R}^n$ containing $\text{dom}(g)$ and g being possibly non-smooth. To avoid ill-posed problems, we assume that the set of optimal solutions $\mathcal{X}$ to problem (1) is non-empty, so that the optimal value F^* of problem (1) is finite. Furthermore, we make the following assumptions.

Assumption 1 The function f in problem (1) satisfies the following:

(a) The gradient ∇f is Lipschitz continuous on an open set $\mathcal{U}$ containing $\text{dom}(g)$; i.e., there exists a constant $L_1 > 0$ such that

$$\|\nabla f(y) - \nabla f(z)\| \le L_1 \|y - z\| \quad \forall y, z \in \mathcal{U}.$$

(b) The Hessian $\nabla^2 f$ is Lipschitz continuous on an open set $\mathcal{U}$ containing $\text{dom}(g)$; i.e., there exists a constant $L_2 > 0$ such that

$$\|\nabla^2 f(y) - \nabla^2 f(z)\| \le L_2 \|y - z\| \quad \forall y, z \in \mathcal{U}.$$

The above assumptions are standard in the analysis of Newton-type methods. As we will see in Sects. 4 and 5, Assumption 1(a) is crucial to the global convergence analysis of IRPN, while Assumption 1(b) is needed for the local convergence analysis. The following result is a direct consequence of Assumption 1; see, e.g., [26, Lemma 1.2.2].

Fact 1 *Under Assumption 1(a), for any $x \in \text{dom}(g)$, we have*

$$\lambda_{\max}(\nabla^2 f(x)) \le L_1.$$

2.1 Optimality conditions and residual functions

We now introduce the optimality condition of problem (1). Given a closed proper convex function $h : \mathbb{R}^n \to (-\infty, +\infty]$, the so-called *proximal operator* $\mathrm{prox}_h : \mathbb{R}^n \to \mathbb{R}^n$ of h is defined by

$$\mathrm{prox}_h(v) := \underset{u \in \mathbb{R}^n}{\mathrm{argmin}} \left\{ \frac{1}{2} \|u - v\|^2 + h(u) \right\}.$$

The proximal operator is the building block of many first-order methods for solving problem (1), including the proximal gradient method and its accelerated versions [32] and (block) coordinate gradient descent methods [39]. Moreover, the proximal operators of many non-smooth functions, such as the indicator function of a closed convex set, the ℓ_1-norm, the grouped LASSO regularizer, the elastic net regularizer, and the nuclear norm, have closed-form representations; see, e.g., [32, Chapter 6]. As is well known, the set of optimal solutions to problem (1) can be characterized using the proximal operator. The following result, which can be found, e.g., in [5, Proposition 3.1], will play an important role in our subsequent development.

Fact 2 *A vector $x \in \mathrm{dom}(g)$ is an optimal solution to problem* (1) *if and only if for any $\tau > 0$,*

$$x = \mathrm{prox}_{\tau g}(x - \tau \nabla f(x)).$$

Let $R : \mathrm{dom}(g) \to \mathbb{R}^n$ be the map given by

$$R(x) := x - \mathrm{prox}_g(x - \nabla f(x)). \tag{3}$$

Fact 2 suggests that we can take $r(x) := \|R(x)\|$ as a *residual function* for problem (1); i.e., the function r satisfies $r(x) \geq 0$ for all $x \in \mathrm{dom}(g)$ and $r(x) = 0$ if and only if $x \in \mathcal{X}$. In addition, the following proposition shows that both R and r are Lipschitz continuous on $\mathrm{dom}(g)$ if Assumption 1(a) holds.

Proposition 1 *Suppose that Assumption 1(a) holds. Then, for any $y, z \in \mathrm{dom}(g)$, we have*

$$|r(y) - r(z)| \leq \|R(y) - R(z)\| \leq (L_1 + 2)\|y - z\|.$$

Proof The first inequality is a direct consequence of the triangle inequality. We now prove the second one. Using the definition of R, we compute

$$\begin{aligned}
\|R(y) - R(z)\| &= \|y - \mathrm{prox}_g(y - \nabla f(y)) - z + \mathrm{prox}_g(z - \nabla f(z))\| \\
&\leq \|y - z\| + \|\mathrm{prox}_g(y - \nabla f(y)) - \mathrm{prox}_g(z - \nabla f(z))\| \\
&\leq 2\|y - z\| + \|\nabla f(y) - \nabla f(z)\| \\
&\leq (L_1 + 2)\|y - z\|,
\end{aligned}$$

where the second inequality follows from the non-expansiveness of the proximal operator (see, e.g., [5, Lemma 2.4]) and the last is by Assumption 1(a).   □

2.1.1 Inner minimization

Recall that in the k-th iteration of an SQA method, one (approximately) minimizes the quadratic model q_k in (2). By letting

$$f_k(x) := f(x^k) + \nabla f(x^k)^T (x - x^k) + \frac{1}{2}(x - x^k)^T H_k(x - x^k),$$

the inner problem reads

$$\min_{x \in \mathbb{R}^n} \{q_k(x) := f_k(x) + g(x)\}, \tag{4}$$

which is also a convex minimization problem of the form (1) with the smooth part being quadratic. Therefore, both the optimality condition and residual function studied earlier for problem (1) can be adapted to the inner problem (4). The following corollary is immediate from Fact 2 and the fact that $\nabla f_k(x) = \nabla f(x^k) + H_k(x - x^k)$.

Corollary 1 *A vector* $x \in \mathrm{dom}(g)$ *is an optimal solution to problem* (4) *if and only if for any* $\tau > 0$,

$$x = \mathrm{prox}_{\tau g}(x - \tau \nabla f_k(x)) = \mathrm{prox}_{\tau g}\left((I - \tau H_k)x - \tau(\nabla f(x^k) - H_k x^k)\right).$$

Similar to the map R in (3), we define the map $R_k : \mathrm{dom}(g) \to \mathbb{R}^n$ by

$$R_k(x) := x - \mathrm{prox}_g(x - \nabla f_k(x)) = x - \mathrm{prox}_g\left((I - H_k)x - (\nabla f(x^k) - H_k x^k)\right).$$

Furthermore, parallel to the residual function r defined for problem (1), we can define the residual function for problem (4) as $r_k(x) := \|R_k(x)\|$. Now, following the lines of the proof of Proposition 1, we can easily show that both R_k and r_k are Lipschitz continuous.

Corollary 2 *For any* $y, z \in \mathbb{R}^n$, *we have*[4]

$$|r_k(y) - r_k(z)| \le \|R_k(y) - R_k(z)\| \le (\lambda_{\max}(H_k) + 2)\|y - z\|.$$

2.2 The Luo–Tseng error bound property

A prevailing assumption in existing convergence analyses of SQA methods for solving problem (1) is the strong convexity of the smooth function f [4,17]. However, such assumption is invalid in many applications (see the discussion in Sect. 1). Instead of

---

[4] Note that Assumption 1(a) is not required for Corollary 2 to hold; *cf.* Proposition 1.

assuming strong convexity, our analysis of IRPN is based on the following local EB property.

Assumption 2 (*Luo–Tseng EB Property*) Let r be the residual function defined in Sect. 2.1. For any $\zeta \geq F^*$, there exist scalars $\kappa > 0$ and $\epsilon > 0$ such that

$$\text{dist}(x, \mathcal{X}) \leq \kappa \cdot r(x) \quad \text{whenever } F(x) \leq \zeta \text{ and } r(x) \leq \epsilon. \tag{5}$$

EBs have long been an important topic and permeate in all aspects of mathematical programming [9,31]. The Luo–Tseng EB property (5) for problem (1) was studied and popularized in a series of papers by Luo and Tseng [22–24].[5] In particular, many useful subclasses of problem (1) have been shown to possess the Luo–Tseng EB property. The early results (see, e.g., [24] and the references therein) mainly focus on the case where the non-smooth function g has a polyhedral epigraph. It is only recently that the validity of the Luo–Tseng EB is established for instances of problem (1) with a non-polyhedral g [38,50] or even with a non-polyhedral optimal solution set [49]. We briefly summarize these results below and refer the readers to the recent work of Zhou and So [49] for a more detailed review of the developments of the Luo–Tseng EB.

Fact 3 *For problem (1), the Luo–Tseng EB (Assumption 2) holds in any of the following scenarios:*

(S1) ([39, Theorem 4]; cf. [30, Theorem 3.1]) f is strongly convex, ∇f is Lipschitz continuous, and g is a closed proper convex function.

(S2) ([39, Lemma 7]; cf. [23, Theorem 2.1]) f takes the form $f(x) = h(Ax) + \langle c, x \rangle$, where $A \in \mathbb{R}^{m \times n}$ and $c \in \mathbb{R}^n$ are given, $h : \mathbb{R}^m \to (-\infty, +\infty)$ is a continuously differentiable function with h being strongly convex and ∇h being Lipschitz continuous on any compact convex subset of $\mathbb{R}^n$, and g has a polyhedral epigraph.

(S3) ([50, Corollaries 1 and 2]; cf. [38, Theorem 2] and [47, Theorem 1]) f takes the form $f(x) = h(Ax)$, where $A \in \mathbb{R}^{m \times n}$, $h : \mathbb{R}^m \to (-\infty, +\infty)$ are as in scenario (S2), and g is the $\ell_{1,p}$-norm regularizer with $p \in [1, 2] \cup \{+\infty\}$ (i.e., $g(x) = \sum_{J \in \mathcal{J}} \omega_J \|x_J\|_p$, where $\mathcal{J}$ is a partition of the index set $\{1, \ldots, n\}$, $x_J \in \mathbb{R}^{|J|}$ is the sub-vector obtained by restricting $x \in \mathbb{R}^n$ to the entries in $J \in \mathcal{J}$, and $\omega_J \geq 0$ is a given parameter; see [50] and the references therein for some background on the $\ell_{1,p}$-norm regularizer).

(S4) ([49, Proposition 12]) f takes the form $f(X) = h(\mathcal{A}(X)) + \langle C, X \rangle$, where $\mathcal{A} : \mathbb{R}^{n \times p} \to \mathbb{R}^m$ is a linear mapping, $C \in \mathbb{R}^{n \times p}$ is a given matrix, $h : \mathbb{R}^m \to (-\infty, +\infty)$ is as in scenario (S2), g is the nuclear norm regularizer (i.e., $g(X)$ equals the sum of all the singular values of X; see [49] and the references therein for some background on the nuclear norm regularizer), and there exists an $X^* \in \mathcal{X}$ such that the following strict complementary-type condition holds (here, ri denotes the relative interior):

$$\mathbf{0} \in \nabla f(X^*) + \text{ri}(\partial g(X^*)).$$

---

[5] A similar EB property has been studied by Pang [30] for linearly constrained variational inequalities.

