Proximal Limited-Memory Quasi-Newton Methods for Nonsmooth Nonconvex Optimization

Simeon vom Dahl* Alberto De Marchi† Christian Kanzow‡

May 11, 2026

We introduce a proximal limited–memory quasi–Newton scheme for minimizing the sum of a continuously differentiable function and a proper, lower semicontinuous and prox-bounded, possibly nonsmooth, function. Both functions might be nonconvex. The method builds upon the computation of scaled proximal operators and is globalized by adaptively updating a regularization parameter based on a criterion of sufficient decrease. We prove global convergence under mild assumptions and then establish convergence of the entire sequence (with rates) under the Kurdyka–Łojasiewicz property. To efficiently solve the subproblems, we exploit the compact representation of limited-memory quasi-Newton updates. We derive also a compact representation of the limited–memory Kleinmichel formula, a rank-one quasi-Newton scheme that preserves positive definiteness under the same condition as the BFGS update. Numerical results show a significant speed up compared to other methods.

Keywords. Nonsmooth and nonconvex optimization; proximal quasi–Newton method; global convergence; limited memory methods; compact representation; Kurdyka–Łojasiewicz inequality; Kleinmichel formula.
AMS Subject Classifications. 65K10, 90C06, 90C26, 90C53.

1 Introduction

We consider the structured optimization problem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \;\; F(x) \coloneqq f(x) + \varphi(x), \tag{P}$$

where $f\colon \mathbb{R}^n \to \mathbb{R}$ and $\varphi\colon \mathbb{R}^n \to \overline{\mathbb{R}} \coloneqq \mathbb{R} \cup \{\infty\}$ are given functions that fulfill the following blanket assumptions.

*University of Würzburg, Institute of Mathematics, Emil-Fischer-Str. 30, 97074 Würzburg, Germany.
EMAIL simeon.vomdahl@uni-wuerzburg.de
†University of the Bundeswehr Munich, Department of Aerospace Engineering, Institute of Applied Mathematics and Scientific Computing, Werner-Heisenberg-Weg 39, 85577 Neubiberg, Germany.
EMAIL alberto.demarchi@unibw.de, ORCID 0000-0002-3545-6898
‡University of Würzburg, Institute of Mathematics, Emil-Fischer-Str. 30, 97074 Würzburg, Germany.
EMAIL christian.kanzow@uni-wuerzburg.de, ORCID 0000-0003-2897-2509

Assumption 1. The conditions below hold for (P):

(a) The function f is continuously differentiable on an open set containing $\operatorname{dom}\varphi$.

(b) The function φ is proper, lower semicontinuous, and prox-bounded (Theorem 2.2).

(c) The objective function $F := f + \varphi$ is bounded from below, i.e., $\inf F > -\infty$.

Under these conditions, $F\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper and lower semicontinuous, though it may be nonsmooth and nonconvex.

Structured optimization problems of the form (P) arise across a wide range of applications in applied mathematics, particularly in high-dimensional settings where regularization is often indispensable. In this setup, f is called the *loss function* and φ the *regularizer*. Over the past few decades, researchers have proposed a wide range of regularizers and paired them with loss functions from many application areas. The most classical regularizer is standard ℓ_1-regularization. It is often applied to linear regression—yielding the classical LASSO problem [39]—or to logistic regression, both of which are ubiquitous in machine learning and statistics. At the same time, interest in nonconvex regularizers has grown significantly in recent years; prominent examples include capped ℓ_1-penalties, SCAD regularization, and MCP regularization. Imaging applications such as image denoising [34] or deblurring form another major domain of interest [6]. In addition to these practical applications, the framework (P) includes the general problem of minimizing a continuously differentiable function over a closed, possibly nonconvex set.

1.1 Related work

Classical proximal methods for (P) are usually analyzed under convexity of the regularizer φ and global Lipschitz continuity of ∇f. Many practically relevant models do not satisfy at least one of these assumptions, and a clear recent trend has therefore been to weaken them as much as possible. At the first-order level, this relaxation happened in two stages. First, convexity of the regularizer was abandoned: Li and Lin [28] proved convergence of accelerated proximal-gradient schemes for nonconvex φ under a globally Lipschitz gradient, and Themelis, Stella, & Patrinos [38] derived the first forward-backward-type method with superlinear rates in the nonconvex setting. Second, the global Lipschitz requirement on ∇f was removed in a sequence of proximal-gradient works [21, 17, 13, 12, 20], in both monotone and nonmonotone variants.

These developments strongly motivate analogous weak assumptions for second-order methods. Incorporating second-order information into proximal algorithms leads to Newton-type methods whose subproblems take the form

$$\operatorname*{minimize}_x \left\{ f(x^k) + \nabla f(x^k)^\top (x - x^k) + \frac{1}{2}(x - x^k)^\top B_k (x - x^k) + \varphi(x) \right\}, \tag{1}$$

where B_k is either the exact Hessian $\nabla^2 f(x^k)$ or a (dense or limited-memory) quasi-Newton approximation.

In the fully convex setting, Lee, Sun & Saunders [27] provided a thorough analysis of exact and inexact proximal Newton and proximal quasi-Newton algorithms. Their theory delivers global convergence together with fast local rates, showing that these methods preserve the key results familiar from smooth optimization. For large-scale applications, limited-memory quasi-Newton methods are usually the preferred choice, since storing full Hessians or dense quasi-Newton

matrices is typically infeasible. Becker & Fadili [4] proposed a limited-memory proximal quasi-Newton framework based on a zero-memory SR1 update, and Scheinberg & Tang [35] gave the first global-rate result for a practical L-BFGS-based scheme.

Growing attention has been recently devoted to substantially weaker global assumptions, similar to the developments for first-order methods. Aravkin, Baraldi & Orban [2] were the first to allow nonconvex regularizers in a proximal quasi-Newton method, proving global convergence for a trust-region scheme under mere prox-boundedness of φ. Most recently, Diouane, Habiboullah, & Orban [14] established global convergence results for a proximal modified quasi-Newton method without requiring convexity of φ, global Lipschitz continuity of ∇f, or uniform boundedness of the model Hessians. For smooth unconstrained optimization, Ueda & Yamashita [40] replaced classical line searches by a single regularization parameter updated through a success-ratio test. Kanzow & Lechner [19] transferred this idea to nonsmooth problems and showed how the compact limited-memory representation of Byrd, Nocedal, & Schnabel [11] can be combined with the fast computation of quasi-Newton proximity operators from [5].

The symmetric rank-one (SR1) update offers inexpensive Sherman–Morrison formulas for both the matrix and its inverse, but it can lose positive definiteness. Kleinmichel [23, 24] proposed a simple modification that preserves positive definiteness under the same condition as the BFGS formula while retaining rank-one updates and the cheap inverse formulas. Despite these advantages, the update has seen little use. Spellucci [36] employed it in numerical tests against his own modified rank-one update. However, to the best of our knowledge, no positive-definite rank-one update has been applied in proximal quasi-Newton methods before.

1.2 Our contributions

We propose a limited-memory regularized proximal quasi-Newton method (RPQN) for problem (P) under Assumption 1. Our method builds on the algorithm introduced in [4], and, more specifically, on [19, 26]. However, their analysis assumes that φ is convex and real-valued, and that ∇f is globally Lipschitz continuous. Our work removes both limitations by allowing extended-valued nonconvex functions φ and requiring only local Lipschitz assumptions for convergence. Thus, our assumptions are comparable to those in [14] and, under additional local Lipschitz continuity, we prove stationarity of accumulation points.

We present the first proximal-type method that systematically employs Kleinmichel's positive-definite rank-one update [23, 24]. Although it retains the cheap Sherman–Morrison inverse of SR1 while guaranteeing positive definiteness, the formula has been largely overlooked. Extending the compact limited-memory framework of [19, 26], we formulate a limited-memory Kleinmichel update, prove its compact representation, and embed it in the semismooth Newton subproblem solver of [19]. The resulting inner systems stay small, so the solver plugs into our RPQN with negligible overhead.

1.3 Organization of the paper

We begin with some preliminary concepts and results in Section 2. A detailed description of our algorithm together with a general convergence theory under minimal assumptions is given in Section 3. The convergence of the entire iteration sequence together with a rate-of-convergence analysis, in particular under the Kurdyka-Łojasiewicz inequality, is shown in Section 4. We discuss the quasi-Newton update by Kleinmichel in Section 5, which preserves positive definiteness under the same conditions as the more famous BFGS method, but has the advantage of being only a rank-one update. An efficient implementation of the overall

algorithm is based on a compact representation of the limited-memory quasi-Newton formula by Kleinmichel, as detailed in Section 6 along with other refinements. Some numerical results including a comparison of different limited-memory quasi-Newton updates are presented in Section 7, before concluding with some final remarks.

1.4 Notation

Throughout, $\mathbb{N} = \{1, 2, \dots\}$ denotes the set of positive integers and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. $\mathbb{R}$ is the real numbers and we write $\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ for the extended reals. For $a, b \in \mathbb{R}$ we use (a, b), $[a, b]$, $[a, b)$ and $(a, b]$ for the usual open, closed and semi-closed intervals.

The sets

$$\mathbb{S}^n := \{A \in \mathbb{R}^{n \times n} \mid A^\top = A\} \quad \text{and} \quad \mathbb{S}^n_{++} := \{A \in \mathbb{S}^n \mid A \succ 0\}$$

collect symmetric and symmetric positive definite matrices, respectively. The identity matrix (with dimensions clear from context) is I.

Vectors $x, y \in \mathbb{R}^n$ are equipped with the Euclidean norm $\|x\|$ and inner product $\langle x, y\rangle := x^\top y$. For a matrix $H \in \mathbb{S}^n$, we let $\|x\|_H^2 := \langle x, Hx\rangle$; which is the squared norm induced by H whenever $H \in \mathbb{S}^n_{++}$. For a closed set $C \subseteq \mathbb{R}^n$ the distance from x to C is $\operatorname{dist}(x, C) := \inf_{y \in C} \|x - y\|$, and $\mathbb{B}_\varepsilon(x) := \{y \in \mathbb{R}^n \mid \|y - x\| \leq \varepsilon\}$ denotes the closed ball of radius ε around x. Given an extended-valued function $f \colon \mathbb{R}^n \to \overline{\mathbb{R}}$, its domain is $\operatorname{dom} f := \{x \in \mathbb{R}^n \mid f(x) < \infty\}$ and f is called proper if $\operatorname{dom} f \neq \emptyset$. Sequences are written $\{x^k\}_{k \in \mathcal{I}}$ for an index set $\mathcal{I} \subseteq \mathbb{N}_0$; we abbreviate this to $\{x^k\}_{\mathcal{I}}$, and simply $\{x^k\}$ when $\mathcal{I} = \mathbb{N}_0$.

2 Preliminaries

We collect here the background material needed for the sections that follow.

2.1 Limiting subdifferential

For an extended-valued function $F \colon \mathbb{R}^n \to \overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ we denote by $\partial F(x)$ its *limiting subdifferential* at x; full definitions can be found in [33, 31]. If F is continuously differentiable at x, then $\partial F(x) = \{\nabla F(x)\}$. If F is convex, then $\partial F(x)$ coincides with the classical convex subdifferential. Otherwise, we only recall the facts that we explicitly use, all taken from [33, Chapter 8]:

- For the structured objective $F = f + \varphi$ and every $x \in \operatorname{dom} \varphi$, $\partial F(x) = \nabla f(x) + \partial \varphi(x)$.
- If $\bar{x}$ is a local minimum of F, then $0 \in \partial F(\bar{x})$.
- The mapping ∂F is *outer semicontinuous* with respect to F-attentive convergence, i.e., if $x^k \to x \in \operatorname{dom} F$, $F(x^k) \to F(x)$, and $v^k \in \partial F(x^k)$ with $v^k \to v$, then $v \in \partial F(x)$.

The third property is often also referred to as *robustness* of the limiting subdifferential. Motivated by the second property, a point x is called *stationary* for (P) if $0 \in \partial F(x)$; the set of all such points is denoted by Ω. Assumption 1(a), together with the outer semicontinuity of $\partial \varphi$, guarantees that Ω is closed, although it may be empty in the present setting.