Note that in scenarios (S2)–(S4), the assumptions on h are the same and can readily be shown to be satisfied by $h(y) = \frac{1}{2}\|y - b\|^2$, which corresponds to least squares regression, and $h(y) = \sum_{i=1}^m \log(1 + e^{-b_i y_i})$ with $b \in \{-1, 1\}^m$, which corresponds to logistic regression. The assumptions on h are also satisfied by loss functions that arise in maximum likelihood estimation (MLE) for learning conditional random fields [48] and MLE under Poisson noise [34]. It follows from Fact 3 that many problems of the form (1) possess the Luo–Tseng EB property, even though they may not be strongly convex.

Prior to this work, the Luo–Tseng EB property (5) has been used to establish the local linear convergence of a number of first-order methods for solving problem (1); see the discussion in Sect. 1. It has also been used to prove that certain primal-dual interior-point path following methods converge superlinearly [37]. However, all the methods in question are quite different in nature from the SQA methods considered in this paper. In Sect. 4, we show how the Luo–Tseng EB property (5) can be used to establish the local superlinear convergence of our proposed IRPN, thus further demonstrating its versatility in convergence analysis.

2.3 Other regularity conditions

Besides the Luo–Tseng EB property (5), other regularity conditions have been used to establish the superlinear convergence of SQA methods for solving problem (1). Let us briefly review some of those conditions here and explain why they are too stringent for the scenarios that we are interested in.

Yen et al. [43] and Zhong et al. [48] introduced the *constant nullspace strong convexity* (CNSC) property of a smooth function. They showed that when the smooth function f possesses the CNSC property and the non-smooth function g satisfies some other regularity conditions, the proximal Newton and proximal quasi-Newton methods in [17] converge quadratically and superlinearly, respectively. Nonetheless, their convergence results are different from those obtained in this paper. Indeed, let $\{x^k\}_{k\geq0}$ be the sequence of iterates generated by the above methods and z^k be the projection of x^k onto the subspace associated with the CNSC property. The results in [43,48] imply that the sequence $\{z^k\}_{k\geq0}$ converges to a point z^* quadratically or superlinearly. However, the convergence rate of $\{x^k\}_{k\geq0}$ remains unclear. In fact, there is no guarantee that $\{x^k\}_{k\geq0}$ or $\{z^k\}_{k\geq0}$ converges to an optimal solution to the problem. We also remark that neither the inexactness of the inner minimization nor the global convergence of the methods was addressed in [43,48].

Dontchev and Rockafellar [6] developed a framework for solving generalized equations, which coincides with the standard SQA method when specialized to the optimality condition $\mathbf{0} \in \nabla f(x) + \partial g(x)$ of problem (1). The corresponding convergence results require the set-valued mapping $\nabla f + \partial g$ to be either *metrically regular* or *strongly metrically sub-regular*; see [6, Chapter 6C]. However, both of these regularity conditions are provably more restrictive than the Luo–Tseng EB property (5). In particular, they require that the problem at hand has a unique optimal solution, which is not satisfied by many instances of problem (1) that are of interest. For example, consider the following two-dimensional ℓ_1-regularized least squares regression problem:

$$\min_{x_1, x_2 \in \mathbb{R}} \left\{ \frac{1}{2}(x_1 + x_2 - 2)^2 + |x_1| + |x_2| \right\}.$$

This problem possesses the Luo–Tseng EB property (5), as it belongs to scenario (S2) in Fact 3. However, it can be verified that the set-valued mapping associated with the optimality condition of this problem is neither metrically regular nor strongly metrically sub-regular.

3 The inexact regularized proximal Newton method

We now describe in detail our proposed algorithm—the inexact regularized proximal Newton (IRPN) method. The algorithm takes as input an initial iterate $x^0 \in \mathbb{R}^n$, a constant $\epsilon_0 > 0$ that controls the solution precision, constants $\theta \in (0, 1/2)$, $\zeta \in (\theta, 1/2)$, $\eta \in (0, 1)$ that are used to specify the inexactness condition and line-search parameters, and constants $c > 0$, $\rho \in [0, 1]$ that are used to form the regularized Hessians $\{H_k\}_{k \geq 0}$. As we will see in Sect. 4, the local convergence rate of IRPN largely depends on the choice of ρ.

At the current iterate x^k, we first construct a quadratic approximation of F at x^k by

$$q_k(x) := f(x^k) + \nabla f(x^k)^T (x - x^k) + \frac{1}{2}(x - x^k)^T H_k (x - x^k) + g(x),$$

where $H_k = \nabla^2 f(x^k) + \mu_k I$ with $\mu_k = c \cdot r(x^k)^\rho$ and r is the residual function defined in Sect. 2.1 for problem (1). Since f is convex and $\mu_k > 0$, the matrix H_k is positive definite for all k. Hence, the quadratic model q_k is strongly convex and has a unique minimizer. However, since the exact minimizer does not admit a closed-form expression in most cases, an iterative algorithm, such as a coordinate minimization method, the coordinate gradient descent method, or the accelerated proximal gradient method, is typically called to find an *approximate* minimizer $\hat{x}^{k+1}$ of the quadratic model q_k. To ensure that the method has the desired convergence properties, we require the vector $\hat{x}^{k+1}$ to satisfy

$$r_k(\hat{x}^{k+1}) \leq \eta \cdot \min\{r(x^k), r(x^k)^{1+\rho}\}, \tag{6a}$$

$$q_k(\hat{x}^{k+1}) - q_k(x^k) \leq \zeta \left(\ell_k(\hat{x}^{k+1}) - \ell_k(x^k) \right), \tag{6b}$$

where r_k is the residual function defined in Sect. 2.1.1 for the inner problem (4) and ℓ_k is the first-order approximation of F at x^k:

$$\ell_k(x) := f(x^k) + \nabla f(x^k)^T (x - x^k) + g(x).$$

We will show in Lemma 1 that the inexactness condition (6) can be satisfied by vectors that are sufficiently close to the exact minimizer of q_k. After finding the approximate minimizer $\hat{x}^{k+1}$, we perform a backtracking line search along the direction $d^k := \hat{x}^{k+1} - x^k$ to obtain a step size $\alpha_k > 0$ that can guarantee a sufficient decrease in

the objective value. The algorithm then steps into the next iterate by setting $x^{k+1} := x^k + \alpha_k d^k$. Finally, we terminate the algorithm when $r(x^k)$ is less than the prescribed precision ϵ_0. We summarize the details of IRPN in Algorithm 1.

Algorithm 1 Inexact Regularized Proximal Newton (IRPN) Method

---

1: **Input:** initial iterate $x^0 \in \mathbb{R}^n$, constants $\epsilon_0 > 0$, $\theta \in (0, 1/2)$, $\zeta \in (\theta, 1/2)$, $\eta \in (0, 1)$, $c > 0$, $\rho \in [0, 1]$, and $\beta \in (0, 1)$.
2: **for** $k = 0, 1, 2, \ldots$ **do** {outer iteration}
3:    compute the value of the residue $r(x^k)$
4:    if $r(x^k) \leq \epsilon_0$, terminate the algorithm and return x^k
5:    form the quadratic model q_k with $H_k = \nabla^2 f(x^k) + \mu_k I$ and $\mu_k = c \cdot r(x^k)^\rho$
6:    {inner iteration} call an inner solver and find an approximate minimizer $\hat{x}^{k+1}$ of q_k that satisfies

$$r_k(\hat{x}^{k+1}) \leq \eta \cdot \min\{r(x^k), r(x^k)^{1+\rho}\} \text{ and } q_k(\hat{x}^{k+1}) - q_k(x^k) \leq \zeta \left(\ell_k(\hat{x}^{k+1}) - \ell_k(x^k) \right) \quad (6)$$

7:    set the search direction $d^k := \hat{x}^{k+1} - x^k$ and find the smallest integer $i \geq 0$ such that

$$F(x^k) - F(x^k + \beta^i d^k) \geq \theta \left(\ell_k(x^k) - \ell_k(x^k + \beta^i d^k) \right) \quad (7)$$

8:    set the step size $\alpha_k = \beta^i$ and the next iterate $x^{k+1} = x^k + \alpha_k d^k$
9: **end for**

The inexactness condition (6) is similar to that proposed in [4]. However, the analysis in [4] cannot be applied to study the convergence behavior of IRPN. Indeed, the global convergence analysis therein requires that $H_k \succeq \lambda I$ with $\lambda > 0$ for all k, while the local convergence rate analysis requires $\nabla^2 f$ to be positive definite at the limit point x^* of the sequence $\{x^k\}_{k \geq 0}$. However, for many of the instances of problem (1) that we are interested in, neither of these requirements are satisfied.

One advantage of our proposed IRPN lies in its flexibility. Indeed, for SQA methods that use regularized Hessians, it is typical to let the regularization parameter μ_k be of the same order as the residue $r(x^k)$ [18,33]; i.e., $\mu_k = c \cdot r(x^k)$. By contrast, we can adjust the order of μ_k according to the parameter ρ. Therefore, IRPN is a tunable family of algorithms parametrized by $\rho \in [0, 1]$. As we will see in Sects. 4 and 5, the parameter ρ plays a dominant role in the local convergence rate of IRPN. Even more, IRPN allows one to choose the inner solver. With the objective being the sum of a strongly convex quadratic function and a non-smooth convex function, each inner problem is a well-structured convex minimization problem. Hence, one can exploit the structure of $\nabla^2 f$ and g to design special inner solvers to speed up the inner minimization.

Before we discuss the convergence properties of IRPN, we need to show that it is well defined. Specifically, we need to argue that the inexactness condition (6) is always feasible and the line search procedure will terminate in a finite number of steps. We first present the following lemma, which ensures that in each iteration, the inexactness condition (6) is satisfied by vectors that are close enough to the exact minimizer of

the inner problem. The proof is identical to [4, Lemma 4.5], but we present it here for completeness.

Lemma 1 *For any iteration k of Algorithm* 1*, the inexactness condition (6) is satisfied by any vector x that is sufficiently close to the exact minimizer $\tilde{x}^{k+1}$ of the quadratic model q_k.*

Proof Consider the iterate x^k in the k-th iteration. We may assume that x^k is not optimal for problem (1), for otherwise the algorithm would have terminated. The continuity of r_k and the fact that $r_k(\tilde{x}^{k+1}) = 0$ imply that condition (6a) is satisfied by any x that is sufficiently close to $\tilde{x}^{k+1}$. Now, let us consider condition (6b). Since $q_k(x) = \ell_k(x) + \frac{1}{2}(x - x^k)^T H_k(x - x^k)$ and $\tilde{x}^{k+1}$ is the minimizer of q_k, we have $H_k(x^k - \tilde{x}^{k+1}) \in \partial \ell_k(\tilde{x}^{k+1})$. This leads to $x^k \neq \tilde{x}^{k+1}$, for otherwise we would have $\mathbf{0} \in \partial \ell_k(x^k)$, which would imply that x^k is an optimal solution to problem (1). Moreover, by the convexity of ℓ_k, we have

$$\ell_k(x^k) \geq \ell_k(\tilde{x}^{k+1}) + (\tilde{x}^{k+1} - x^k)^T H_k(\tilde{x}^{k+1} - x^k). \tag{8}$$

Since $H_k = \nabla^2 f(x^k) + \mu_k I$ for some $\mu_k > 0$, H_k is positive definite. This, together with the fact that $x^k \neq \tilde{x}^{k+1}$, implies that $\ell_k(x^k) > \ell_k(\tilde{x}^{k+1})$. Furthermore, we have

$$q_k(\tilde{x}^{k+1}) - q_k(x^k) = \ell_k(\tilde{x}^{k+1}) + \frac{1}{2}(\tilde{x}^{k+1} - x^k)^T H_k(\tilde{x}^{k+1} - x^k) - \ell_k(x^k)$$

$$\leq \frac{1}{2}\left(\ell_k(\tilde{x}^{k+1}) - \ell_k(x^k)\right)$$

$$< \zeta\left(\ell_k(\tilde{x}^{k+1}) - \ell_k(x^k)\right),$$

where the first inequality is due to (8) and the last is due to $\zeta \in (\theta, 1/2)$ and $\ell_k(x^k) > \ell_k(\tilde{x}^{k+1})$. Therefore, by the continuity of q_k and ℓ_k, condition (6b) is satisfied by any x that is close enough to $\tilde{x}^{k+1}$. □

We next show that the backtracking line search is well defined.

Lemma 2 *Suppose that Assumption* 1*(a) holds. Then, for any iteration k of Algorithm* 1*, there exists an integer $i \geq 0$ such that the descent condition (7) is satisfied. Moreover, the step size α_k obtained from the line search strategy satisfies*

$$\alpha_k \geq \frac{\beta(1 - \theta)\mu_k}{(1 - \zeta)L_1}.$$

Proof By definition, we have

$$q_k(\hat{x}^{k+1}) - q_k(x^k) = \ell_k(\hat{x}^{k+1}) - \ell_k(x^k) + \frac{1}{2}(\hat{x}^{k+1} - x^k)^T H_k(\hat{x}^{k+1} - x^k).$$

Since the approximate minimizer $\hat{x}^{k+1}$ satisfies the inexactness condition (6), the above implies that

$$\ell_k(x^k) - \ell_k(\hat{x}^{k+1}) \geq \frac{1}{2(1-\zeta)} (\hat{x}^{k+1} - x^k)^T H_k (\hat{x}^{k+1} - x^k)$$

$$\geq \frac{\mu_k}{2(1-\zeta)} \|\hat{x}^{k+1} - x^k\|^2. \tag{9}$$

Due to the convexity of ℓ_k, for any $\alpha \in [0, 1]$, we have

$$\ell_k(x^k) - \ell_k(x^k + \alpha d^k) \geq \alpha \left(\ell_k(x^k) - \ell_k(x^k + d^k) \right).$$

Combining the above two inequalities and noting that $\hat{x}^{k+1} = x^k + d^k$, we obtain

$$\ell_k(x^k) - \ell_k(x^k + \alpha d^k) \geq \frac{\alpha \mu_k}{2(1-\zeta)} \|d^k\|^2. \tag{10}$$

Moreover, since ∇f is Lipschitz continuous by Assumption 1(a), we have

$$f(x^k + \alpha d^k) - f(x^k) \leq \alpha \nabla f(x^k)^T d^k + \frac{\alpha^2 L_1}{2} \|d^k\|^2,$$

which, by the definition of ℓ_k, leads to

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

It then follows from the above inequality and (10) that

$$F(x^k) - F(x^k + \alpha d^k) - \theta \left(\ell_k(x^k) - \ell_k(x^k + \alpha d^k) \right)$$

$$\geq (1-\theta) \left(\ell_k(x^k) - \ell_k(x^k + \alpha d^k) \right) - \frac{\alpha^2 L_1}{2} \|d^k\|^2$$

$$\geq \frac{(1-\theta)\alpha \mu_k}{2(1-\zeta)} \|d^k\|^2 - \frac{\alpha^2 L_1}{2} \|d^k\|^2$$

$$= \frac{\alpha}{2} \left(\frac{1-\theta}{1-\zeta} \mu_k - \alpha L_1 \right) \|d^k\|^2.$$

Hence, as long as α satisfies $\alpha < (1-\theta)\mu_k/(1-\zeta)L_1$, the descent condition (7) is satisfied. Since the backtracking line search multiplies the step length by $\beta \in (0, 1)$ after each trial, the line search strategy will output an α_k that satisfies $\alpha_k \geq \beta(1-\theta)\mu_k/(1-\zeta)L_1$ in a finite number of steps. □

Combining Lemmas 1 and 2, we conclude that IRPN is well defined.

4 Convergence analysis of IRPN

In this section, we establish the global convergence and local convergence rate of IRPN. Let $\{x^k\}_{k \geq 0}$ be the sequence of iterates generated by Algorithm 1. We first

present the following result, which shows that the residues $\{r(x^k)\}_{k \geq 0}$ eventually vanish.

Proposition 2 *Suppose that Assumption 1(a) holds. Then, we have*

$$\lim_{k \to \infty} r(x^k) = 0.$$

In particular, every accumulation point of the sequence $\{x^k\}_{k \geq 0}$ is an optimal solution to problem (1).

Proof From (10), we obtain

$$\ell_k(x^k) - \ell_k(x^k + \alpha d^k) \geq \frac{\alpha \mu_k}{2(1 - \zeta)} \|d^k\|^2 \geq 0.$$

Hence, due to the descent condition (7) and the assumption that $F^* > -\infty$, we have $\lim_{k \to \infty} \left(\ell_k(x^k) - \ell_k(x^k + \alpha_k d^k) \right) = 0$. This, together with (10) again, implies that

$$\lim_{k \to \infty} \alpha_k \mu_k \|d^k\|^2 = 0. \tag{11}$$

On the other hand, we have $r_k(\hat{x}^{k+1}) \leq \eta \cdot r(x^k)$ from the inexactness condition (6). It then follows that

$$
\begin{aligned}
(1 - \eta) r(x^k) &\leq r(x^k) - r_k(\hat{x}^{k+1}) \\
&= r_k(x^k) - r_k(\hat{x}^{k+1}) \\
&\leq (\lambda_{\max}(H_k) + 2) \|d^k\| \\
&\leq (L_1 + \mu_k + 2) \|d^k\|,
\end{aligned}
\tag{12}
$$

where the first equality follows from $r_k(x^k) = r(x^k)$, the second inequality is by Corollary 2, and the last inequality is due to Fact 1. By combining (11) and (12) with Lemma 2, we obtain

$$\lim_{k \to \infty} \frac{\mu_k^2}{(L_1 + \mu_k + 2)^2} r(x^k)^2 = 0.$$

Since $\mu_k = c \cdot r(x^k)^\rho$, we conclude that $\lim_{k \to \infty} r(x^k) = 0$. □

Although Proposition 2 and the Luo–Tseng EB property (5) together imply that $\text{dist}(x^k, \mathcal{X}) \to 0$, the latter does not guarantee the convergence of the sequence $\{x^k\}_{k \geq 0}$. Hence, some extra arguments are needed to establish the global convergence of IRPN. As it turns out, by invoking the Luo–Tseng EB property (5), we can bound the local rate at which the sequence $\{\text{dist}(x^k, \mathcal{X})\}_{k \geq 0}$ tends to zero, from which we can establish not only the convergence but also the local rate of convergence of the sequence $\{x^k\}_{k \geq 0}$. To begin, let us present the following result, whose proof can be found in Sect. 4.2.

Proposition 3 *Suppose that Assumptions* 1 *and* 2 *hold. Then, for sufficiently large k, we have*

(i) $\mathrm{dist}(x^{k+1}, \mathcal{X}) \leq \gamma \cdot \mathrm{dist}(x^k, \mathcal{X})$ *for some* $\gamma \in (0, 1)$ *if we take* $\rho = 0$, $c \leq \frac{\kappa}{4}$, *and* $\eta \leq \frac{1}{2(L_1+3)^2}$;

(ii) $\mathrm{dist}(x^{k+1}, \mathcal{X}) \leq O(\mathrm{dist}(x^k, \mathcal{X})^{1+\rho})$ *if we take* $\rho \in (0, 1)$;

(iii) $\mathrm{dist}(x^{k+1}, \mathcal{X}) \leq O(\mathrm{dist}(x^k, \mathcal{X})^2)$ *if we take* $\rho = 1$, $c \geq \kappa L_2$, *and* $\eta \leq \frac{\kappa^2 L_2}{2(L_1+1)}$.

As remarked previously, the setting of Proposition 3 precludes the use of techniques that are based on either the positive definiteness of the Hessian of f at an optimal solution or the uniqueness of the optimal solution to prove it. Our techniques, which heavily exploit the Luo–Tseng EB property (5), are different from those in [4,17] and should find further applications in the convergence rate analysis of other second-order methods in the absence of strong convexity.

Using Proposition 3, we can then prove that the sequence of iterates generated by IRPN converges to an optimal solution to problem (1) and establish the local rate of convergence.

Theorem 1 (Global convergence and local rate of convergence of IRPN) *Suppose that Assumptions* 1 *and* 2 *hold. Then, in all three cases of Proposition* 3, *the sequence* $\{x^k\}_{k \geq 0}$ *converges to some* $x^* \in \mathcal{X}$. *Moreover, the convergence rate is at least*

(i) *R-linear if* $\rho = 0$, $c \leq \frac{\kappa}{4}$, *and* $\eta \leq \frac{1}{2(L_1+2)^2}$;

(ii) *Q-superlinear with order* $1 + \rho$ *if* $\rho \in (0, 1)$;

(iii) *Q-quadratic if* $\rho = 1$, $c \geq \kappa L_2$, *and* $\eta \leq \frac{\kappa^2 L_2}{2(L_1+1)}$.

Remarkably, Theorem 1 shows that IRPN can attain a superlinear or even quadratic rate of convergence without strong convexity. The proof of Theorem 1 can be found in Sect. 4.3.

Although Theorem 1 shows that a larger ρ will result in a faster convergence rate with respect to the outer iteration counter k, it does not necessarily mean that a larger ρ will lead to a better overall complexity. The reason is that a larger ρ results in a more stringent inexactness condition (see (6)) and thus the time consumed by each iteration is longer. Hence, despite the faster convergence rate, the best choice of $\rho \in [0, 1]$ depends on the problem and the inner solver (see also Sect. 5). Nonetheless, Theorem 1 provides a complete characterization of the convergence rate of IRPN in terms of ρ and facilitates the flexibility of IRPN. Developing an empirical or analytic approach to tuning ρ is definitely a topic worth pursuing.