2.2 Proximity Operator

Here we briefly recall the proximity operator and Moreau envelope, along with the notion of prox-boundedness, which is central to our analysis later. All results in this section are taken from [33, Section 1.G], where the reader is referred for a full treatment.

Definition 2.1. Let $\varphi\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper and lower semicontinuous, and let $\gamma > 0$. The *Moreau envelope* and *proximity operator* of φ with parameter γ are

$$\operatorname{env}_\varphi^\gamma(x) := \inf_{y\in\mathbb{R}^n} \left\{ \varphi(y) + \frac{1}{2\gamma}\|y-x\|^2 \right\},$$
$$\operatorname{prox}_\varphi^\gamma(x) := \operatorname*{argmin}_{y\in\mathbb{R}^n} \left\{ \varphi(y) + \frac{1}{2\gamma}\|y-x\|^2 \right\}.$$

For $\gamma = 1$, we write $\operatorname{env}_\varphi := \operatorname{env}_\varphi^1$ and $\operatorname{prox}_\varphi := \operatorname{prox}_\varphi^1$.

In general, $\operatorname{prox}_\varphi^\gamma$ is set-valued, whereas for convex φ it is single-valued. If φ is an indicator function δ_C, then $\operatorname{prox}_\varphi^\gamma$ coincides with the projection onto C. The definitions extend naturally to a positive definite matrix $H \in \mathbb{S}^n_{++}$ by replacing $\frac{1}{2\gamma}\|\cdot\|^2$ with $\frac{1}{2}\|\cdot\|_H^2$, yielding $\operatorname{env}_\varphi^H$ and $\operatorname{prox}_\varphi^H$, reducing to the scalar definition with $H = \gamma^{-1}I$. To formulate existence properties of proximal points in this general setting, we next recall prox-boundedness.

Definition 2.2. A function $\varphi\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ is called *prox-bounded* if there exists $\gamma > 0$ such that $\operatorname{env}_\varphi^\gamma(x) > -\infty$ for some $x \in \mathbb{R}^n$. The supremum of all such γ is called the *threshold of prox-boundedness* of φ and is denoted by γ_φ.

The characterizations given in [33, Exercise 1.24] show that prox-boundedness is clearly a mild condition satisfied by a broad class of functions. In particular, every proper, lower semicontinuous function that is bounded from below is also prox-bounded (with $\gamma_\varphi = +\infty$).

The following proposition (part of [33, Theorem 1.25]) summarizes key regularity properties of the Moreau envelope and proximal mapping under prox-boundedness.

Proposition 2.3. *Let $\varphi\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, lower semicontinuous, and prox-bounded. Then for every $\gamma \in (0, \gamma_\varphi)$, the set $\operatorname{prox}_\varphi^\gamma(x)$ is nonempty and compact for all $x \in \mathbb{R}^n$, whereas the value $\operatorname{env}_\varphi^\gamma(x)$ is finite and depends continuously on (γ, x).*

Note that Theorem 2.3 sharpens Theorem 2.2 in an important way: for every $\gamma \in (0, \gamma_\varphi)$, the envelope is finite *for all* $x \in \mathbb{R}^n$. Hence, the qualifier *some* in Theorem 2.2 can effectively be read as *all*.

2.3 Kurdyka–Łojasiewicz property

We now state the definition of the celebrated Kurdyka–Łojasiewicz (KL) property, which will play a central role in our subsequent convergence analysis. Since the nonsmooth version was introduced in [9], it has been invoked countless times as a key assumption in the local convergence theory of algorithms for problem (P). We refer the reader also to the recent textbook [18, Chapter 8] for a comprehensive discussion and motivation of this KL property.

Definition 2.4 (Kurdyka–Łojasiewicz)**.** Let $g\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper and lower semicontinuous. We say that g has the *KL property* at $x^* \in \{x \in \mathbb{R}^n \mid \partial g(x) \neq \emptyset\}$ if there exist a constant $\eta > 0$, a neighborhood $U \subset \mathbb{R}^n$ of x^*, and a continuous concave function $\chi\colon [0,\eta] \to [0,\infty)$ which is

continuously differentiable on $(0, \eta)$ and satisfies $\chi(0) = 0$ as well as $\chi'(t) > 0$ for all $t \in (0, \eta)$ such that the so-called *KL inequality*

$$\chi'\left(g(x) - g(x^*)\right) \operatorname{dist}\left(0, \partial g(x)\right) \geq 1$$

holds for all $x \in U \cap \{x \in \mathbb{R}^n \mid g(x^*) < g(x) < g(x^*) + \eta\}$. The function χ from above is referred to as the *desingularization function*. In case it can be chosen to be the function $t \mapsto ct^{1-\theta}$ with $\theta \in [0, 1)$ for some $c > 0$, then g is said to have the KL property of exponent θ at x^*. If g has the KL property (of exponent θ) everywhere, it is called a KL function (of exponent θ).

By [3, Lemma 2.1], every proper, lower semicontinuous function $g\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ already enjoys the KL property at all non-critical points x, i.e., those with $0 \notin \partial g(x)$. Subsequent work has identified functions that satisfy the property everywhere. In particular, Bolte et al. [9] showed that all *tame* functions—those definable on o-minimal structures—are KL. This covers real polynomials, p-norms, exponentials, logarithms, and many others. Because tame functions are closed under finite sums and compositions, this yields a large family of KL functions [3].

To simplify the KL-based convergence-rate analysis later on, we will make use of the following technical lemma from [1, Lemma 1].

Lemma 2.5. *Let $\{a_j\}_{j\in\mathbb{N}}$ be a non-negative, monotonically decreasing sequence converging to 0, and suppose that for all sufficiently large j,*

$$a_j^\alpha \leq \beta\,(a_j - a_{j+1}), \tag{2}$$

for some constants $\alpha, \beta > 0$. Then the following statements hold:

(a) *If $\alpha \in (0, 1]$, then $\{a_j\}$ converges linearly to 0 with rate $1 - \frac{1}{\beta}$.*

(b) *If $\alpha > 1$, then there exists a constant $C > 0$ such that*

$$a_j \leq C j^{-\frac{1}{\alpha - 1}} \qquad \text{for all } j \text{ sufficiently large.}$$

3 Methodology and global convergence

This section introduces our regularized proximal quasi-Newton method and presents convergence results under a hierarchy of assumptions.

3.1 Algorithm

Consider a fixed iteration $k \in \mathbb{N}_0$, and let $x^k \in \mathbb{R}^n$ denote the current iterate. Proximal Newton–type methods compute the next iterate by solving the subproblem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad q_k(x) \coloneqq f(x^k) + \nabla f(x^k)^\top (x - x^k) + \frac{1}{2}(x - x^k)^\top B_k (x - x^k) + \varphi(x),$$

where the symmetric matrix B_k captures second-order information of f. A *proximal Newton* method takes $B_k = \nabla^2 f(x^k)$, whereas a *proximal quasi-Newton* method uses an approximation $B_k \approx \nabla^2 f(x^k)$. Because, in general, the Hessian surrogate B_k may be indefinite, q_k is not necessarily convex. We apply a positive spectral shift

$$G_k \;\coloneqq\; B_k + \mu_k I, \qquad \mu_k > 0, \tag{3}$$

and instead solve the regularized subproblem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad \hat{q}_k(x) \coloneqq f(x^k) + \nabla f(x^k)^\top (x - x^k) + \frac{1}{2}(x - x^k)^\top G_k (x - x^k) + \varphi(x). \tag{4}$$

At every iteration of Algorithm 1 we search for a minimizer $\hat{x}^k$ of (4). If successful, we write

$$d^k \coloneqq \hat{x}^k - x^k. \tag{5}$$

Our globalization strategy adaptively updates the regularization parameter μ_k. This approach was introduced in [40] and further used in [22] for smooth unconstrained problems, and more recently extended to nonsmooth problems in [26, 41, 2, 14]. To evaluate the candidate iterate $\hat{x}^k$, we compare the actual decrease in the objective with the decrease predicted by the quadratic model, namely

$$\text{ared}_k \coloneqq F(x^k) - F(\hat{x}^k), \qquad \text{pred}_k \coloneqq F(x^k) - q_k(\hat{x}^k). \tag{6}$$

Algorithm 1 Regularized Proximal Quasi-Newton Method

Choose parameters $c_1 \in (0,1)$; $\sigma_2 > 1$; $0 < \mu_{\min} \leq \mu_{\max} < \infty$.
Choose $x^0 \in \operatorname{dom}\varphi$, $B_0 \in \mathbb{R}^{n\times n}$, $\mu_0 \in [\mu_{\min}, \mu_{\max}]$.
for $k = 0, 1, 2, \ldots$ **do**
    Set $G_k = B_k + \mu_k I$.
    Search for a solution $\hat{x}^k$ of the regularized subproblem (4).
    if $\hat{x}^k$ is found **then**
        Compute ared_k and pred_k as in (6).
    end if
    if $\hat{x}^k$ is found **and** $\text{ared}_k \geq c_1 \text{pred}_k$ **then**
        Set $x^{k+1} = \hat{x}^k$ and choose $B_{k+1} \in \mathbb{R}^{n\times n}$, $\mu_{k+1} \in [\mu_{\min}, \mu_{\max}]$. ▷ successful
    else
        Set $x^{k+1} = x^k$, $B_{k+1} = B_k$ and $\mu_{k+1} = \sigma_2 \mu_k$. ▷ unsuccessful
    end if
end for

At every iteration Algorithm 1 tries to obtain a solution of the regularized proximal quasi-Newton subproblem (4). If a candidate is found and passes the acceptance tests in Algorithm 1, then it is accepted as the next iterate; otherwise the current point is retained and the regularization parameter is increased.

Parameter μ_k does not require strict update rules: Algorithm 1 corresponds to picking any element from a user-defined compact (yet arbitrarily large) interval $[\mu_{\min}, \mu_{\max}]$. The upper bound $\mu_{\max}$ guarantees boundedness of all μ_k following a successful iteration. Later, in Theorem 3.12, we will see that this transfers to a crucial local boundedness property for μ_k.

Note that B_k is defined in a way that it remains constant in unsuccessful iterations. This is natural since B_k gives an approximation of the (not necessarily existing) Hessian of f at x^k, so that no change in x^k implies no change in the approximation B_k. Note that keeping B_k fixed in unsuccessful iterations k also reduces the computational costs in the subsequent iteration.

In what follows, we denote by

$$\begin{aligned}
\mathcal{K} &\coloneqq \left\{ k \in \mathbb{N}_0 \,\middle|\, \hat{x}^k \text{ is found at Algorithm 1} \right\}, \\
\mathcal{S} &\coloneqq \{ k \in \mathcal{K} \,|\, \text{iteration } k \text{ is successful} \}, \\
\mathcal{U} &\coloneqq \{ k \in \mathcal{K} \,|\, \text{iteration } k \text{ is unsuccessful} \},
\end{aligned}$$

the sets of *computed*, *successful*, and *unsuccessful* iterations of Algorithm 1.

3.2 Convergence analysis

The remainder of this section is devoted to the global convergence analysis of Algorithm 1. Throughout our convergence analysis, we make the implicit assumption that, for all iterations $k \in \mathbb{N}_0$, a solution $\hat{x}^k$ is found at Algorithm 1 of Algorithm 1 whenever subproblem (4) admits one.

We begin with a standard structural observation: every subproblem solution can be expressed as a scaled proximal step. Although this is not needed for the convergence proofs, it is the key relation underlying the subproblem solver in Section 6.2.

Lemma 3.1. *Suppose that Assumption 1 holds. Let $k \in \mathcal{K}$ and suppose $G_k \in \mathbb{S}^n$. Then the inclusion $\hat{x}^k \in \operatorname{prox}_{\varphi}^{G_k}\left(x^k - G_k^{-1}\nabla f(x^k)\right)$ holds.*

Proof. By the definition of the scaled proximal operator, we have

$$\begin{aligned}
\operatorname{prox}_{\varphi}^{G_k}\left(x^k - G_k^{-1}\nabla f(x^k)\right) &= \operatorname*{argmin}_{x\in\mathbb{R}^n}\left\{\varphi(x) + \frac{1}{2}\left\|x - (x^k - G_k^{-1}\nabla f(x^k))\right\|_{G_k}^2\right\} \\
&= \operatorname*{argmin}_{x\in\mathbb{R}^n}\left\{\varphi(x) + \frac{1}{2}\left(x - x^k + G_k^{-1}\nabla f(x^k)\right)^\top G_k\left(x - x^k + G_k^{-1}\nabla f(x^k)\right)\right\} \\
&= \operatorname*{argmin}_{x\in\mathbb{R}^n}\left\{\varphi(x) + \frac{1}{2}\left[(x-x^k)^\top G_k(x-x^k) + 2(x-x^k)^\top G_k\left(G_k^{-1}\nabla f(x^k)\right)\right]\right\} \\
&= \operatorname*{argmin}_{x\in\mathbb{R}^n}\left\{\varphi(x) + \nabla f(x^k)^\top(x-x^k) + \frac{1}{2}(x-x^k)^\top G_k(x-x^k)\right\} = \operatorname*{argmin}_{x\in\mathbb{R}^n}\hat{q}_k(x) \ni \hat{x}^k,
\end{aligned}$$

where we repeatedly used the fact that adding a constant to the objective function does not change its minimizer. This concludes the proof. □

Using the relationship between q_k and $\hat{q}_k$, we now establish an explicit lower estimate for pred_k together with two criteria that guarantee stationarity of x^k. Recall that d^k is defined in (5).

Lemma 3.2. *Suppose that Assumption 1 holds. For all $k \in \mathcal{K}$ it holds that $\operatorname{pred}_k \geq \frac{\mu_k}{2}\|d^k\|^2$. Furthermore, we have*

$$\operatorname{pred}_k = 0 \quad \Longrightarrow \quad \|d^k\| = 0 \quad \Longrightarrow \quad x^k \text{ is a stationary point of (P).}$$

Proof. By definition, it holds that $\hat{q}_k(x^k) = F(x^k)$. Hence, because $\hat{x}^k$ minimizes $\hat{q}_k$, we have $F(x^k) - \hat{q}_k(\hat{x}^k) \geq F(x^k) - \hat{q}_k(x^k) = 0$. Using the identity $q_k(x) = \hat{q}_k(x) - \frac{\mu_k}{2}\|x - x^k\|^2$, we therefore get

$$\operatorname{pred}_k = F(x^k) - q_k(\hat{x}^k) = F(x^k) - \hat{q}_k(\hat{x}^k) + \frac{\mu_k}{2}\|d^k\|^2 \geq \frac{\mu_k}{2}\|d^k\|^2. \tag{7}$$

Since $\mu_k > 0$, $\operatorname{pred}_k = 0$ implies $\|d^k\| = 0$. Then it follows that

$$0 \in \partial\hat{q}_k(\hat{x}^k) = \partial\hat{q}_k(x^k) = \nabla f(x^k) + \partial\varphi(x^k) = \partial F(x^k).$$

Hence, x^k is a stationary point of (P). □

The following bound is a direct consequence of Theorem 3.2 and will be used to handle unsuccessful iterations when the regularization parameter becomes large.

Lemma 3.3. *For all $k \in \mathcal{K}$ and any $\omega > 0$ it holds that*

$$\sqrt{\mu_k \text{pred}_k} \geq \omega \implies \text{pred}_k \geq \frac{\omega}{\sqrt{2}} \|d^k\|.$$

Proof. Theorem 3.2 together with $\sqrt{\mu_k \text{pred}_k} \geq \omega$ implies

$$\text{pred}_k \geq \sqrt{\text{pred}_k} \frac{\omega}{\sqrt{\mu_k}} \geq \sqrt{\frac{1}{2}\mu_k \|d^k\|^2} \frac{\omega}{\sqrt{\mu_k}} = \frac{\omega}{\sqrt{2}} \|d^k\|,$$

where the second inequality is due to (7). □

Together with the update mechanism in Algorithm 1, the previous lemmas imply three basic properties that will be used repeatedly in the convergence analysis.

Lemma 3.4. *Suppose that Assumption 1 holds. Then Algorithm 1 satisfies the following properties:*

(a) For all $k \in \mathbb{N}_0$ it holds that $\mu_k \geq \mu_{\min}$.

(b) For all $k \in \mathcal{K}$ it holds that $\text{pred}_k \geq \frac{\mu_{\min}}{2} \|d^k\|^2$.

(c) The sequence $\{F(x^k)\}$ is monotonically decreasing.

Proof. Statement (a) follows inductively from the possible updates for μ_k in Algorithm 1. Statement (b) then simply follows from Theorem 3.2 and (a). Let $k \in \mathbb{N}_0$. If k is an unsuccessful iteration, then it simply holds that $F(x^{k+1}) = F(x^k)$. If k is successful, then it follows from (b) that

$$F(x^k) - F(x^{k+1}) = \text{ared}_k \geq c_1 \text{pred}_k \geq 0.$$

This shows that the sequence $\{F(x^k)\}$ is monotonically decreasing. □

Our global convergence analysis is organized with a hierarchy of regularity assumptions. Starting from Assumption 1, we first establish elementary properties of the computed steps d^k and arrive at Theorem 3.7. We then incorporate boundedness of the matrices $\{B_k\}$ to obtain a true stationarity conclusion; see Theorem 3.11. In the final step, a local Lipschitz assumption on ∇f allows us to promote these results to stationarity of all accumulation points; see Theorem 3.15.

Lemma 3.5. *Suppose that Assumption 1 holds and Algorithm 1 performs infinitely many successful iterations. Then it holds that $\lim_{k \in \mathcal{S}} \|d^k\| = 0$.*

Proof. By definition, for every $k \in \mathcal{S}$,

$$F(x^k) - F(\hat{x}^k) = \text{ared}_k \geq c_1 \text{pred}_k \geq \frac{c_1 \mu_{\min}}{2} \|d^k\|^2,$$

where we used Theorem 3.4(b) in the last inequality. Since F is bounded from below by Assumption 1, summation over k yields

$$\infty > \sum_{k=0}^{\infty} \left[F(x^k) - F(x^{k+1})\right] = \sum_{k \in \mathcal{S}} \left[F(x^k) - F(\hat{x}^k)\right] \geq \frac{c_1 \mu_{\min}}{2} \sum_{k \in \mathcal{S}} \|d^k\|^2.$$

Therefore $\lim_{k \in \mathcal{S}} \|d^k\| = 0$. □

The following result concerns the case with finitely many successful iterations.

Lemma 3.6. *Suppose that Assumption 1 holds. The set $\mathcal{K}$ is infinite and for every index set $\mathcal{L} \subset \mathbb{N}_0$ with $\{x^k\}_\mathcal{L}$ and $\{B_k\}_\mathcal{L}$ bounded and $\{\mu_k\}_\mathcal{L} \to \infty$, the set $\mathcal{L}' \coloneqq \mathcal{L} \cap \mathcal{K}$ is also infinite and it holds that $\{d^k\}_{\mathcal{L}'} \to 0$.*

Proof. By Theorems 2.2 and 2.3, the Moreau envelope $\mathrm{env}_\varphi^\gamma$ is real-valued and continuous and satisfies, for all sufficiently small $\gamma > 0$, $\mathrm{env}_\varphi^\gamma(z) > -\infty$ for all $z \in \mathbb{R}^n$. Hence, $\varphi(x) \geq \mathrm{env}_\varphi^\gamma(x^k) - \frac{1}{2\gamma}\|x - x^k\|^2$ for all $x \in \mathbb{R}^n$, from which we obtain

$$\begin{aligned}
\hat{q}_k(x) &= q_k(x) + \frac{\mu_k}{2}\|x - x^k\|^2 \\
&= f(x^k) + \nabla f(x^k)^\top (x - x^k) + \frac{1}{2}(x - x^k)^\top B_k (x - x^k) + \varphi(x) + \frac{\mu_k}{2}\|x - x^k\|^2 \\
&\geq f(x^k) + \nabla f(x^k)^\top (x - x^k) + \frac{1}{2}(x - x^k)^\top B_k (x - x^k) \qquad (8) \\
&\quad + \frac{1}{2}\left(\mu_k - \frac{1}{\gamma}\right)\|x - x^k\|^2 + \mathrm{env}_\varphi^\gamma(x^k) \\
&\geq f(x^k) - \|\nabla f(x^k)\|\|x - x^k\| - \frac{1}{2}\|B_k\|\|x - x^k\|^2 + \frac{1}{2}\left(\mu_k - \frac{1}{\gamma}\right)\|x - x^k\|^2 + \mathrm{env}_\varphi^\gamma(x^k) \\
&= f(x^k) - \|\nabla f(x^k)\|\|x - x^k\| + \frac{1}{2}\left(\mu_k - \frac{1}{\gamma} - \|B_k\|\right)\|x - x^k\|^2 + \mathrm{env}_\varphi^\gamma(x^k), \qquad (9)
\end{aligned}$$

where we used prox-boundedness of φ in the first inequality, and the Cauchy-Schwarz inequality in the second. This shows that for $\mu_k > \frac{1}{\gamma} + \|B_k\|$, $\hat{q}_k$ is coercive and together with lower semi-continuity of $\hat{q}_k$ it follows by Weierstrass' theorem that $\hat{q}_k$ has at least one minimizer. By our implicit assumption, such a minimizer $\hat{x}^k$ is found by Algorithm 1. Assume, by contradiction, that $\mathcal{K}$ is only a finite set. Then all iterates are eventually unsuccessful. However, then it must hold that $\mu_k > \frac{1}{\gamma} + \|B_k\|$ for k sufficiently large, a contradiction. Hence, $\mathcal{K}$ is an infinite set. For an index set $\mathcal{L} \subset \mathbb{N}_0$ with $\{x^k\}_\mathcal{L}$ and $\{B_k\}_\mathcal{L}$ bounded and $\{\mu_k\}_\mathcal{L} \to \infty$, it follows immediately from (8) that $k \in \mathcal{L}'$ for $k \in \mathcal{L}$ sufficiently large. For all $k \in \mathcal{L}'$ it holds that $F(x^k) = \hat{q}_k(x^k) \geq \hat{q}_k(\hat{x}^k)$, which together with (8) implies $\{d^k\}_{\mathcal{L}'} \to 0$, as otherwise the aforementioned properties of $\mathcal{L}$ together with continuity of ∇f and $\mathrm{env}_\varphi^\gamma$ would yield a contradiction to $F(x^k) \leq F(x^0)$, which follows from Theorem 3.4(c). □

Corollary 3.7. *Suppose that Assumption 1 holds. Then it holds that*

$$\liminf_{k \in \mathcal{K}} \|d^k\| = 0.$$

Proof. This follows immediately from Theorem 3.5 for the case with infinitely many successful iterations, and from Theorem 3.6 otherwise. □

A small step norm $\|d^k\|$ is often used in practice as a stopping criterion, and Theorem 3.2 explains why this is reasonable in the present setting. Accordingly, the previous corollary can already be viewed as a (weak) global convergence statement under the basic assumptions alone, in particular, without assuming boundedness of the sequence $\{B_k\}$. That said, vanishing step lengths are not the same as a quantitative stationarity estimate: even though $\|d^k\| = 0$ forces x^k to be stationary, one cannot control $\mathrm{dist}(0, \partial F(x^k))$ by $\|d^k\|$ without extra structure. For this reason, we now strengthen the assumptions.

Assumption 2. The sequence $\{B_k\}$ is bounded, i.e., there exists $M_B > 0$ with $\|B_k\| \leq M_B$ for all $k \in \mathbb{N}_0$.

We now present two technical results ahead of Theorem 3.10.