Since Proposition 3 plays a crucial role in obtaining our convergence results and its proof is rather tedious, let us give an overview of the proof here. Let $\bar{x}^k$ be the projection of x^k onto the optimal solution set $\mathcal{X}$. Also, let $\tilde{x}^{k+1}$ and $\hat{x}^{k+1}$ be the exact and approximate minimizers of the quadratic model q_k, respectively; see Fig. 1. Our goal is to show that $\mathrm{dist}(x^{k+1}, \mathcal{X}) = O(\mathrm{dist}(x^k, \mathcal{X})^{1+\rho})$. Towards that end, we first prove that the distances $\|\tilde{x}^{k+1} - x^k\|$ and $\|\tilde{x}^{k+1} - \hat{x}^{k+1}\|$ are comparable, in the sense that they are both of $O(\mathrm{dist}(x^k, \mathcal{X}))$; see Lemmas 4 and 5. This implies that $\|\hat{x}^{k+1} - x^k\|$ is also of $O(\mathrm{dist}(x^k, \mathcal{X}))$. Then, we show that eventually the step sizes computed by Algorithm 1 will all equal 1, which means that $x^{k+1} = \hat{x}^{k+1}$ for all
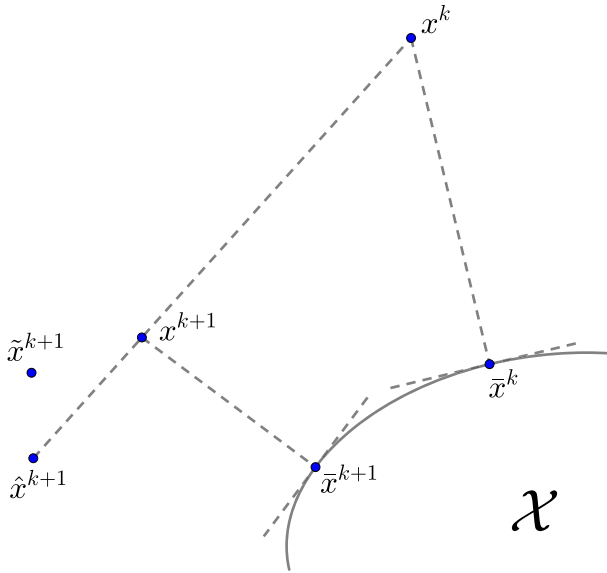
Fig. 1 Illustration of the geometry in the proof of Proposition 3

sufficiently large k; see Lemma 6. As a result, we have $\|x^{k+1} - x^k\| = O(\text{dist}(x^k, \mathcal{X}))$ for all sufficiently large k. Finally, we show that $r(x^{k+1}) = O(\|x^{k+1} - x^k\|^{1+\rho})$ and invoke the Luo–Tseng EB property (5) to conclude that

$$\text{dist}(x^{k+1}, \mathcal{X}) \le \kappa \cdot r(x^{k+1}) = O(\|\hat{x}^{k+1} - x^k\|^{1+\rho}) = O(\text{dist}(x^k, \mathcal{X})^{1+\rho}).$$

The rest of this section is devoted to proving Proposition 3 and Theorem 1. We first prove some technical lemmas in Sect. 4.1. Then, we prove Proposition 3 and Theorem 1 in Sects. 4.2 and 4.3, respectively.

4.1 Technical lemmas

Throughout this subsection, Assumptions 1 and 2 are in force. Recall that $\bar{x}^k$ is the projection of x^k onto the optimal solution set $\mathcal{X}$; $\tilde{x}^{k+1}$ and $\hat{x}^{k+1}$ are the exact and approximate minimizers of the quadratic model q_k, respectively.

Lemma 3 *It holds for all k that*

$$\|\tilde{x}^{k+1} - \bar{x}^k\| \le \frac{1}{\mu_k} \|\nabla f(\bar{x}^k) - \nabla f(x^k) - H_k(\bar{x}^k - x^k)\|.$$

Proof If $\tilde{x}^{k+1} = \bar{x}^k$, then the inequality holds trivially. Hence, suppose that $\tilde{x}^{k+1} \ne \bar{x}^k$. As $\bar{x}^k \in \mathcal{X}$, we have $\mathbf{0} \in \nabla f(\bar{x}^k) + \partial g(\bar{x}^k)$. Moreover, the definition of $\tilde{x}^{k+1}$ yields $\mathbf{0} \in \nabla f(x^k) + H_k(\tilde{x}^{k+1} - x^k) + \partial g(\tilde{x}^{k+1})$. Using these and the monotonicity of the subdifferential mapping ∂g of the closed proper convex function g, we obtain

$$0 \leq \left\langle \nabla f(x^k) - \nabla f(\bar{x}^k) + H_k(\tilde{x}^{k+1} - x^k), \bar{x}^k - \tilde{x}^{k+1} \right\rangle$$

$$= \left\langle \nabla f(x^k) - \nabla f(\bar{x}^k) - H_k(x^k - \bar{x}^k), \bar{x}^k - \tilde{x}^{k+1} \right\rangle$$

$$- \left\langle H_k(\bar{x}^k - \tilde{x}^{k+1}), \bar{x}^k - \tilde{x}^{k+1} \right\rangle$$

$$\leq \|\nabla f(x^k) - \nabla f(\bar{x}^k) - H_k(x^k - \bar{x}^k)\| \cdot \|\bar{x}^k - \tilde{x}^{k+1}\| - \mu_k \|\bar{x}^k - \tilde{x}^{k+1}\|^2,$$

which yields the desired inequality. $\qquad\square$

We remark here that we did not use any special structure of H_k in the proof of Lemma 3 except the positive definiteness of H_k. Therefore, the same result will hold for any positive definite approximate Hessian H_k, with μ_k replaced by the minimum eigenvalue $\lambda_{\min}(H_k)$ of H_k.

Lemma 4 *It holds for all k that*

$$\|\tilde{x}^{k+1} - x^k\| \leq \left(\frac{L_2}{2\mu_k} \mathrm{dist}(x^k, \mathcal{X}) + 2 \right) \mathrm{dist}(x^k, \mathcal{X}).$$

Proof Using the triangle inequality, the definition of μ_k, and Assumption 1(b), we compute

$$\|\nabla f(\bar{x}^k) - \nabla f(x^k) - H_k(\bar{x}^k - x^k)\|$$
$$\leq \|\nabla f(\bar{x}^k) - \nabla f(x^k) - \nabla^2 f(x^k)(\bar{x}^k - x^k)\| + \mu_k \|\bar{x}^k - x^k\|$$
$$\leq \frac{L_2}{2} \|\bar{x}^k - x^k\|^2 + \mu_k \|\bar{x}^k - x^k\|.$$

This, together with Lemma 3, yields

$$\|\tilde{x}^{k+1} - x^k\| \leq \|\tilde{x}^{k+1} - \bar{x}^k\| + \mathrm{dist}(x^k, \mathcal{X})$$

$$\leq \frac{1}{\mu_k} \|\nabla f(\bar{x}^k) - \nabla f(x^k) - H_k(\bar{x}^k - x^k)\| + \mathrm{dist}(x^k, \mathcal{X})$$

$$\leq \frac{L_2}{2\mu_k} \|\bar{x}^k - x^k\|^2 + \|\bar{x}^k - x^k\| + \mathrm{dist}(x^k, \mathcal{X})$$

$$= \left(\frac{L_2}{2\mu_k} \mathrm{dist}(x^k, \mathcal{X}) + 2 \right) \mathrm{dist}(x^k, \mathcal{X}),$$

as desired. $\qquad\square$

Lemma 5 *It holds for all k that*

$$\|\hat{x}^{k+1} - \tilde{x}^{k+1}\| \leq \frac{\eta(L_1 + |1 - \mu_k|)}{c} r(x^k) + \eta \cdot r(x^k)^{1+\rho}.$$

Proof The inequality holds trivially if $\hat{x}^{k+1} = \tilde{x}^{k+1}$. Hence, we assume that $\hat{x}^{k+1} \neq \tilde{x}^{k+1}$. Recall that $R_k(\hat{x}^{k+1}) = \hat{x}^{k+1} - \mathrm{prox}_g(\hat{x}^{k+1} - \nabla f_k(\hat{x}^{k+1}))$. Thus,

$$R_k(\hat{x}^{k+1}) - \nabla f_k(\hat{x}^{k+1}) \in \partial g\left(\hat{x}^{k+1} - R_k(\hat{x}^{k+1})\right).$$

Since $q_k = f_k + g$, we have

$$R_k(\hat{x}^{k+1}) + \nabla f_k\left(\hat{x}^{k+1} - R_k(\hat{x}^{k+1})\right) - \nabla f_k(\hat{x}^{k+1}) \in \partial q_k\left(\hat{x}^{k+1} - R_k(\hat{x}^{k+1})\right).$$

Using the fact that $\nabla f_k(x) = \nabla f(x^k) + H_k(x - x^k)$, the above leads to

$$(I - H_k)R_k(\hat{x}^{k+1}) \in \partial q_k(\hat{x}^{k+1} - R_k(\hat{x}^{k+1})).$$

On the other hand, we have $\mathbf{0} \in \partial q_k(\tilde{x}^{k+1})$. Since q_k is μ_k-strongly convex, the subdifferential mapping ∂q_k is μ_k-strongly monotone and thus

$$\left\langle (I - H_k)R_k(\hat{x}^{k+1}), \hat{x}^{k+1} - R_k(\hat{x}^{k+1}) - \tilde{x}^{k+1} \right\rangle \geq \mu_k \|\hat{x}^{k+1} - R_k(\hat{x}^{k+1}) - \tilde{x}^{k+1}\|^2.$$

Upon applying the Cauchy–Schwarz inequality to the above, we obtain

$$\begin{aligned}
\|\hat{x}^{k+1} - R_k(\hat{x}^{k+1}) - \tilde{x}^{k+1}\| &\leq \frac{1}{\mu_k}\left\|(I - H_k)R_k(\hat{x}^{k+1})\right\| \\
&\leq \frac{1}{\mu_k}\|\nabla^2 f(x^k) - (1 - \mu_k)I\| \cdot r_k(\hat{x}^{k+1}) \\
&\leq \frac{\eta(L_1 + |1 - \mu_k|)}{\mu_k} r(x^k)^{1+\rho} \\
&\leq \frac{\eta(L_1 + |1 - \mu_k|)}{c} r(x^k),
\end{aligned}$$

where the third inequality is due to Fact 1 and the inexactness condition (6), and the last is due to the definition of μ_k. Hence, we obtain

$$\begin{aligned}
\|\hat{x}^{k+1} - \tilde{x}^{k+1}\| &\leq \|\hat{x}^{k+1} - R_k(\hat{x}^{k+1}) - \tilde{x}^{k+1}\| + \|R_k(\hat{x}^{k+1})\| \\
&\leq \frac{\eta(L_1 + |1 - \mu_k|)}{c} r(x^k) + \eta \cdot r(x^k)^{1+\rho},
\end{aligned}$$

as desired. $\qquad\square$

Next, we show that Algorithm 1 eventually takes a unit step size; i.e., $\alpha_k = 1$ and $x^{k+1} = x^k + d^k$ for all sufficiently large k.