Lemma 3.8. *Suppose that Assumptions 1 and 2 hold. For an infinite set $\mathcal{L} \subset \mathcal{U}$, if $\{x^k\}_\mathcal{L}$ is bounded and $\{\mu_k\}_\mathcal{L} \to \infty$, then it holds that $\lim_{k\in\mathcal{L}} \mu_k \|d^k\| = 0$.*

Proof. Let us assume, by contradiction, that there exists some $\omega > 0$ such that $\mu_k\|d^k\| \geq \omega$ on a subset $\mathcal{L}' \subset \mathcal{L}$. From Theorem 3.6 we know that $\{d^k\}_{\mathcal{L}'} \to 0$, whereas (7) gives

$$\text{pred}_k \geq \frac{\mu_k}{2}\|d^k\|^2 \geq \frac{\omega}{2}\|d^k\| \tag{10}$$

for all $k \in \mathcal{L}'$. By Taylor's theorem, there exists ξ^k on the line segment between x^k and $\hat{x}^k$ with $f(\hat{x}^k) = f(x^k) + \nabla f(\xi^k)^\top d^k$. From the boundedness of $\{x^k\}_{\mathcal{L}'}$ and $\{d^k\}_{\mathcal{L}'} \to 0$ it follows that both x^k and ξ^k belong to a sufficiently large compact set for all $k \in \mathcal{L}'$. The continuity of ∇f implies uniform continuity on this compact set and, hence, from $\{\|\xi^k - x^k\|\}_{\mathcal{L}'} \to 0$ it follows that

$$\|\nabla f(\xi^k) - \nabla f(x^k)\| \to_{\mathcal{L}'} 0. \tag{11}$$

Therefore, we get

$$\begin{aligned}
|\rho_k - 1| &= \left|\frac{\text{ared}_k - \text{pred}_k}{\text{pred}_k}\right| = \frac{\left|f(\hat{x}^k) - f(x^k) - \nabla f(x^k)^\top d^k - \frac{1}{2}\left(d^k\right)^\top B_k d^k\right|}{\text{pred}_k} \\
&\leq \frac{\left|f(\hat{x}^k) - f(x^k) - \nabla f(x^k)^\top d^k\right|}{\text{pred}_k} + \frac{\left|\left(d^k\right)^\top B_k d^k\right|}{2\text{pred}_k} \\
&= \frac{\left|(\nabla f(\xi^k) - \nabla f(x^k))^\top d^k\right|}{\text{pred}_k} + \frac{\left|\left(d^k\right)^\top B_k d^k\right|}{2\text{pred}_k} \\
&\leq \frac{\|\nabla f(\xi^k) - \nabla f(x^k)\|\|d^k\|}{\text{pred}_k} + \frac{\|B_k\|\|d^k\|^2}{2\text{pred}_k} \leq \frac{2\|\nabla f(\xi^k) - \nabla f(x^k)\|}{\omega} + \frac{\|B_k\|\|d^k\|}{\omega}
\end{aligned}$$

for all $k \in \mathcal{L}'$ sufficiently large, where we used the Cauchy-Schwarz inequality in the second, and (10) in the third inequality. By $\{d^k\}_{\mathcal{L}'} \to 0$, boundedness of $\{B_k\}$ and (11), it follows that the right-hand side converges to zero, which implies that $\{\rho_k\}_{\mathcal{L}'} \to 1$. Hence, it holds that $\text{ared}_k \geq c_1 \text{pred}_k$ for $k \in \mathcal{L}'$ sufficiently large. This yields a contradiction as such a k would be a successful iteration. □

An immediate consequence is the finite-success case: if only finitely many iterations are successful, then both $\{x^k\}$ and $\{B_k\}$ eventually stop changing, while $\mu_k \to \infty$, and therefore $\{d^k\}_\mathcal{K} \to 0$.

Lemma 3.9. *Suppose Assumption 1 holds. If $\{x^k\}$ is unbounded, then $\liminf_{k\in\mathcal{K}} \mu_k\|d^k\| = 0$.*

Proof. We prove the statement by contraposition. Suppose that there exists $\omega > 0$ with $\mu_k\|d^k\| \geq \omega$ for all $k \in \mathcal{K}$. Then it follows that

$$\begin{aligned}
\infty > F(x^0) - \inf F \geq \sum_{k=0}^{\infty} F(x^k) - F(x^{k+1}) &= \sum_{k\in\mathcal{S}} F(x^k) - F(x^{k+1}) \\
&\geq c_1 \sum_{k\in\mathcal{S}} \text{pred}_k \geq \frac{c_1\omega}{2} \sum_{k\in\mathcal{S}} \|x^{k+1} - x^k\| = \frac{c_1\omega}{2} \sum_{k=0}^{\infty} \|x^{k+1} - x^k\|,
\end{aligned}$$

where the last inequality is due to (5) and (7). Thus $\{x^k\}$ is a Cauchy sequence, and thus converges, contradicting the assumption that $\{x^k\}$ is unbounded. □

Theorem 3.10. *Suppose Assumptions 1 and 2 are satisfied. Then* $\liminf_{k\in\mathcal{K}} \mu_k \|d^k\| = 0$ *holds.*

Proof. If $\{x^k\}$ is unbounded, the result follows from Theorem 3.9. Hence, assume that $\{x^k\}$ is bounded for the remaining part of the proof. If $\{\mu_k\}$ is bounded, Algorithm 1 performs infinitely many successful iterations and the result follows from Theorem 3.5. It remains to assume that $\{\mu_k\}$ is unbounded. Then, there are infinitely many unsuccessful iterations and we can find an index set $\mathcal{L}' \subset \mathcal{K}$ such that $\{\mu_k\}_{\mathcal{L}'} \to \infty$. Eventually, for $k \in \mathcal{L}'$ large enough, it holds that $k-1 \in \mathcal{U}$, otherwise $\mu_k \leq \mu_{\max}$ by the update rule at Algorithm 1 of Algorithm 1. Without loss of generality, we can assume that $k-1 \in \mathcal{U}$ for all $k \in \mathcal{L}'$. Now we define $\mathcal{L} := \{k-1 \mid k \in \mathcal{L}'\}$, for which $\{x^k\}_{\mathcal{L}}$ remains bounded and $\{\mu_k\}_{\mathcal{L}} \to \infty$. Therefore, Theorem 3.8 yields $\lim_{k\in\mathcal{L}} \mu_k \|d^k\| = 0$, concluding the proof. □

Convergence in terms of the distance of $\partial F(\hat{x}^k)$ from zero can be established when the smooth term f has additional regularity, without assuming existence of accumulation points.

Theorem 3.11. *Suppose that Assumptions 1 and 2 hold. If* ∇f *is uniformly continuous, then it holds that*

$$\liminf_{k\in\mathcal{K}} \operatorname{dist}(0, \partial F(\hat{x}^k)) = 0.$$

Proof. By Theorem 3.6, $\mathcal{K}$ is an infinite set. For every iteration $k \in \mathcal{K}$ it holds that

$$\begin{aligned} 0 \in \partial \hat{q}_k(\hat{x}^k) = \nabla f(x^k) + G_k d^k + \partial\varphi(\hat{x}^k) \\ \iff \quad \nabla f(\hat{x}^k) - \nabla f(x^k) - G_k d^k \in \nabla f(\hat{x}^k) + \partial\varphi(\hat{x}^k) = \partial F(\hat{x}^k). \end{aligned}$$

Therefore we get

$$\operatorname{dist}(0, \partial F(\hat{x}^k)) \leq \|\nabla f(\hat{x}^k) - \nabla f(x^k) - G_k d^k\| \leq \|\nabla f(\hat{x}^k) - \nabla f(x^k)\| + (\|B_k\| + \mu_k)\|d^k\|.$$

According to Theorem 3.10 it holds that $\liminf_{k\in\mathcal{K}} \mu_k \|d^k\| = 0$. Together with the uniform continuity of ∇f and the boundedness of $\{B_k\}$, taking the $\liminf_{k\in\mathcal{K}}$ on both sides yields the result. □

3.3 Stationarity of accumulation points

Under one additional *local* regularity assumption, we can derive a stronger convergence statement. In particular, this allows us to prove local boundedness of the regularization parameter and, eventually, stationarity of every accumulation point of the generated sequence. See [21, 13, 12] for related proximal-gradient analyses.

Assumption 3. ∇f is locally Lipschitz continuous on an open set containing $\operatorname{dom}\varphi$, i.e., for all $x \in \operatorname{dom}\varphi$ there exist $L, \varepsilon > 0$ such that

$$\|\nabla f(y) - \nabla f(z)\| \leq L\|y - z\|$$

for all $y, z \in \mathbb{B}_\varepsilon(x)$.

Lemma 3.12. *Suppose that Assumptions 1 to 3 hold. Then, for every accumulation point* x^* *of* $\{x^k\}$*, there exist constants* $\varepsilon > 0$ *and* $\bar{\mu} > 0$ *with* $\mu_k \leq \bar{\mu}$ *for all* $k \in \mathbb{N}_0$ *with* $x^k \in \mathbb{B}_\varepsilon(x^*)$.

Proof. Consider an accumulation point x^* of $\{x^k\}$. Assume, by contradiction, that for all $\varepsilon > 0$ and $\bar{\mu} > 0$ there exists $k \in \mathbb{N}_0$ with $x^k \in \mathbb{B}_\varepsilon(x^*)$ and $\mu_k > \bar{\mu}$. Then, owing to Assumption 3, there exists a local Lipschitz constant L^* of ∇f in $\mathbb{B}_\varepsilon(x^*)$. Moreover, there exists a subset $\mathcal{L} \subset \mathbb{N}_0$ with $\{x^k\}_\mathcal{L} \to x^*$ and $\{\mu_k\}_\mathcal{L} \to \infty$. For all $k \in \mathcal{L}$ with $\mu_k > \mu_{\max}$, $k-1$ was unsuccessful with $\mu_{k-1} = \sigma_2^{-1}\mu_k$ and $x^{k-1} = x^k$ by Algorithm 1 of Algorithm 1. This shows that we can further assume, without loss of generality, that $\mathcal{L} \subset \mathcal{U}$. Theorem 3.6 together with Assumption 2 assures that we can assume, without loss of generality, that $\mathcal{L} \subset \mathcal{K}$. Furthermore, Theorem 3.6 also implies $\lim_{k\in\mathcal{L}} \|d^k\| = 0$. By Taylor's theorem there exists ξ^k on the line segment between x^k and $\hat{x}^k$ such that $f(\hat{x}^k) = f(x^k) + \nabla f(\xi^k)^\top d^k$. Consequently, for $k \in \mathcal{L}$ sufficiently large it holds that

$$\begin{aligned}\mathrm{pred}_k - \mathrm{ared}_k &= f(\hat{x}^k) - f(x^k) - \langle \nabla f(x^k), d^k\rangle - \frac{1}{2}\langle d^k, B_k d^k\rangle \\ &= \langle \nabla f(\xi^k) - \nabla f(x^k), d^k\rangle - \frac{1}{2}\langle d^k, B_k d^k\rangle \\ &\le \|\nabla f(\xi^k) - \nabla f(x^k)\|\|d^k\| + \frac{1}{2}\|B_k\|\|d^k\|^2 \le \left(L^* + \frac{1}{2}\|B_k\|\right)\|d^k\|^2,\end{aligned}$$

where we used the Cauchy-Schwarz inequality in the first, and the Lipschitz constant L^* in the second inequality. Hence, together with Theorem 3.2 this implies

$$\mathrm{ared}_k - c_1\mathrm{pred}_k = (1-c_1)\mathrm{pred}_k - (\mathrm{pred}_k - \mathrm{ared}_k) \ge \frac{1}{2}\big((1-c_1)\mu_k - 2L^* - \|B_k\|\big)\|d^k\|^2.$$

By Assumption 2 this shows that, for $k \in \mathcal{L}$ sufficiently large, it holds that $\rho_k \ge c_1$, which is a contradiction to $k \in \mathcal{U}$. □

Recall that φ, and therefore F, is only lower semicontinuous and need not be continuous. Hence, convergence of a subsequence $\{x^k\}$ does not automatically imply convergence of the associated function values $\{F(x^k)\}$ to $F(x^*)$. Even so, the next result establishes a stronger conclusion, made possible by the specific structure of the proximal-type iteration.

Lemma 3.13. *Suppose that Assumptions 1 to 3 hold. Let x^* be an accumulation point of $\{x^k\}$. Then the entire sequence $\{F(x^k)\}$ converges to $F(x^*)$.*

Proof. By Theorem 3.4(c), the sequence $\{F(x^k)\}$ is monotonically decreasing and, by Assumption 1, bounded from below. Hence, $\{F(x^k)\}$ is convergent to some finite number $F_* > -\infty$. There exists a subset $\mathcal{L} \subset \mathbb{N}_0$ with $\{x^k\}_\mathcal{L} \to x^*$. Without loss of generality we can assume that $\mathcal{L} \subset \mathcal{S}$. Due to the lower semicontinuity of F, we get

$$F_* = \lim_{k\to\infty} F(x^k) = \liminf_{k\in\mathcal{L}} F(x^k) \ge F(x^*).$$

Thus it suffices to show that

$$F_* \le F(x^*). \tag{12}$$

In order to verify (12), recall that, for each $k \in \mathcal{L}$, $x^{k+1} = \hat{x}^k$ solves the subproblem (4). In particular, we therefore obtain

$$\begin{aligned}\hat{q}_k(x^{k+1}) &= f(x^k) + \nabla f(x^k)^\top(x^{k+1} - x^k) + \frac{1}{2}(x^{k+1} - x^k)^\top G_k(x^{k+1} - x^k) + \varphi(x^{k+1}) \\ &\le f(x^k) + \nabla f(x^k)^\top(x^* - x^k) + \frac{1}{2}(x^* - x^k)^\top G_k(x^* - x^k) + \varphi(x^*).\end{aligned} \tag{13}$$

From Theorem 3.5 we know that $\{d^k\}_{\mathcal{L}} \to 0$. Taking into account that $d^k = x^{k+1} - x^k$ for $k \in \mathcal{S}$, this implies $\{x^{k+1}\}_{\mathcal{L}} \to x^*$. Furthermore, Theorem 3.12 together with the boundedness of $\{B_k\}$ then yields the boundedness of the subsequence $\{G_k\}_{k\in\mathcal{L}}$. Now, taking the limit in (13), the previous considerations yield

$$\begin{aligned}
F_* &= \lim_{k\to\infty} F(x^k) = \lim_{k\in\mathcal{L}} F(x^{k+1}) = \lim_{k\in\mathcal{L}} \left[f(x^{k+1}) + \varphi(x^{k+1})\right] = \lim_{k\in\mathcal{L}} \left[f(x^k) + \varphi(x^{k+1})\right] \\
&= \lim_{k\in\mathcal{L}} \left[f(x^k) + \nabla f(x^k)^\top (x^{k+1} - x^k) + \frac{1}{2}(x^{k+1} - x^k)^\top G_k (x^{k+1} - x^k) + \varphi(x^{k+1})\right] \\
&\leq \lim_{k\in\mathcal{L}} \left[f(x^k) + \nabla f(x^k)^\top (x^* - x^k) + \frac{1}{2}(x^* - x^k)^\top G_k (x^* - x^k) + \varphi(x^*)\right] \\
&= f(x^*) + \varphi(x^*) = F(x^*),
\end{aligned}$$

where the inequality is due to (13). Altogether, this verifies (12) and completes the proof. □

We have another technical lemma before turning to Theorem 3.15.

Lemma 3.14. *Suppose that Assumptions 1 and 2 hold. Then for every accumulation point x^* of $\{x^k\}$ there exists a subset $\mathcal{L} \subset \mathcal{K}$ with $\{x^k\}_{\mathcal{L}} \to x^*$ and $\lim_{k\in\mathcal{L}} \mu_k \|d^k\| = 0$.*

Proof. If Algorithm 1 performs only finitely many successful iterations, then the result follows immediately from Theorems 3.6 and 3.8. Now suppose that $\mathcal{S}$ is an infinite set. Since x^* is an accumulation point of $\{x^k\}$, there exists $\mathcal{L} \subset \mathbb{N}_0$ with $\{x^k\}_{\mathcal{L}} \to x^*$. We construct the set $\mathcal{L}' \subset \mathcal{S}$ by replacing unsuccessful iterations $k \in \mathcal{L}$ with the last successful iteration before k. As $\mathcal{S}$ is an infinite set, the same holds for $\mathcal{L}'$. From Theorem 3.5 it follows that $\{x^k\}_{\mathcal{L}'} \to x^*$. If $\{\mu_k\}_{\mathcal{L}'}$ is bounded, the result follows immediately with Theorem 3.5. If $\{\mu_k\}_{\mathcal{L}'}$ is unbounded, passing to a subsequence if necessary, we can assume that $k - 1 \in \mathcal{U}$ for all $k \in \mathcal{L}'$, as for $k - 1 \in \mathcal{S}$ it would be $\mu_k \leq \mu_{\max}$ by Algorithm 1. Now consider the set $\mathcal{U}' := \{k - 1 \,|\, k \in \mathcal{L}'\}$. Clearly, for $\mathcal{U}'$ it also holds that $\{x^k\}_{\mathcal{U}'} \to x^*$ and $\{\mu_k\}_{\mathcal{U}'} \to \infty$. By Theorem 3.6, again passing to a subsequence if necessary, we have $\mathcal{U}' \subset \mathcal{U}$. Then the result follows from Theorem 3.8. □

Theorem 3.15. *Suppose that Assumptions 1 to 3 hold. Then all accumulation points of $\{x^k\}$ are stationary points of (P).*

Proof. From Theorem 3.14 we know that there exists a subset $\mathcal{L} \subset \mathcal{K}$ with $\{x^k\}_{\mathcal{L}} \to x^*$ and $\lim_{k\in\mathcal{L}} \mu_k \|d^k\| = \lim_{k\in\mathcal{L}} \|d^k\| = 0$. From Theorem 3.13 we know that $\{F(x^k)\} \to F(x^*)$ and we get

$$\lim_{k\to\infty} \varphi(x^k) = \lim_{k\to\infty} F(x^k) - f(x^k) = F(x^*) - f(x^*) = \varphi(x^*).$$

Optimality of $\hat{x}^k$ with respect to (4) combined with the Cauchy-Schwarz inequality gives

$$\begin{aligned}
F(x^k) = \hat{q}_k(x^k) \geq \hat{q}_k(\hat{x}^k) &= f(x^k) + \nabla f(x^k)^\top d^k + \frac{1}{2}(d^k)^\top G_k d^k + \varphi(\hat{x}^k) \\
&\geq f(x^k) - \|\nabla f(x^k)\| \|d^k\| + \frac{\mu_k - \|B_k\|}{2} \|d^k\|^2 + \varphi(\hat{x}^k),
\end{aligned}$$

hence

$$\varphi(x^k) \geq \varphi(\hat{x}^k) - \|\nabla f(x^k)\| \|d^k\| + \frac{\mu_k - \|B_k\|}{2} \|d^k\|^2$$

for all $k \in \mathcal{L}$. Taking the lim sup along $\mathcal{L}$ gives

$$\liminf_{k\in\mathcal{L}} \varphi(\hat{x}^k) \geq \varphi(x^*) \geq \limsup_{k\in\mathcal{L}} \varphi(\hat{x}^k),$$

where the first inequality is due to lower semicontinuity of φ and $\{\hat{x}^k\}_{\mathcal{L}} \to x^*$. Therefore, $\lim_{k\in\mathcal{L}} \varphi(\hat{x}^k) = \varphi(x^*)$. Finally, since $\hat{x}^k \in \operatorname{argmin} \hat{q}_k$, the optimality condition results in

$$0 \in \partial \hat{q}_k(\hat{x}^k) = \nabla f(x^k) + (B_k + \mu_k I)d^k + \partial\varphi(\hat{x}^k)$$

for all $k \in \mathcal{L}$. Using the aforementioned limits along with the outer semicontinuity (with respect to φ-attentive convergence) of the limiting subdifferential, taking the limit along $\mathcal{L}$ gives

$$0 \in \nabla f(x^*) + \partial\varphi(x^*) = \partial F(x^*),$$

concluding the proof. □

4 Convergence under the KL property

The first aim of this section is to prove convergence of the entire sequence $\{x^k\}$ generated by Algorithm 1 under the following (fairly mild) assumptions (in addition to Assumptions 1 to 3).

Assumption 4. Consider problem (P) and Algorithm 1.

(a) The set Ω of stationary points of (P) is nonempty and there exists an accumulation point $x^* \in \Omega$ of $\{x^k\}_{\mathcal{S}}$.

(b) F has the KL property at x^* with desingularization function χ, constant $\eta > 0$ and neighborhood U of x^*, see Definition 2.4.

The KL property is next to the error bound condition nowadays the most famous property to show convergence and rate-of-convergence results for proximal methods. In recent years, a lot of research has been done on the relation of those two concepts. A major milestone has been [10], where for convex lsc functions the equivalence between the KL property of exponent 1/2 and an error bound is established. In recent years, the KL property has become the default assumption in modern convergence analysis. One of the reasons for that is that many of the objective functions in practical applications are KL with known exponents.

In the following, we let $L^* > 0$ be the local Lipschitz constant of ∇f in $\mathbb{B}_{\varepsilon_0}(x^*)$ for some radius $\varepsilon_0 > 0$. Then, by Theorem 3.12, there exists $\bar{\mu} > 0$ and $0 < \varepsilon_1 < \varepsilon_0$ such that $\mu_k \leq \bar{\mu}$ for all $k \in \mathbb{N}_0$ with $x^k \in \mathbb{B}_{\varepsilon_1}(x^*)$.

Lemma 4.1. *Suppose that Assumptions 1 to 4 hold. Define* $\beta := \frac{c_1\mu_{\min}}{2(M_B+\bar{\mu}+L^*)}$ *and*

$$\alpha_k := \|x^k - x^*\| + \sqrt{\frac{8}{c_1\mu_{\min}}\left(F(x^k) - F(x^*)\right)} + \frac{1}{\beta}\chi\left(F(x^k) - F(x^*)\right).$$

It holds that $\liminf_{k\in\mathcal{S}} \alpha_k = 0$.

Proof. According to Assumption 4(a) there exists a subsequence of $\{x^k\}_{\mathcal{S}}$ converging to x^*. Along such a subsequence, the first summand tends to 0 by convergence of x^k to x^*; the second tends to 0 by Theorem 3.13; and the third tends to 0 because the desingularization function χ is continuous at the origin. Hence $\alpha_k \to 0$ along that subsequence. □

Lemma 4.2. *Suppose that Assumptions 1 to 4 hold. For every iteration $k \in \mathcal{S}$ with $x^k \in \mathbb{B}_{\varepsilon_1}(x^*)$ and $\|d^k\| \leq \varepsilon_0 - \varepsilon_1$ it holds that*

$$\operatorname{dist}\left(0, \partial F(x^{k+1})\right) \leq (M_B + \bar{\mu} + L^*)\,\|x^{k+1} - x^k\|.$$

Proof. Since $k \in \mathcal{S}$, $x^{k+1} = \hat{x}^k$ solves (4), whose stationarity condition gives

$$0 \in \nabla f(x^k) + G_k(x^{k+1} - x^k) + \partial\varphi(x^{k+1}).$$

Hence

$$-G_k(x^{k+1} - x^k) + \nabla f(x^{k+1}) - \nabla f(x^k) \in \nabla f(x^{k+1}) + \partial\varphi(x^{k+1}) = \partial F(x^{k+1}), \tag{14}$$

where we used the sum rule for the limiting subdifferential. Then, by $\|d^k\| \leq \varepsilon_0 - \varepsilon_1$, it holds that $x^{k+1} \in \mathbb{B}_{\varepsilon_0}(x^*)$. Assumption 2 and Theorem 3.12 guarantee that $\|G_k\| \leq M_B + \bar{\mu}$. Then, using (14) together with Assumption 3, we obtain

$$\begin{aligned}\operatorname{dist}\left(0, \partial F(x^{k+1})\right) &\leq \| - G_k(x^{k+1} - x^k) + \nabla f(x^{k+1}) - \nabla f(x^k)\| \\ &\leq (M_B + \bar{\mu})\|x^{k+1} - x^k\| + L^*\|x^{k+1} - x^k\| = (M_B + \bar{\mu} + L^*)\|x^{k+1} - x^k\|,\end{aligned}$$

which completes the proof. □

The following result guarantees convergence of the entire sequence $\{x^k\}$ under Assumptions 1 to 4. Note that, in general, these assumptions do not necessarily imply the local uniqueness of a solution. A key departure from the proof of [17, Thm 4.5] is that our induction runs only over iterations in $\mathcal{S}$ from k_0 (as defined in the proof) onward, rather than over all $k \geq k_0$.