Lemma 6 *There exists an integer $k_0 \geq 0$ such that $\alpha_k = 1$ for all $k \geq k_0$ in either of the following cases:*

 (i) $\rho \in [0, 1)$.
(ii) $\rho = 1$ and c, η satisfy $2\eta L_2(L_1 + 2) + \kappa^2 L_2^2 + 2c\kappa L_2 \leq 6c^2$.

The inequality in (ii) is satisfied if we take $c \geq \kappa L_2$ and $\eta \leq \frac{\kappa^2 L_2}{2(L_1+1)}$.

Proof By the Fundamental Theorem of Calculus and Assumption 1(b), we compute

$$
\begin{aligned}
&f(\hat{x}^{k+1}) - f(x^k) \\
&= \int_0^1 \nabla f\left(x^k + t(\hat{x}^{k+1} - x^k)\right)^T (\hat{x}^{k+1} - x^k)\, dt \\
&= \int_0^1 \left[\nabla f\left(x^k + t(\hat{x}^{k+1} - x^k)\right) - \nabla f(x^k)\right]^T (\hat{x}^{k+1} - x^k)\, dt \\
&\quad + \nabla f(x^k)^T (\hat{x}^{k+1} - x^k) \\
&= \int_0^1 \int_0^1 t(\hat{x}^{k+1} - x^k)^T \left[\nabla^2 f\left(x^k + st(\hat{x}^{k+1} - x^k)\right) - \nabla^2 f(x^k)\right](\hat{x}^{k+1} - x^k)\, ds\, dt \\
&\quad + \int_0^1 t(\hat{x}^{k+1} - x^k)^T \nabla^2 f(x^k)(\hat{x}^{k+1} - x^k)\, dt + \nabla f(x^k)^T \left(\hat{x}^{k+1} - x^k\right) \\
&\leq \frac{L_2}{6}\|\hat{x}^{k+1} - x^k\|^3 + \frac{1}{2}(\hat{x}^{k+1} - x^k)^T \nabla^2 f(x^k)(\hat{x}^{k+1} - x^k) + \nabla f(x^k)^T(\hat{x}^{k+1} - x^k).
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
&F(\hat{x}^{k+1}) - F(x^k) + q_k(x^k) - q_k(\hat{x}^{k+1}) \\
&\quad = f(\hat{x}^{k+1}) - f(x^k) - \nabla f(x^k)^T (\hat{x}^{k+1} - x^k) \\
&\qquad - \frac{1}{2}(\hat{x}^{k+1} - x^k)^T \nabla^2 f(x^k)(\hat{x}^{k+1} - x^k) - \frac{\mu_k}{2}\|\hat{x}^{k+1} - x^k\|^2 \\
&\quad \leq \frac{L_2}{6}\|\hat{x}^{k+1} - x^k\|^3 - \frac{\mu_k}{2}\|\hat{x}^{k+1} - x^k\|^2.
\end{aligned}
\tag{13}
$$

It then follows from (6b), (9), and (13) that

$$
\begin{aligned}
&F(\hat{x}^{k+1}) - F(x^k) \\
&= \left(F(\hat{x}^{k+1}) - F(x^k) + q_k(x^k) - q_k(\hat{x}^{k+1})\right) + \left(q_k(\hat{x}^{k+1}) - q_k(x^k)\right) \\
&\leq \frac{L_2}{6}\|d^k\|^3 - \frac{\mu_k}{2}\|d^k\|^2 + \zeta\left(\ell_k(\hat{x}^{k+1}) - \ell_k(x^k)\right) \\
&= \frac{L_2}{6}\|d^k\|^3 - \frac{\mu_k}{2}\|d^k\|^2 + \theta\left(\ell_k(\hat{x}^{k+1}) - \ell_k(x^k)\right) + (\zeta - \theta)\left(\ell_k(\hat{x}^{k+1}) - \ell_k(x^k)\right) \\
&\leq \frac{L_2}{6}\|d^k\|^3 + \theta\left(\ell_k(\hat{x}^{k+1}) - \ell_k(x^k)\right) - \left(1 + \frac{\zeta - \theta}{1 - \zeta}\right)\frac{\mu_k}{2}\|d^k\|^2.
\end{aligned}
$$

In particular, if

$$
\|d^k\| = \|\hat{x}^{k+1} - x^k\| \leq \left(1 + \frac{\zeta - \theta}{1 - \zeta}\right)\frac{3c \cdot r(x^k)^\rho}{L_2},
\tag{14}
$$

then $F(\hat{x}^{k+1}) - F(x^k) \leq \theta\left(\ell_k(\hat{x}^{k+1}) - \ell_k(x^k)\right)$, which, according to the line search strategy in Algorithm 1, would lead to a unit step size $\alpha_k = 1$. Hence, it remains to determine for what choices of the parameters c, η would inequality (14) hold. By Lemmas 4 and 5, for all sufficiently large k,

$$\|\hat{x}^{k+1} - x^k\| \leq \frac{\eta(L_1 + |1 - \mu_k|)}{c} r(x^k) + \eta \cdot r(x^k)^{1+\rho}$$
$$+ \left(\frac{L_2}{2c \cdot r(x^k)^\rho} \text{dist}(x^k, \mathcal{X}) + 2 \right) \text{dist}(x^k, \mathcal{X})$$
$$\leq \frac{\eta(L_1 + |1 - \mu_k|)}{c} r(x^k) + \eta \cdot r(x^k)^{1+\rho}$$
$$+ \frac{\kappa^2 L_2}{2c} r(x^k)^{2-\rho} + 2\kappa \cdot r(x^k).$$

Hence, a sufficient condition for inequality (14) to hold is

$$\frac{\eta(L_1 + |1 - \mu_k|)}{c} r(x^k) + \eta \cdot r(x^k)^{1+\rho} + \frac{\kappa^2 L_2}{2c} r(x^k)^{2-\rho} + 2\kappa \cdot r(x^k)$$
$$\leq \left(1 + \frac{\zeta - \theta}{1 - \zeta} \right) \frac{3c \cdot r(x^k)^\rho}{L_2}. \tag{15}$$

If $\rho \in [0, 1)$, then since $r(x^k) \to 0$, we see that inequality (15) holds for all sufficiently large k. On the other hand, if $\rho = 1$, then since $\zeta > \theta$ and $\mu_k \in (0, 1)$ for all sufficiently large k, inequality (15) holds for all sufficiently large k if

$$\frac{\eta(L_1 + 1)}{c} + \frac{\kappa^2 L_2}{2c} + 2\kappa \leq \frac{3c}{L_2},$$

which can be satisfied by taking $c \geq \kappa L_2$ and $\eta \leq \frac{\kappa^2 L_2}{2(L_1+1)}$. □

4.2 Proof of Proposition 3

By Lemma 6, in all three cases of Proposition 3, we have $\alpha_k = 1$ and hence $\hat{x}^{k+1} = x^{k+1}$ for all sufficiently large k. Consequently, all the results derived in Sect. 4.1 will continue to hold if we replace $\hat{x}^{k+1}$ by x^{k+1} when k is sufficiently large. From Proposition 1, we have

$$\|x^{k+1} - x^k\| \leq \frac{\eta(L_1 + |1 - \mu_k|)}{c} r(x^k) + \eta \cdot r(x^k)^{1+\rho}$$
$$+ \left(\frac{L_2}{2c \cdot r(x^k)^\rho} \text{dist}(x^k, \mathcal{X}) + 2 \right) \text{dist}(x^k, \mathcal{X})$$
$$\leq \left(\frac{\eta(L_1 + |1 - \mu_k|)(L_1 + 2)}{c} + \eta(L_1 + 2)r(x^k)^\rho \right.$$
$$+ \left. \frac{L_2}{2c \cdot r(x^k)^\rho} \text{dist}(x^k, \mathcal{X}) + 2 \right) \text{dist}(x^k, \mathcal{X}). \tag{16}$$

By the Luo–Tseng EB property (5), we have $\text{dist}(x, \mathcal{X}) \leq \kappa \cdot r(x)$ whenever $r(x)$ is sufficiently small. Since $r(x^k) \to 0$, we see from (16) that for all sufficiently large k,

$$\|x^{k+1} - x^k\| = O(\text{dist}(x^k, \mathcal{X})).$$

Using the non-expansiveness of prox_g and the fact that $\nabla f_k(x^{k+1}) = \nabla f(x^k) + H_k(x^{k+1} - x^k)$, we have

$$
\begin{aligned}
\|\text{prox}_g(x^{k+1} &- \nabla f(x^{k+1})) - \text{prox}_g(x^{k+1} - \nabla f_k(x^{k+1}))\| \\
&\leq \|\nabla f(x^{k+1}) - \nabla f(x^k) - H_k(x^{k+1} - x^k)\|.
\end{aligned}
\tag{17}
$$

It follows that for all sufficiently large k,

$$
\begin{aligned}
\text{dist}(x^{k+1}, \mathcal{X}) &\leq \kappa \cdot r(x^{k+1}) \\
&\leq \kappa \left(\|R(x^{k+1}) - R_k(x^{k+1})\| + \|R_k(x^{k+1})\| \right) \\
&\leq \kappa \|\text{prox}_g(x^{k+1} - \nabla f(x^{k+1})) - \text{prox}_g(x^{k+1} - \nabla f_k(x^{k+1}))\| \\
&\quad + \kappa\eta \cdot r(x^k)^{1+\rho} \\
&\leq \kappa \|\nabla f(x^{k+1}) - \nabla f(x^k) - H_k(x^{k+1} - x^k)\| + \kappa\eta \cdot r(x^k)^{1+\rho} \\
&\leq \frac{\kappa L_2}{2} \|x^{k+1} - x^k\|^2 + \kappa\mu_k \|x^{k+1} - x^k\| + \kappa\eta \cdot r(x^k)^{1+\rho} \\
&\leq \frac{\kappa L_2}{2} \|x^{k+1} - x^k\|^2 + c\kappa(L_1 + 2)^\rho \text{dist}(x^k, \mathcal{X})^\rho \|x^{k+1} - x^k\| \\
&\quad + \kappa\eta(L_1 + 2)^{1+\rho} \text{dist}(x^k, \mathcal{X})^{1+\rho} \\
&\leq O(\text{dist}(x^k, \mathcal{X})^{1+\rho}),
\end{aligned}
$$

(18)

(19)

where the third inequality is due to the inexactness condition (6), the fourth follows from (17), the sixth is due to the definition of μ_k and Proposition 1, and the last follows from (16).