Theorem 4.3. *Suppose that Assumptions 1 to 4 hold. Then the entire sequence $\{x^k\}$ converges to x^*.*

Proof. The sequence $\{F(x^k)\}$ is monotonically decreasing (Theorem 3.4(c)) and converges to $F(x^*)$ by Theorem 3.13; in particular, $F(x^k) \geq F(x^*)$ for all $k \in \mathbb{N}_0$. Now suppose $F(x^k) = F(x^*)$ for some $k \in \mathbb{N}_0$. By monotonicity, also $F(x^{k+1}) = F(x^*)$. First assume that iteration k is successful. Then

$$0 = F(x^k) - F(x^{k+1}) = \operatorname{ared}_k \geq c_1 \operatorname{pred}_k \geq \frac{c_1 \mu_{\min}}{2}\|x^{k+1} - x^k\|^2 \geq 0, \tag{15}$$

where the second inequality uses Theorem 3.4(b). Hence $x^{k+1} = x^k$, and this holds trivially if iteration k is unsuccessful, i.e., the sequence $\{x^k\}$ is eventually constant. Since, by Assumption 4(a), a subsequence of $\{x^k\}_{\mathcal{S}}$ converges to x^*, it follows that $x^k = x^*$ for all sufficiently large $k \in \mathbb{N}_0$.

For the remainder of the proof, assume $F(x^k) > F(x^*)$ for all $k \in \mathbb{N}_0$. According to Theorem 4.1, there exists $\mathcal{S}' \subset \mathcal{S}$ with $\{\alpha_k\}_{\mathcal{S}'} \to 0$. Together with Theorem 3.5 and Theorem 3.13, we can choose $k_0 \in \mathcal{S}'$ sufficiently large such that

$$\alpha_{k_0} \in (0, \varepsilon_1), \quad \mathbb{B}_{\alpha_{k_0}}(x^*) \subset U, \quad F(x^{k_0}) < F(x^*) + \eta, \tag{16}$$

and $\|d^k\| \leq \varepsilon_0 - \varepsilon_1$ for all successful iterations $k \geq k_0$, where η denotes the constant from the desingularization function $\chi\colon [0, \eta] \to [0, \infty)$ associated with the KL property at x^*. Since $\chi(0) = 0$ and $\chi'(t) > 0$ for all $t \in (0, \eta)$, $\chi\left(F(x^k) - F(x^*)\right) \geq 0$ for all $k \geq k_0$. Denote by $s_0, s_1, s_2, \ldots$ the successful iterations, starting with $s_0 := k_0$. As iterates are not updated during unsuccessful iterations,

$$x^{s_{j+1}} = x^{s_j + 1} \quad \text{for all } j \in \mathbb{N}_0. \tag{17}$$

We claim that for all $j \in \mathbb{N}_0$:

(a) $x^{s_j} \in \mathbb{B}_{\alpha_{k_0}}(x^*)$,

(b) $\|x^{k_0} - x^*\| + \sum_{i=0}^{j} \|x^{s_{i+1}} - x^{s_i}\| \le \alpha_{k_0}$, which is equivalent to

$$\sum_{i=0}^{j} \|x^{s_{i+1}} - x^{s_i}\| \le \sqrt{\frac{8}{c_1 \mu_{\min}} \big(F(x^{k_0}) - F(x^*)\big)} + \frac{1}{\beta} \chi\Big(F(x^{k_0}) - F(x^*)\Big). \tag{18}$$

We prove (a)–(b) by induction. For $j = 0$, (a) holds by the definition of α_{k_0}. Moreover, (15) together with the monotonicity of $\{F(x^k)\}$ and (17) yields

$$\|x^{s_1} - x^{s_0}\| = \|x^{k_0+1} - x^{k_0}\| \le \sqrt{\frac{2}{c_1 \mu_{\min}} \big(F(x^{k_0}) - F(x^{k_0+1})\big)} \le \sqrt{\frac{2}{c_1 \mu_{\min}} \big(F(x^{k_0}) - F(x^*)\big)}. \tag{19}$$

Hence (b) also holds for $j = 0$. Assume (a)–(b) hold for $i = 0, \dots, j$ with some $j \ge 0$. By the triangle inequality and the induction hypothesis,

$$\|x^{s_{j+1}} - x^*\| \le \|x^{k_0} - x^*\| + \sum_{i=0}^{j} \|x^{s_{i+1}} - x^{s_i}\| \le \alpha_{k_0}, \tag{20}$$

so (a) holds with $j + 1$ in place of j. For (b), first note that (16) implies $F(x^*) < F(x^k) < F(x^*) + \eta$ for all $k \ge k_0$. Since F has the KL property at x^* and $x^{s_i} \in \mathbb{B}_{\alpha_{k_0}}(x^*) \subset U$ for all $i \in \{0, \dots, j+1\}$ by the induction hypothesis and (20), we have

$$\chi'\big(F(x^{s_i}) - F(x^*)\big) \operatorname{dist}\big(0, \partial F(x^{s_i})\big) \ge 1 \tag{21}$$

for all $i = 0, \dots, j+1$. Theorem 4.2 and (17) give

$$\begin{aligned} \operatorname{dist}\big(0, \partial F(x^{s_{i+1}})\big) = \operatorname{dist}\big(0, \partial F(x^{s_i+1})\big) &\le (M_B + \bar{\mu} + L^*)\|x^{s_i+1} - x^{s_i}\| \\ &= (M_B + \bar{\mu} + L^*)\|x^{s_{i+1}} - x^{s_i}\| \end{aligned}$$

for all $i = 0, \dots, j+1$. In view of (21), this yields

$$\chi'\big(F(x^{s_i}) - F(x^*)\big) \ge \frac{1}{\operatorname{dist}\big(0, \partial F(x^{s_i})\big)} \ge \frac{1}{(M_B + \bar{\mu} + L^*)\|x^{s_i} - x^{s_{i-1}}\|} \tag{22}$$

for all $i = 1, \dots, j+2$. For convenience, set

$$\Delta_{i,l} := \chi\big(F(x^{s_i}) - F(x^*)\big) - \chi\big(F(x^{s_l}) - F(x^*)\big)$$

for $i, l = 0, \dots, j+2$. Then, by concavity of χ, we obtain

$$\begin{aligned} \Delta_{i,i+1} \ge \chi'\big(F(x^{s_i}) - F(x^*)\big)\big(F(x^{s_i}) - F(x^{s_i+1})\big) &\overset{(22)}{\ge} \frac{F(x^{s_i}) - F(x^{s_i+1})}{(M_B + \bar{\mu} + L^*)\|x^{s_i} - x^{s_{i-1}}\|} \\ &\overset{(15)}{\ge} \frac{c_1 \mu_{\min} \|x^{s_i+1} - x^{s_i}\|^2}{2(M_B + \bar{\mu} + L^*)\|x^{s_i} - x^{s_{i-1}}\|} \overset{(17)}{=} \beta \frac{\|x^{s_{i+1}} - x^{s_i}\|^2}{\|x^{s_i} - x^{s_{i-1}}\|} \end{aligned}$$

for all $i \in \{1, \dots, j+1\}$, where β is the constant from Theorem 4.1. Since $a + b \ge 2\sqrt{ab}$ for all $a, b \ge 0$, it follows that

$$\frac{1}{\beta}\Delta_{i,i+1} + \|x^{s_i} - x^{s_{i-1}}\| \ge 2\sqrt{\frac{1}{\beta}\Delta_{i,i+1}\|x^{s_i} - x^{s_{i-1}}\|} \ge 2\|x^{s_{i+1}} - x^{s_i}\|$$

for all $i \in \{1, \dots, j+1\}$. Summation gives

$$2\sum_{i=1}^{j+1} \|x^{s_{i+1}} - x^{s_i}\| \le \sum_{i=1}^{j+1} \|x^{s_i} - x^{s_{i-1}}\| + \frac{1}{\beta}\sum_{i=1}^{j+1} \Delta_{i,i+1} = \sum_{i=0}^{j} \|x^{s_{i+1}} - x^{s_i}\| + \frac{1}{\beta}\Delta_{1,j+2}$$
$$\le \sum_{i=1}^{j+1} \|x^{s_{i+1}} - x^{s_i}\| + \|x^{s_1} - x^{s_0}\| + \frac{1}{\beta}\Delta_{1,j+2}.$$

Subtracting the first term on the right, using (19), and the nonnegativity and monotonicity of χ, we obtain

$$\sum_{i=1}^{j+1} \|x^{s_{i+1}} - x^{s_i}\| \le \sqrt{\frac{2}{c_1\mu_{\min}}\big(F(x^{k_0}) - F(x^*)\big)} + \frac{1}{\beta}\chi\big(F(x^{k_0}) - F(x^*)\big).$$

Adding $\|x^{s_1} - x^{s_0}\|$ to both sides and using (19) again yields

$$\sum_{i=0}^{j+1} \|x^{s_{i+1}} - x^{s_i}\| \le \sqrt{\frac{8}{c_1\mu_{\min}}\big(F(x^{k_0}) - F(x^*)\big)} + \frac{1}{\beta}\chi\big(F(x^{k_0}) - F(x^*)\big).$$

Thus (b) holds with $j+1$ in place of j, completing the induction. In particular, (a) implies $x^k \in \mathbb{B}_{\alpha_{k_0}}(x^*)$ for all successful $k \ge k_0$. Since iterates are not updated during unsuccessful iterations, this in fact holds for all $k \ge k_0$. Letting $\{k_0\}_{\mathcal{S}'} \to \infty$, we get convergence of the entire sequence $\{x^k\}$ to x^*. □

We now turn from convergence itself to quantitative rates, assuming a KL desingularization of the form $\chi(t) = c\,t^{1-\theta}$. The analysis naturally applies to the successful iterations, because those are the indices where objective decrease is explicitly controlled (and because the iterates remain constant at unsuccessful iterations). A direct extension to the full sequence is generally unavailable: unsuccessful iterations may still occur at arbitrarily large indices and interrupt any global Q-type rate statement for all indices. The first statement of the next theorem resolves this mechanism by showing that such interruptions can only happen in uniformly bounded blocks.

Theorem 4.4. *Suppose that Assumptions 1 to 4 hold and that F has the KL property at x^* with desingularization function $\chi(t) = c\,t^{1-\theta}$ for some $c > 0$ and $\theta \in (0,1)$. Let $s_0, s_1, s_2, \dots$ denote the successful iterations, with $s_0 := k_0$. Then the sequence $\{s_{j+1} - s_j\}$ is bounded and the following statements hold:*

(a) If $\theta \in (0, \frac{1}{2})$, then $\{F(x^{s_k})\}$ converges Q-superlinearly to $F(x^*)$ with rate $\frac{1}{2\theta}$, and $\{x^{s_k}\}$ converges R-superlinearly to x^* with the same rate.

(b) If $\theta = \frac{1}{2}$, then $\{F(x^{s_k})\}$ converges Q-linearly to $F(x^*)$, and $\{x^{s_k}\}$ converges R-linearly to x^*.

(c) If $\theta \in (\frac{1}{2}, 1)$, then there exist constants $C_1, C_2 > 0$ such that for all sufficiently large k,

$$F(x^{s_k}) - F(x^*) \le C_1\, k^{-\frac{1}{2\theta-1}}, \qquad \|x^{s_k} - x^*\| \le C_2\, k^{-\frac{1-\theta}{2\theta-1}}.$$

Proof. Suppose there exists an index set $\mathcal{J} \subset \mathbb{N}_0$ with $\{s_{j+1} - s_j\}_{\mathcal{J}} \to +\infty$. Then $\mu_{s_{j+1}} = \sigma_2^{s_{j+1}-s_j}\mu_{s_j} \ge \sigma_2^{s_{j+1}-s_j}\mu_{\min} \to_{\mathcal{J}} +\infty$, a contradiction to Theorem 4.3 and Theorem 3.12. Hence,

there exists $\bar{s} \in \mathbb{N}$ with $s_{j+1} - s_j \leq \bar{s}$ for all $j \in \mathbb{N}_0$. This means that the number of consecutive unsuccessful iterations after k_0 is bounded.