For the case where $\rho = 0$, in order to establish the linear convergence of the sequence $\{\text{dist}(x^k, \mathcal{X})\}_{k \geq 0}$, we need the constant in the big-O notation in (19) to be strictly less than 1. Upon inspecting (16) and (18), we see that this can be guaranteed if

$$
\kappa\eta(L_1 + 2) + c\kappa \left(\frac{\eta(L_1 + |1 - c|)(L_1 + 2)}{c} + \eta(L_1 + 2) + 2 \right) < 1.
$$

The above inequality can be satisfied by taking $c \leq \frac{\kappa}{4}$ and $\eta \leq \frac{1}{2(L_1+3)^2}$. $\qquad \square$

4.3 Proof of Theorem 1

By the Luo–Tseng EB property (5) and Proposition 2, we have $\text{dist}(x^k, \mathcal{X}) \to 0$. This implies that in all three cases of Proposition 3, there exists an integer $K_1 \geq 0$ such that $\text{dist}(x^{k+1}, \mathcal{X}) \leq \gamma \cdot \text{dist}(x^k, \mathcal{X})$ for all $k \geq K_1$, where $\gamma \in (0, 1)$ is the constant in case (i) of Proposition 3. In addition, by (16), there exist a real number $\sigma > 0$ and an integer $K_2 \geq 0$ such that $\|x^{k+1} - x^k\| \leq \sigma \cdot \text{dist}(x^k, \mathcal{X})$ for all $k \geq K_2$. Lastly,

since $\text{dist}(x^k, \mathcal{X}) \to 0$, given any $\epsilon > 0$, there exists an integer $K_3 \geq 0$ such that $\text{dist}(x^k, \mathcal{X}) \leq (1 - \gamma)\epsilon/\sigma$ for all $k \geq K_3$. Upon taking $K = \max\{K_1, K_2, K_3\}$, we have

$$\|x^{k_1} - x^{k_0}\| \leq \sum_{i=k_0}^{k_1-1} \|x^{i+1} - x^i\| \leq \sigma \sum_{i=k_0}^{k_1-1} \text{dist}(x^i, \mathcal{X})$$

$$\leq \sigma \sum_{i=0}^{\infty} \gamma^i \cdot \text{dist}(x^{k_0}, \mathcal{X}) \leq \frac{\sigma}{1 - \gamma} \cdot \text{dist}(x^{k_0}, \mathcal{X}) \leq \epsilon$$

for all $k_1 \geq k_0 \geq K$. It follows that $\{x^k\}_{k \geq 0}$ is a Cauchy sequence and hence converges to some x^*. This, together with $\text{dist}(x^k, \mathcal{X}) \to 0$ and the fact that $\mathcal{X}$ is closed, implies that $x^* \in \mathcal{X}$.

Now, it remains to establish the convergence rate of the sequence $\{x^k\}_{k \geq 0}$ in cases (i), (ii), and (iii). From the above analysis, it is immediate that

$$\|x^{k+i} - x^{k+1}\| \leq \frac{\sigma}{1 - \gamma} \cdot \text{dist}(x^{k+1}, \mathcal{X})$$

for all integers $i \geq 1$ and $k \geq K$. Upon taking $i \to \infty$, we have

$$\|x^{k+1} - x^*\| \leq \frac{\sigma}{1 - \gamma} \cdot \text{dist}(x^{k+1}, \mathcal{X}). \tag{20}$$

In case (i), Proposition 3 implies that the sequence $\{\text{dist}(x^{k+1}, \mathcal{X})\}_{k \geq 0}$ converges R-linearly to 0. Hence, by (20), the sequence $\{x^k\}_{k \geq 0}$ converges R-linearly to x^*. In case (ii), using Proposition 3 and (20), we obtain

$$\|x^{k+1} - x^*\| = O(\text{dist}(x^k, \mathcal{X})^{1+\rho}) \leq O(\|x^k - x^*\|^{1+\rho}),$$

which implies that the sequence $\{x^k\}_{k \geq 0}$ converges Q-superlinearly to x^* with order $1 + \rho$. The same arguments show that the sequence $\{x^k\}_{k \geq 0}$ converges Q-quadratically to x^* in case (iii). This completes the proof. $\qquad\square$

5 Numerical experiments

In this section, we study the numerical performance of our proposed IRPN and compare it with some existing algorithms. We focus on the ℓ_1-regularized logistic regression problem, which takes the following form:

$$\min_{x \in \mathbb{R}^n} \left\{ F(x) = \underbrace{\frac{1}{m} \sum_{i=1}^{m} \log\left(1 + \exp(-b_i \cdot a_i^T x)\right)}_{f(x)} + \underbrace{\lambda \|x\|_1}_{g(x)} \right\}. \tag{21}$$

Table 1 Tested data sets

Data set	n	m	nnz
colon-cancer	2000	62	124,000
rcv1	47,236	20,242	1,498,952
news20	1,355,191	19,996	9,097,916

Here, $a_1, \ldots, a_m \in \mathbb{R}^n$ are given data samples; $b_1, \ldots, b_m \in \{1, -1\}$ are given labels; $\lambda > 0$ is a given regularization parameter. Problem (21) arises in linear classification tasks in machine learning and is a standard benchmark problem for testing the efficiency of different algorithms for solving problem (1). In our experiments, we use the data sets colon-cancer, rcv1, and news20 from [20] and set $\lambda = 5 \times 10^{-4}$. The sizes of and the numbers of non-zero entries in these data sets are listed in Table 1. Since $m < n$ in all three data sets, the corresponding objective functions F in (21) are not strongly convex. However, due to Fact 3, problem (21) possesses the Luo–Tseng EB property (5). Hence, both Assumptions 1 and 2 are satisfied by problem (21) and our convergence analysis of IRPN applies.

All experiments are coded in MATLAB (R2017b) and run on a Dell desktop with a 3.50-GHz Intel Core E3-1270 v3 processor and 32 GB of RAM.[6] We next present the list of tested algorithms and discuss their implementation details below.

FISTA: The description of this algorithm can be found in [1]. In our experiments, we use the constant step size $1/L_1$, where $L_1 = \|A\|^2/(4m)$ and $A = [a_1 \cdots a_m] \in \mathbb{R}^{n \times m}$.[7] It can be verified that L_1 is the Lipschitz constant of ∇f in (21). In addition, we restart the algorithm if $(y^{k-1} - x^k)^T(x^k - x^{k-1}) > 0$ for some k, where y^{k-1} is the extrapolation point in iteration k. Such restarting strategy has been found to be empirically efficient [28,40] and it indeed improves the performance of FISTA for solving problem (21) in our experiments.

SpaRSA: This algorithm is a non-monotone proximal gradient method, whose description can be found in [41]. Following the notations in [41], we set in our experiments $\sigma = 10^{-4}$, $\alpha_{\min} = 10^{-8}$, $\alpha_{\max} = 10^8$, $M = 5$, $\alpha_0 = 1$, and

$$\alpha_k = \max\left\{\alpha_{\min}, \min\left\{\alpha_{\max}, \frac{\|\Delta x\|^2}{|\Delta G^T \Delta x|}\right\}\right\}$$

for $k \geq 1$, where $\Delta x = x^k - x^{k-1}$ and $\Delta G = \nabla f(x^k) - \nabla f(x^{k-1})$.

CGD: This algorithm is the coordinate gradient descent method, whose general description can be found in [39]. It has been successfully applied to solve problem (21); see, e.g., [44,46]. Given the current iterate x^k, we randomly choose a coordinate index $i_k \in \{1, \ldots, n\}$ and compute a descent direction $d^k \in \mathbb{R}^n$ by letting $d_i^k = 0$ for all $i \neq i_k$ and

---

[6] The code can be downloaded from https://github.com/ZiruiZhou/IRPN.

[7] $\|A\|^2$ is computed via the MATLAB code lambda = eigs(A*A',1,'LM').

$$d_{i_k}^k = \underset{s \in \mathbb{R}}{\operatorname{argmin}} \left\{ \frac{1}{2} H_{i_k i_k} s^2 + g_{i_k} s + \lambda |x_{i_k}^k + s| \right\},$$

where $H_{i_k i_k} = [\nabla^2 f(x^k)]_{i_k i_k} + \nu$ with $\nu = 10^{-6}$ and $g_{i_k} = [\nabla f(x^k)]_{i_k}$. Then, we choose α_k to be the largest element in $\{\beta^j\}_{j \geq 0}$ satisfying

$$F(x^k + \alpha_k d^k) \leq F(x^k) + \sigma \alpha_k \left(\nabla f(x^k)^T d^k + \lambda \|x^k + \alpha_k d^k\|_1 - \lambda \|x^k\|_1 \right),$$

where we set $\beta = 0.25$ and $\sigma = 0.5$ in our experiments, and set $x^{k+1} = x^k + \alpha_k d^k$. To make the number of iterations of CGD comparable to those of FISTA and SpaRSA, we refer to every n coordinate updates as an iteration of CGD. In addition, while the operations in each iteration of FISTA and SpaRSA can be vectorized in MATLAB, those of CGD cannot and require a *for* loop, which is inefficient in MATLAB. To remedy this issue, we implement CGD in MATLAB as a C source MEX-file.

IRPN: This is the family of inexact SQA methods proposed in this paper, whose description can be found in Algorithm 1. In our experiments, we set $\theta = \beta = 0.25$, $\zeta = 0.4$, $c = 10^{-6}$, and $\eta = 0.5$. We test the three values 0, 0.5, and 1 for ρ, which, according to Theorem 1, result in the local linear, superlinear, and quadratic convergence of IRPN, respectively. Notice that in the context of problem (21), each inner problem of IRPN is to minimize the sum of a strongly convex quadratic function and the non-smooth convex function $\lambda \|x\|_1$. Hence, it can be solved using the coordinate descent algorithm with simple updates; see, e.g., [45]. Due to the same reason as CGD, we code the inner solver of IRPN in MATLAB as a C source MEX-file. Moreover, we refer to every n coordinate updates as an iteration of the inner solver; i.e., an inner iteration of IRPN.

newGLMNET: The description of this algorithm can be found in [45]. In our experiments, we implement it as a special case of IRPN. In particular, we set $c = 10^{-12}$ as suggested in [45] and $\rho = 0$ in Algorithm 1. Moreover, instead of the stopping criterion (6), newGLMNET uses a heuristic stopping criterion for the inner problem [45]. Specifically, in the k-th iteration, the inner solver is terminated when $r_k(\hat{x}^{k+1}) \leq \epsilon_{\text{in}}$, where $\epsilon_{\text{in}} = 0.1$ initially and $\epsilon_{\text{in}} = \epsilon_{\text{in}}/4$ if the stopping criterion is satisfied after only one inner iteration.

It is worth mentioning the computation of the Hessian-vector product for the logistic loss function, which is explained in detail in [45] and is crucial to the success of Newton-type methods for solving problem (21). For the logistic loss function f in (21), one can verify by a direct computation that $\nabla^2 f(x) = ADA^T$, where $A = [a_1 \cdots a_m] \in \mathbb{R}^{n \times m}$ and D is a diagonal matrix with

$$D_{ii} = \frac{\exp(-b_i \cdot a_i^T x)}{\left(\exp(-b_i \cdot a_i^T x) + 1 \right)^2}, \quad i = 1, \ldots, m.$$

Notice that A is sparse in many real data sets such as colon-cancer, rcv1, and news20. Hence, $\nabla^2 f(x)v$ for any $v \in \mathbb{R}^n$ can be evaluated efficiently in our experiments.