We now show convergence rates for the subsequence of successful iterations. Together with Theorems 3.5 and 3.13, Theorem 4.3 guarantees that there exists $j_0 \in \mathbb{N}_0$ such that all successful iterates $k \geq s_{j_0}$ satisfy

$$x^k \in \mathbb{B}_{\varepsilon_1}(x^*) \cap U \cap \{x \in \mathbb{R}^n \mid F(x^*) < F(x) < F(x^*) + \eta\} \quad \text{and} \quad \|d^k\| \leq \varepsilon_0 - \varepsilon_1.$$

Throughout, we use relation (17), which holds since unsuccessful iterations do not change the iterates. For every $j \in \mathbb{N}_0$, since s_j is successful, Theorem 3.4(b) yields

$$\begin{aligned} F(x^{s_j}) - F(x^{s_j+1}) = F(x^{s_j}) - F(x^{s_j+1}) = \text{ared}_{s_j} \geq c_1 \text{pred}_{s_j} \geq \frac{c_1 \mu_{\min}}{2} \|d^{s_j}\|^2 \\ = \frac{c_1 \mu_{\min}}{2} \|x^{s_j+1} - x^{s_j}\|^2 = \frac{c_1 \mu_{\min}}{2} \|x^{s_{j+1}} - x^{s_j}\|^2. \quad (23) \end{aligned}$$

Theorem 4.2 gives, for all $j \geq j_0$,

$$\begin{aligned} \operatorname{dist}\big(0, \partial F(x^{s_j+1})\big) = \operatorname{dist}\big(0, \partial F(x^{s_j+1})\big) \leq (M_B + \bar{\mu} + L^*) \|x^{s_j+1} - x^{s_j}\| \\ = (M_B + \bar{\mu} + L^*) \|x^{s_{j+1}} - x^{s_j}\|. \quad (24) \end{aligned}$$

Since F has the KL property at x^* with $\chi(t) = c\, t^{1-\theta}$, we have for all $j \geq j_0$,

$$\begin{aligned} 1 \leq \chi'\left(F(x^{s_{j+1}}) - F(x^*)\right) \operatorname{dist}\left(0, \partial F(x^{s_{j+1}})\right) \\ = c(1-\theta)\big(F(x^{s_{j+1}}) - F(x^*)\big)^{-\theta} \operatorname{dist}\big(0, \partial F(x^{s_{j+1}})\big). \quad (25) \end{aligned}$$

Equivalently,

$$\operatorname{dist}(0, \partial F(x^{s_{j+1}})) \;\geq\; \frac{1}{c(1-\theta)}\big(F(x^{s_{j+1}}) - F(x^*)\big)^{\theta}.$$

Combining this with (24) yields

$$\|x^{s_{j+1}} - x^{s_j}\| \geq \frac{(F(x^{s_{j+1}}) - F(x^*))^{\theta}}{c(1-\theta)(M_B + \bar{\mu} + L^*)}. \quad (26)$$

By defining

$$\tau := \left(\frac{2c^2(1-\theta)^2(M_B + \bar{\mu} + L^*)^2}{c_1 \mu_{\min}}\right)^{\frac{1}{2\theta}}$$

and using (26) and (23), we obtain

$$\begin{aligned} F(x^{s_{j+1}}) - F(x^*) &\leq \big(c(1-\theta)(M_B + \bar{\mu} + L^*)\big)^{\frac{1}{\theta}} \|x^{s_{j+1}} - x^{s_j}\|^{\frac{1}{\theta}} \\ &\leq \big(c(1-\theta)(M_B + \bar{\mu} + L^*)\big)^{\frac{1}{\theta}} \left(\frac{2}{c_1 \mu_{\min}}\right)^{\frac{1}{2\theta}} \big(F(x^{s_j}) - F(x^{s_{j+1}})\big)^{\frac{1}{2\theta}} \\ &= \tau\big(F(x^{s_j}) - F(x^{s_{j+1}})\big)^{\frac{1}{2\theta}} \leq \tau\big(F(x^{s_j}) - F(x^*)\big)^{\frac{1}{2\theta}}. \quad (27) \end{aligned}$$

This implies Q-superlinear convergence of $\{F(x^{s_j})\}$ when $\theta \in (0, \frac{1}{2})$. Then, by denoting $a_j := F(x^{s_j}) - F(x^*)$, (27) gives, for $j \geq j_0$,

$$a_{j+1}^{2\theta} \leq \tau^{2\theta}(a_j - a_{j+1}),$$

which is precisely the setting of Theorem 2.5. Hence, for $\theta = \frac{1}{2}$, $\{F(x^{s_j})\}$ converges Q-linearly to $\{F(x^*)\}$ and for $\theta \in \left(\frac{1}{2}, 1\right)$ we get the inequality

$$F(x^{s_j}) - F(x^*) \leq C_1 j^{-\frac{1}{2\theta-1}}$$

for some $C_1 > 0$ and j sufficiently large.

We now verify the statements for the sequence $\{x^{s_j}\}$. Because $\{x^k\}$ converges to x^*, the statements from the induction in the proof of Theorem 4.3 remain valid when k_0 is replaced by any sufficiently large successful iteration s_k. In particular, (18) implies

$$\sum_{j=k}^{l} \|x^{s_{j+1}} - x^{s_j}\| \leq \sqrt{\frac{8}{c_1 \mu_{\min}} \big(F(x^{s_k}) - F(x^*)\big)} + \frac{1}{\beta} \chi\big(F(x^{s_k}) - F(x^*)\big)$$

for all $l \geq k$. Hence,

$$\begin{aligned}
\|x^{s_l} - x^{s_k}\| \leq \sum_{j=k}^{l-1} \|x^{s_{j+1}} - x^{s_j}\| &\leq \sqrt{\frac{8}{c_1 \mu_{\min}} \big(F(x^{s_k}) - F(x^*)\big)} + \frac{1}{\beta} \chi\big(F(x^{s_k}) - F(x^*)\big) \\
&= \sqrt{\frac{8}{c_1 \mu_{\min}} \big(F(x^{s_k}) - F(x^*)\big)} + \frac{c}{\beta} \big(F(x^{s_k}) - F(x^*)\big)^{1-\theta} \\
&\leq \left(\sqrt{\frac{8}{c_1 \mu_{\min}}} + \frac{c}{\beta}\right) \big(F(x^{s_k}) - F(x^*)\big)^{\min\{1/2, 1-\theta\}}.
\end{aligned}$$

Letting $l \to \infty$, the convergence rates for $\{x^{s_k}\}$ follow by substituting the corresponding rates already established for $\{F(x^{s_k})\}$ in the cases $\theta \in (0, \frac{1}{2})$, $\theta = \frac{1}{2}$, and $\theta \in (\frac{1}{2}, 1)$. □

5 Limited-memory Kleinmichel matrices

This section is devoted to a lesser-known quasi-Newton method, the *Kleinmichel formula*. This method is a rank-one method that, unlike the famous SR1 formula, guarantees positive definiteness under the same circumstances as the BFGS formula. The Kleinmichel formula is defined by

$$H_{k+1} = \gamma_k \left[H_k + \frac{\big(y^k - \gamma_k H_k d^k\big)\big(y^k - \gamma_k H_k d^k\big)^\top}{\gamma_k \langle y^k - \gamma_k H_k d^k, d^k \rangle} \right], \tag{28}$$

where $\gamma_k > 0$ is a free parameter and $y^k := \nabla f(x^{k+1}) - \nabla f(x^k)$. Note that $\gamma_k = 1$ yields exactly the SR1-formula. For the sake of completeness, we recount this important property, with a simple proof, from the original [23, Satz 1] (in German).

Proposition 5.1. *Suppose that H_k is positive definite, $\langle y^k, d^k \rangle > 0$ and $\gamma_k \in \left(0, \frac{\langle y^k, d^k \rangle}{\langle d^k, H_k d^k \rangle}\right)$. Then H_{k+1} as defined by the Kleinmichel formula (28) is positive definite.*

Proof. For $\gamma_k > 0$ and positive definite H_k, it is clear that H_{k+1} is positive definite if $\langle y^k - \gamma_k H_k d^k, d^k \rangle > 0$ holds. However, from $\gamma_k < \frac{\langle y^k, d^k \rangle}{\langle d^k, H_k d^k \rangle}$, it immediately follows that

$$\langle y^k - \gamma_k H_k d^k, d^k \rangle = \langle y^k, d^k \rangle - \gamma_k \langle d^k, H_k d^k \rangle > 0,$$

which completes the proof. □

So, even though the Kleinmichel update is a rank-one update, it preserves positive definiteness as long as $\langle y^k, d^k\rangle > 0$ holds, just like the BFGS update. The initial matrix H_0 is usually chosen to have a simple structure (a positive multiple of the identity matrix, for instance).

In large-scale applications it is not practical to store and update a dense matrix H_k explicitly. Instead, one usually employs a limited-memory strategy: at iteration k the matrix H_{k+1} is reconstructed from a simple initialization $H_{k,0}$ and only the m most recent quasi-Newton pairs (d^j, y^j), $j = k-m, \ldots, k-1$, for some fixed memory size $m \ll k$. This idea goes back to Nocedal's fundamental work on L-BFGS [32] and leads to limited-memory quasi-Newton methods with memory parameter m. In this setting the initialization $H_{k,0}$ may depend on the current iterate k and is not necessarily equal to the original matrix H_0.

A crucial ingredient for the efficient numerical use of such methods is the existence of a compact representation of the form

$$H_k = H_{k,0} + Q_k M_k^{-1} Q_k^\top,$$

where $H_{k,0} \in \mathbb{S}^n_{++}$ is the initialization matrix, $Q_k \in \mathbb{R}^{n\times s}$ has only few columns ($s \ll n$), and $M_k \in \mathbb{S}^s$ is nonsingular. Byrd, Nocedal & Schnabel [11, Theorems 2.3 and 5.1] showed that the L-BFGS and L-SR1 matrices admit such a representation.

In order to derive an analogous representation for the Kleinmichel update, it is convenient to introduce the step and gradient-difference matrices

$$S_k \coloneqq \begin{bmatrix} d^0 & d^1 & \ldots & d^{k-1}\end{bmatrix} \in \mathbb{R}^{n\times k} \quad \text{and} \quad Y_k \coloneqq \begin{bmatrix} y^0 & y^1 & \ldots & y^{k-1}\end{bmatrix} \in \mathbb{R}^{n\times k}. \tag{29}$$

The following theorem provides a compact representation of H_k in terms of H_0, S_k and Y_k. It plays the same role for the Kleinmichel formula as the results in [11] do for the BFGS and SR1 updates. In a limited-memory implementation one simply replaces S_k and Y_k by the matrices containing the m most recent columns, exactly as in the L-BFGS and L-SR1 cases.

Theorem 5.2. *Let the symmetric matrix H_0 be updated k times by means of the Kleinmichel formula using the pairs $\{d^j, y^j\}_{j=0}^{k-1}$, and assume that each update is well-defined, namely $\langle y^j - \gamma_j H_j d^j, d^j\rangle \neq 0$, $j = 0, \ldots, k-1$. Then the resulting matrix H_k is given by*

$$H_k = \overline{\gamma}_k H_0 + Q_k M_k^{-1} Q_k^\top, \tag{30}$$

where S_k and Y_k are defined in (29), $\overline{\gamma}_k \coloneqq \prod_{i=0}^{k-1} \gamma_i$ for all $k \geq 1$, the nonsingular matrix M_k is defined recursively by

$$M_1 = \langle q^0, d^0\rangle, \quad M_{j+1} = \begin{bmatrix} \gamma_j^{-1} M_j & Q_j^\top d^j \\ (Q_j^\top d^j)^\top & \langle q^j, d^j\rangle \end{bmatrix}, \quad j \geq 1,$$

and $Q_k \coloneqq [q^0, \ldots, q^{k-1}] \coloneqq Y_k - \overline{\gamma}_k H_0 S_k$.