We initialize all the tested algorithms by the same point. In order to see the impact of different initial points on the performance of these algorithms, we choose $x^0 = 0$

Table 2 Computational results on colon-cancer

Tol.	Index	FISTA*	SpaRSA	CGD	newG.	IRPN $\rho = 0$	IRPN $\rho = 0.5$	IRPN $\rho = 1$
(a) $x^0 = 0$								
10^{-4}	Outer iter.	–	–	–	10	6	4	4
	Inner iter.	904	104	18	33	26	37	87
	Time	0.41	0.03	0.08	0.03	**0.02**	0.03	0.06
10^{-6}	Outer iter.	–	–	–	19	14	5	5
	Inner iter.	3856	397	55	92	84	85	183
	Time	1.78	0.13	0.32	**0.06**	**0.06**	**0.06**	0.13
10^{-8}	Outer iter.	–	–	–	25	24	6	6
	Inner iter.	5202	756	100	144	162	142	273
	Time	2.30	0.23	0.40	**0.08**	0.14	0.10	0.18
(b) $x^0 = 10\xi$								
10^{-4}	Outer iter.	–	–	–	14	11	10	7
	Inner iter.	1185	124	33	39	39	110	126
	Time	0.43	0.04	0.13	**0.03**	**0.03**	0.09	0.19
10^{-6}	Outer iter.	–	–	–	25	17	9	8
	Inner iter.	3536	422	71	101	85	113	268
	Time	1.50	0.12	0.32	0.08	**0.06**	0.09	0.27
10^{-8}	Outer iter.	–	–	–	25	24	6	6
	Inner iter.	4932	727	100	173	175	118	373
	Time	2.18	0.25	0.43	0.11	0.13	**0.08**	0.30

*The time for computing the Lipschitz constant L_1 in FISTA is 0.04

and $x^0 = 10\xi$, where the entries of ξ are sampled *i.i.d.* from the standard normal distribution. All the algorithms are terminated if the iterate x^k satisfies $r(x^k) \leq \epsilon_0$, where $\epsilon_0 = 10^{-4}, 10^{-6}$, and 10^{-8} in our experiments.

The computational results on the data sets colon-cancer, rcv1, and news20 are presented in Tables 2, 3, and 4, respectively. Each table has two sub-tables, which show the results with the two initial points $x^0 = 0$ and $x^0 = 10\xi$, respectively. In each sub-table, we record the number of outer iterations, the number of inner iterations, and the CPU time (in seconds) of all the tested algorithms for achieving the required accuracy. We note that FISTA is the only one among these algorithms that requires the Lipschitz constant L_1. The time for computing L_1 is added to the time of FISTA and can be found at the bottom of each table. From the bolded entries in Tables 2, 3, and 4, we can see that the SQA methods (newGLMNET and IRPN) perform the best among all the tested algorithms. In particular, we have the following observations:

(a) The SQA methods substantially outperform FISTA.
(b) SpaRSA and CGD are competitive with the SQA methods when $\epsilon_0 = 10^{-4}$, while the SQA methods outperform SpaRSA and CGD when $\epsilon_0 = 10^{-6}$ and 10^{-8}.

Table 3 Computational results on rcv1

Tol.	Index	FISTA*	SpaRSA	CGD	newG.	IRPN $\rho = 0$	IRPN $\rho = 0.5$	IRPN $\rho = 1$
(a) $x^0 = 0$								
10^{-4}	Outer iter.	–	–	–	9	5	3	3
	Inner iter.	113	30	8	20	19	20	34
	Time	39.13	0.29	0.46	0.32	0.27	**0.26**	0.42
10^{-6}	Outer iter.	–	–	–	17	11	5	5
	Inner iter.	377	280	35	55	48	36	106
	Time	45.42	2.60	2.05	0.77	0.64	**0.48**	1.21
10^{-8}	Outer iter.	–	–	–	23	16	5	5
	Inner iter.	697	549	61	90	77	61	126
	Time	46.67	5.08	3.51	1.18	1.03	**0.70**	1.32
(b) $x^0 = 10\xi$								
10^{-4}	Outer iter.	–	–	–	12	8	5	5
	inner iter.	308	50	15	25	23	27	44
	time	41.58	0.48	0.88	0.41	**0.35**	0.37	0.57
10^{-6}	Outer iter.	–	–	–	19	12	7	7
	Inner iter.	595	286	47	54	47	57	112
	Time	45.25	2.65	2.77	0.77	**0.66**	0.68	1.24
10^{-8}	Outer iter.	–	–	–	26	19	8	8
	Inner iter.	1009	591	67	101	87	97	141
	Time	51.31	5.47	3.85	1.34	1.17	**1.14**	1.54

*The time for computing the Lipschitz constant L_1 in FISTA is 36.14

(c) As the parameter ρ in IRPN increases from 0 to 1, the number of outer iterations needed by IRPN decreases, while the number of inner iterations within each outer iteration increases. The overall performance of $\rho = 0$ is competitive with that of $\rho = 0.5$ and better than that of $\rho = 1$.

It is worth noting that the number of inner iterations needed by CGD is consistently the least among the tested algorithms. However, since it requires a line search procedure after every coordinate update, its overall performance is worse than that of the SQA methods, which only require one line search procedure in each outer iteration. Such drawback of CGD is also noted in [45].

Before we end this section, we present in Fig. 2 the convergence behavior of IRPN with $\rho \in \{0, 0.5, 1\}$ for the data sets rcv1 and news20. The left two sub-figures show the convergence of the sequence of residues $\{r(x^k)\}_{k \geq 0}$ against the number of outer iterations, where $\{x^k\}_{k \geq 0}$ is the sequence of iterates generated by IRPN. It can be observed that the sequence $\{r(x^k)\}_{k \geq 0}$ exhibits a linear convergence rate when $\rho = 0$ and a superlinear rate when $\rho = 0.5$ and 1. This, together with the Luo–Tseng EB property (5), implies that the sequence $\{\text{dist}(x^k, \mathcal{X})\}_{k \geq 0}$ converges at least linearly when $\rho = 0$ and at least superlinearly when $\rho = 0.5$ and 1, which corroborates our

Table 4 Computational results on news20

Tol.	Index	FISTA*	SpaRSA	CGD	newG.	IRPN $\rho = 0$	IRPN $\rho = 0.5$	IRPN $\rho = 1$
(a) $x^0 = 0$								
10^{-4}	Outer iter.	–	–	–	9	5	3	3
	Inner iter.	182	48	33	41	37	54	127
	Time	77.68	5.63	12.13	4.92	**4.16**	5.68	12.87
10^{-6}	Outer iter.	–	–	–	17	12	4	4
	Inner iter.	566	220	98	115	100	107	223
	Time	131.01	25.31	36.21	12.58	**10.99**	10.99	22.40
10^{-8}	Outer iter.	–	–	–	23	18	4	4
	Inner iter.	978	580	155	189	170	157	248
	Time	188.77	67.62	57.23	20.15	18.51	**16.36**	24.84
(b) $x^0 = 10\xi$								
10^{-4}	Outer iter.	–	–	–	13	7	6	5
	Inner iter.	647	296	57	76	67	89	95
	Time	150.79	35.31	21.10	7.74	**7.40**	9.06	9.60
10^{-6}	Outer iter.	–	–	–	20	14	7	6
	Inner iter.	1055	302	108	166	154	155	179
	Time	207.24	35.88	40.04	16.41	16.17	**15.41**	17.54
10^{-8}	Outer iter.	–	–	–	26	21	7	7
	Inner iter.	1430	946	182	212	189	191	286
	Time	258.89	108.36	67.34	21.03	19.72	**18.72**	27.77

*The time for computing the Lipschitz constant L_1 in FISTA is 51.85

results in Proposition 3. Nonetheless, a faster convergence rate of the outer iterates does not imply a better overall performance of IRPN, as a larger ρ leads to a more stringent inexactness condition (see (6)) and hence requires more inner iterations within each outer iteration. With that being said, we present the right two sub-figures in Fig. 2 for comparing the overall performance of IRPN with $\rho \in \{0, 0.5, 1\}$. In particular, we show the convergence behavior of the sequence of residues $\{r(x^k)\}_{k \geq 0}$ against the CPU time used, where $\{x^k\}_{k \geq 0}$ is the sequence of *all* iterates (including both the inner and outer ones) generated by IRPN. It can be observed that in terms of overall performance, IRPN with $\rho = 0$ and 0.5 are comparable to each other, and they both outperform IRPN with $\rho = 1$.

6 Conclusions

In this paper, we proposed a new family of inexact SQA methods called the *inexact regularized proximal Newton* (IRPN) method for minimizing the sum of two closed proper convex functions, one of which is smooth and the other is possibly non-smooth. Compared with some prior SQA methods, IRPN is more flexible and has a much more
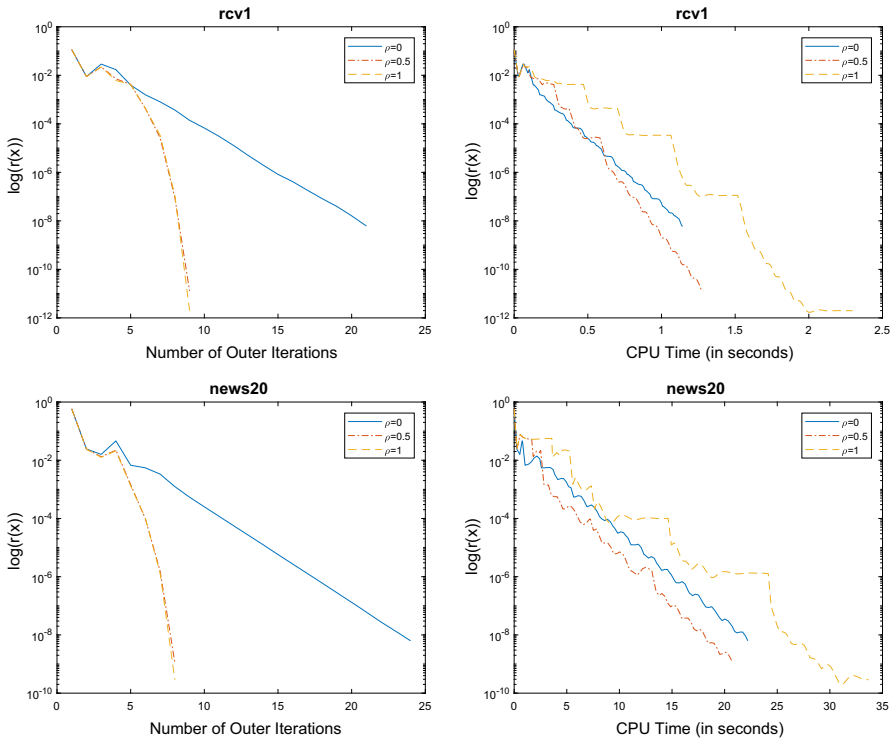
Fig. 2 Convergence behavior of IRPN with $\rho \in \{0, 0.5, 1\}$

satisfying convergence theory. In particular, IRPN does not require an exact inner solver or the strong convexity of the smooth part of the objective function. Moreover, it can be shown to converge globally to an optimal solution and the local convergence rate is linear, superlinear, or even quadratic, depending on the choice of parameters of the algorithm. The key to our analysis is the Luo–Tseng error bound property. Although such property has played a fundamental role in establishing the linear convergence of various first-order methods, to the best of our knowledge, this is the first work to exploit such property to establish the superlinear convergence of SQA-type methods for non-smooth convex minimization. We compared our proposed IRPN with several popular and efficient algorithms by applying them to the ℓ_1-regularized logistic regression problem. Experiment results indicate that IRPN achieves the desired levels of accuracy in shorter training time and less number of iterations than the other algorithms. This shows that IRPN not only has strong theoretical guarantees but also a superior numerical performance.

References

1. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)
2. Becker, S., Fadili, J.: A quasi-Newton proximal splitting method. In Pereira, F.C.N., Burges, C.J.C., Bottou, L., Weinberger, K. Q. (eds), Advances in Neural Information Processing Systems 25: Proceedings of the 2012 Conference, pp. 2618–2626 (2012)
3. Bhatia, R.: Matrix Analysis, Volume 169 of Graduate. Springer, New York (1997)
4. Byrd, R.H., Nocedal, J., Oztoprak, F.: An inexact successive quadratic approximation method for L-1 regularized optimization. Math. Program. Ser. B **157**(2), 375–396 (2016)
5. Combettes, P.L., Wajs, V.R.: Signal recovery by proximal forward-backward splitting. Multiscale Model. Simul. **4**(4), 1168–1200 (2005)
6. Dontchev, A.L., Rockafellar, R.T.: Implicit Functions and Solution Mappings. Springer Monographs in Mathematics. Springer, New York (2009)
7. Facchinei, F., Fischer, A., Herrich, M.: A family of Newton methods for nonsmooth constrained systems with nonisolated solutions. Math. Methods Oper. Res. **77**(3), 433–443 (2013)
8. Facchinei, F., Fischer, A., Herrich, M.: An LP-Newton method: nonsmooth equations, KKT systems, and nonisolated solutions. Math. Program. Ser. A **146**(1–2), 1–36 (2014)
9. Facchinei, F., Pang, J.-S.: Finite–Dimensional Variational Inequalities and Complementarity Problems, vol. 1. Springer, New York (2003)
10. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J.: LIBLINEAR: a library for large linear classification. J. Mach. Learn. Res. **9**(Aug), 1871–1874 (2008)
11. Fischer, A.: Local behavior of an iterative framework for generalized equations with nonisolated solutions. Math. Program. Ser. A **94**(1), 91–124 (2002)
12. Fischer, A., Herrich, M., Izmailov, A.F., Solodov, M.V.: A globally convergent LP-Newton method. SIAM J. Optim. **26**(4), 2012–2033 (2016)
13. Friedman, J., Hastie, T., Höfling, H., Tibshirani, R.: Pathwise coordinate optimization. Ann. Appl. Stat. **1**(2), 302–332 (2007)
14. Hou, K., Zhou, Z., So, A.M.-C., Luo, Z.-Q.: On the linear convergence of the proximal gradient method for trace norm regularization. In Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K. Q., (eds), Advances in Neural Information Processing Systems 26: Proceedings of the 2013 Conference, pp. 710–718 (2013)
15. Hsieh, C.-J., Dhillon, I. S., Ravikumar, P. K., Sustik, M. A.: Sparse inverse covariance matrix estimation using quadratic approximation. In: Shawe-Taylor, J., Zemel, R. S., Bartlett, P., Pereira, F.C.N., Weinberger, K.Q. (eds), Advances in Neural Information Processing Systems 24: Proceedings of the 2011 Conference, pp. 2330–2338 (2011)
16. Karimi, H., Nutini, J., Schmidt, M.: Linear convergence of gradient and proximal-gradient Methods under the Polyak-Łojasiewicz Condition. In: Frasconi, P., Landwehr, N., Manco, G., Vreeken, J. (eds) Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2016), Part I, Vol. 9851 of Lecture Notes in Artificial Intelligence, pp. 795–811. Springer International Publishing AG, Cham, Switzerland (2016)
17. Lee, J.D., Sun, Y., Saunders, M.A.: Proximal Newton-type methods for minimizing composite functions. SIAM J. Optim. **24**(3), 1420–1443 (2014)
18. Li, D.-H., Fukushima, M., Qi, L., Yamashita, N.: Regularized Newton methods for convex minimization problems with singular solutions. Comput. Optim. Appl. **28**(2), 131–147 (2004)
19. Li, G., Pong, T.K.: Calculus of the exponent of Kurdyka–Łojasiewicz inequality and its applications to linear convergence of first-order methods. Found. Comput. Math. (2017). https://doi.org/10.1007/s10208-017-9366-8
20. LIBSVM Data: Classification, Regression, and Multi-label. https://www.csie.ntu.edu.tw/cjlin/libsvmtools/datasets/
21. Liu, H., So, A. M.-C., Wu, W.: Quadratic optimization with orthogonality constraint: explicit Łojasiewicz exponent and linear convergence of retraction-based line-search and stochastic variance-reduced gradient methods. Preprint (2017)
22. Luo, Z.-Q., Tseng, P.: Error bound and convergence analysis of matrix splitting algorithms for the affine variational inequality problem. SIAM J. Optim. **2**(1), 43–54 (1992)
23. Luo, Z.-Q., Tseng, P.: On the linear convergence of descent methods for convex essentially smooth minimization. SIAM J. Control Optim. **30**(2), 408–425 (1992)

24. Luo, Z.-Q., Tseng, P.: Error bounds and convergence analysis of feasible descent methods: a general approach. Ann. Oper. Res. **46**(1), 157–178 (1993)
25. Moré, J.J.: The Levenberg–Marquardt algorithm: implementation and theory. In: Watson, G.A. (ed.) Numerical Analysis, Volume 630 of Lecture Notes in Mathematics, pp. 105–116. Springer, Berlin (1978)
26. Nesterov, Yu.: Introductory Lectures on Convex Optimization: A Basic Course. Kluwer, Boston (2004)
27. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer Series in Operations Research and Financial Engineering, second edn. Springer, New York (2006)
28. O'Donoghue, B., Candès, E.: Adaptive restart for accelerated gradient schemes. Found. Comput. Math. **15**(3), 715–732 (2015)
29. Olsen, P.A., Oztoprak, F., Nocedal, J., Rennie, S.: Newton-like methods for sparse inverse covariance estimation. In: Pereira, F.C.N., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds), Advances in Neural Information Processing Systems 25: Proceedings of the 2012 Conference, pp. 755–763 (2012)
30. Pang, J.-S.: A posteriori error bounds for the linearly-constrained variational inequality problem. Math. Oper. Res. **12**(3), 474–484 (1987)
31. Pang, J.-S.: Error bounds in mathematical programming. Math. Program. **79**(1–3), 299–332 (1997)
32. Parikh, N., Boyd, S.: Proximal algorithms. Foundations and Trends® in Optimization **1**(3), 127–239 (2014)
33. Qi, H., Sun, D.: A quadratically convergent newton method for computing the nearest correlation matrix. SIAM J. Matrix Anal. Appl. **28**(2), 360–385 (2006)
34. Sardy, S., Antoniadis, A., Tseng, P.: Automatic smoothing with wavelets for a wide class of distributions. J. Comput. Gr. Stat. **13**(2), 399–421 (2004)
35. Scheinberg, K., Tang, X.: Practical inexact proximal quasi-newton method with global complexity analysis. Math. Program. Ser. A **160**(1–2), 495–529 (2016)
36. Schmidt, M., van den Berg, E., Friedlander, M.P., Murphy, K.: Optimizing costly functions with simple constraints: a limited-memory projected quasi-Newton algorithm. In: Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS 2009), pp. 456–463 (2009)
37. Tseng, P.: Error bounds and superlinear convergence analysis of some Newton-type methods in optimization. In: Nonlinear Optimization and Related Topics, vol. 36 of Applied Optimization, pp. 445–462. Springer, Dordrecht (2000)
38. Tseng, P.: Approximation accuracy, gradient methods, and error bound for structured convex optimization. Math. Program. Ser. B **125**(2), 263–295 (2010)
39. Tseng, P., Yun, S.: A coordinate gradient descent method for nonsmooth separable minimization. Math. Program. Ser. B **117**(1–2), 387–423 (2009)
40. Wen, B., Chen, X., Pong, T.K.: Linear convergence of proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth minimization problems. SIAM J. Optim. **27**(1), 124–145 (2017)
41. Wright, S.J., Nowak, R.D., Figueiredo, M.A.T.: Sparse reconstruction by separable approximation. IEEE Trans. Signal Process. **57**(7), 2479–2493 (2009)
42. Yamashita, N., Fukushima, M.: On the rate of convergence of the Levenberg–Marquardt method. In: Alefeld, G., Chen, X. (eds.) Topics in Numerical Analysis, Volume 15 of Computing Supplement, pp. 239–249. Springer, Wien (2001)
43. Yen, I. E.-H., Hsieh, C.-J., Ravikumar, P. K., Dhillon, I. S.: Constant nullspace strong convexity and fast convergence of proximal methods under high-dimensional settings. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds), Advances in Neural Information Processing Systems 27: Proceedings of the 2014 Conference, pp. 1008–1016 (2014)
44. Yuan, G.-X., Chang, K.-W., Hsieh, C.-J., Lin, C.-J.: A comparison of optimization methods and software for large-scale L1-regularized linear classification. J. Mach. Learn. Res. **11**(Nov), 3183–3234 (2010)
45. Yuan, G.-X., Ho, C.-H., Lin, C.-J.: An improved GLMNET for L1-regularized logistic regression. J. Mach. Learn. Res. **13**(1), 1999–2030 (2012)
46. Yun, S., Toh, K.-C.: A coordinate gradient descent method for ℓ_1-regularized convex minimization. Comput. Optim. Appl. **48**(2), 273–307 (2011)
47. Zhang, H., Jiang, J., Luo, Z.-Q.: On the linear convergence of a proximal gradient method for a class of nonsmooth convex minimization problems. J. Oper. Res. Soc. China **1**(2), 163–186 (2013)
48. Zhong, K., Yen, I.E.-H., Dhillon, I.S., Ravikumar, P.: Proximal quasi–Newton for computationally intensive ℓ_1–regularized M-estimators. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds) Advances in Neural Information Processing Systems 27: Proceedings of the 2014 Conference, pp. 2375–2383 (2014)

49. Zhou, Z., So, A.M.-C.: A unified approach to error bounds for structured convex optimization problems. Math. Program. Ser. A **165**(2), 689–728 (2017)
50. Zhou, Z., Zhang, Q., So, A.M.-C.: $\ell_{1,p}$-norm regularization: error bounds and convergence rate analysis of first-order methods. In: Proceedings of the 32nd International Conference on Machine Learning (ICML 2015), pp. 1501–1510 (2015)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.