Proof. We proceed by induction. When $k = 1$ the right-hand side of (30) is

$$\overline{\gamma}_1 H_0 + Q_1 M_1^{-1} Q_1^\top = \gamma_0 H_0 + \frac{(y^0 - \gamma_0 H_0 d^0)(y^0 - \gamma_0 H_0 d^0)^\top}{\langle y^0 - \gamma_0 H_0 d^0, d^0\rangle} = H_1.$$

Let us now assume that (30) holds for some k. Moreover, define

$$\delta_k \coloneqq \langle q^k, d^k\rangle - \gamma_k (Q_k^\top d^k)^\top M_k^{-1} Q_k^\top d^k.$$

By definition of Q_k, $\overline{\gamma}_k$ and the formula for H_k, it holds that

$$\begin{aligned}\delta_k &:= \langle q^k, d^k\rangle - \gamma_k (Q_k^\top d^k)^\top M_k^{-1} Q_k^\top d^k \\ &= \langle y^k, d^k\rangle - \overline{\gamma}_{k+1}\langle d^k, H_0 d^k\rangle - \gamma_k (Q_k^\top d^k)^\top M_k^{-1} Q_k^\top d^k \\ &= \langle y^k, d^k\rangle - \gamma_k \langle d^k, \left(\overline{\gamma}_k H_0 + Q_k M_k^{-1} Q_k^\top\right) d^k\rangle \\ &= \langle y^k - \gamma_k H_k d^k, d^k\rangle \neq 0\end{aligned}$$

by assumption. Applying the Kleinmichel update to H_k we have

$$\begin{aligned}H_{k+1} &= \gamma_k \left(\overline{\gamma}_k H_0 + Q_k M_k^{-1} Q_k^\top\right) + \frac{(q^k - \gamma_k Q_k M_k^{-1} Q_k^\top d^k)(q^k - \gamma_k Q_k M_k^{-1} Q_k^\top d^k)^\top}{\langle q^k, d^k\rangle - \gamma_k \langle d^k, Q_k M_k^{-1} Q_k^\top d^k\rangle} \\ &= \overline{\gamma}_{k+1} H_0 + \frac{1}{\delta_k}\Big[\delta_k \gamma_k Q_k M_k^{-1} Q_k^\top + q^k (q^k)^\top - \gamma_k q^k (d^k)^\top Q_k M_k^{-1} Q_k^\top \\ &\qquad - \gamma_k Q_k M_k^{-1} Q_k^\top d^k (q^k)^\top + \gamma_k^2 (Q_k M_k^{-1} Q_k^\top d^k)(Q_k M_k^{-1} Q_k^\top d^k)^\top\Big],\end{aligned}$$

which can be expressed as

$$\begin{aligned}H_{k+1} &= \overline{\gamma}_{k+1} H_0 \\ &\quad + \frac{1}{\delta_k}\begin{bmatrix} Q_k & q^k\end{bmatrix}\begin{bmatrix}\gamma_k M_k^{-1}(\delta_k I + \gamma_k Q_k^\top d^k (d^k)^\top Q_k M_k^{-1}) & -\gamma_k M_k^{-1} Q_k^\top d^k \\ -\gamma_k (d^k)^\top Q_k M_k^{-1} & I\end{bmatrix}\begin{bmatrix} Q_k^\top \\ (q^k)^\top\end{bmatrix}. \qquad (31)\end{aligned}$$

By direct multiplication we obtain

$$\begin{bmatrix}\gamma_k^{-1} M_k & Q_k^\top d^k \\ (d^k)^\top Q_k & \langle q^k, d^k\rangle\end{bmatrix}\begin{bmatrix}\gamma_k M_k^{-1}(\delta_k I + \gamma_k Q_k^\top d^k (d^k)^\top Q_k M_k^{-1}) & -\gamma_k M_k^{-1} Q_k^\top d^k \\ -\gamma_k (d^k)^\top Q_k M_k^{-1} & I\end{bmatrix}\frac{1}{\delta_k} = I, \qquad (32)$$

since

$$\begin{aligned}\gamma_k (d^k)^\top Q_k M_k^{-1}\big(\delta_k I + \gamma_k Q_k^\top d^k (d^k)^\top Q_k M_k^{-1}\big) - \gamma_k (q^k)^\top d^k (d^k)^\top Q_k M_k^{-1} \\ = \gamma_k \big(\delta_k + \gamma_k (d^k)^\top Q_k M_k^{-1} Q_k^\top d^k - \langle q^k, d^k\rangle\big)(d^k)^\top Q_k M_k^{-1} = 0\end{aligned}$$

by definition of δ_k. Therefore, M_{k+1} is invertible with M_{k+1}^{-1} given by the second matrix in (32). However, this is also the matrix appearing in (31) and hence we see that (31) is exactly (30) with $k+1$ instead of k, which establishes the result. □

6 Algorithmic details

This section collects design choices and features for our implementation `RPQN` of Algorithm 1.

6.1 Termination condition

Algorithm 1 is complemented with a criterion for declaring that an iterate x^k is sufficiently close to stationarity. Upon a successful iteration, we compute the residual

$$r_k := \mu_k \|x^{k+1} - x^k\| = \mu_k \|d^k\|, \qquad (33)$$

which provides a metric for monitoring the quality of x^k; see Theorem 3.2 and [37, 21]. As soon as the residual r_k falls below a user-specified tolerance, Algorithm 1 returns the current iterate x^k and terminates.

6.2 Solution of the subproblems

Subproblems arising at Algorithm 1 are tackled in essentially the same way as in [26, Section 3.3], which exploits compact representations of quasi-Newton approximations to efficiently solve the scaled proximal subproblem (4). For completeness we briefly summarize the approach and refer to [26] for further details.

In each iteration of Algorithm 1 we must compute the solution of subproblem (4), which by Theorem 3.1 is equivalent to computing

$$\hat{x}^k \in \operatorname{prox}_{\varphi}^{G_k}\left(x^k - G_k^{-1}\nabla f(x^k)\right),$$

where $G_k \coloneqq B_k + \mu_k I$ and $B_k \approx \nabla^2 f(x^k)$. If B_k is obtained by a limited-memory BFGS, SR1, or Kleinmichel update, then it admits a compact representation

$$B_k = B_{k,0} + Q_k M_k^{-1} Q_k^\top, \tag{34}$$

see [26, Section 3.3.1] and Section 5. Following [26, Section 3.3.2], this representation can be rewritten in the form

$$B_k = B_{k,0} + U_{k,1}U_{k,1}^\top - U_{k,2}U_{k,2}^\top,$$

with matrices $U_{k,i} \in \mathbb{R}^{n\times r_i}$ of small rank $r_i > 0$ $(i = 1, 2)$. Consequently,

$$G_k = B_k + \mu_k I = B_{k,0} + \mu_k I + U_{k,1}U_{k,1}^\top - U_{k,2}U_{k,2}^\top, \tag{35}$$

so that we can apply the following result of Becker, Fadili & Ochs [5, Corollary 3.6], stated here akin to [26, Theorem 3.21].

Theorem 6.1. *Let $H = H_0 + U_1U_1^\top - U_2U_2^\top \in \mathbb{S}_{++}^n$, with $H_0 \in \mathbb{S}_{++}^n$ and $U_i \in \mathbb{R}^{n\times r_i}$ with rank r_i $(i = 1, 2)$. Set $H_1 = H_0 + U_1U_1^\top$. Then the following holds:*

$$\operatorname{prox}_{\varphi}^{H}(y) = \operatorname{prox}_{\varphi}^{H_0}\left(y + H_1^{-1}U_2\alpha_2^* - H_0^{-1}U_1\alpha_1^*\right),$$

where $\alpha_i^* \in \mathbb{R}^{r_i}$, $i = 1, 2$, are the unique zeros of the coupled system $\Xi(\alpha) = 0$, where $\Xi\colon \mathbb{R}^{r_1+r_2} \to \mathbb{R}^{r_1+r_2}$ is defined via

$$\Xi(\alpha) \coloneqq \begin{pmatrix} U_1^\top\left(y + H_1^{-1}U_2\alpha_2 - \operatorname{prox}_{\varphi}^{H_0}(y + H_1^{-1}U_2\alpha_2 - H_0^{-1}U_1\alpha_1)\right) + \alpha_1 \\ U_2^\top\left(y - \operatorname{prox}_{\varphi}^{H_0}(y + H_1^{-1}U_2\alpha_2 - H_0^{-1}U_1\alpha_1)\right) + \alpha_2 \end{pmatrix}.$$

Applied with $H \coloneqq G_k$ from representation (35) and $y \coloneqq x^k - G_k^{-1}\nabla f(x^k)$, Theorem 6.1 shows that, once $\operatorname{prox}_{\varphi}^{H_0}$ can be computed analytically (which is often the case when H_0 is a scaled identity matrix), the computation of $\hat{x}^k \in \operatorname{prox}_{\varphi}^{G_k}(y)$ reduces to solving the low-dimensional and strongly monotone system $\Xi(\alpha) = 0$ in $\mathbb{R}^{r_1+r_2}$, where r_1 and r_2 are typically very small compared with n.

The mapping Ξ in Theorem 6.1 is Newton differentiable for a variety of regularizers φ, and a generalized derivative of Ξ can be expressed explicitly in terms of a generalized derivative of $\operatorname{prox}_{\varphi}^{H_0}$; see [26, Proposition 3.22] for details. For many typical choices of φ, such as the ℓ_1- and ℓ_2-norms and group-sparsity penalties, this generalized derivative can be computed analytically.

6.3 Skipping quasi-Newton updates

Since the BFGS and Kleinmichel updates remain positive definite when $\langle d^k, y^k \rangle > 0$, it is standard practice to skip the update whenever

$$\langle d^k, y^k \rangle < \varepsilon_{\text{QN}} \|d^k\|^2$$

for some prescribed $\varepsilon_{\text{QN}} > 0$. In particular, for the Kleinmichel update we set $\gamma_k = \frac{\langle y^k, d^k \rangle}{2\langle d^k, H_k d^k \rangle}$ at each iteration $k \in \mathbb{N}_0$, in accordance with Theorem 5.1. Throughout the numerical experiments, we set $\varepsilon_{\text{QN}} = 10^{-8}$. In the case of the SR1 update, poorly conditioned steps are automatically excluded by a strategy described in [26, Section 3.3.2].

The initialization matrix $H_{k,0}$ for the limited-memory quasi-Newton updates is chosen as

$$H_{k,0} = \frac{\langle y^k, y^k \rangle}{\langle d^k, y^k \rangle} I,$$

following the recommendation of Liu & Nocedal [29]. In particular, when no curvature pairs are stored (i.e., zero memory), the resulting algorithm reduces to a pure proximal-gradient method.

6.4 Parameter updates

Algorithm 1 does not specify how to update the regularization parameter μ at successful iterations; it concedes any value in the user-specified interval $[\mu_{\min}, \mu_{\max}]$. Our implementation `RPQN` follows classical trust-region methods, distinguishing between successful and highly successful steps. Specifically, Algorithm 1 is executed according to

$$\mu_{k+1} = \begin{cases} \sigma_1 \mu_k & \text{if } \text{ared}_k \geq c_2 \text{pred}_k, \\ \mu_k & \text{otherwise} \end{cases} \tag{36}$$

for some fixed $\sigma_1 \in (0, 1)$ and $c_2 \in (c_1, 1)$. Included also in R2N [14], this strategy allows to reduce the regularization, possibly speeding up convergence, when the model seems to fit the actual problem particularly well. Otherwise, the regularization parameter is kept constant to avoid unwarranted unsuccessful iterations.

6.5 Nonmonotone globalization

In order to impose less conservative steps, with the aim of reducing the number of unsuccessful iterations, we incorporate a nonmonotone globalization mechanism. While [7, 8, 42, 21, 14] adopt the *max* kind of nonmonotonicity introduced by Grippo, Lampariello & Lucidi [16], we use instead the *average* scheme investigated in [43, 38, 12], which does not interfere with convergence properties of the monotone counterpart.

Let $\eta_{\text{nm}} \in (0, 1]$ be a given monotonicity factor and, for all $k \in \mathbb{N}$, define recursively the merit sequence $\{\Phi_k\}$ by

$$\Phi_k := \begin{cases} F(x^0) & \text{if } k = 0, \\ \eta_{\text{nm}} F(x^k) + (1 - \eta_{\text{nm}}) \Phi_{k-1} & \text{if } k > 0. \end{cases}$$

Then, by evaluating the actual improvement ared_k at Algorithm 1 according to

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

Algorithm 1 enforces sufficient decrease with respect to the merit Φ_k and not the current objective $F(x^k)$. The scheme falls back to monotone decrease when $\eta_{\text{nm}} = 1$, otherwise the fact that $\Phi_k \geq F(x^k)$ by [12, Lemma 4.2] relaxes the condition for accepting a tentative update $\hat{x}^k$, allowing larger steps and possibly faster convergence in practice.

7 Numerical results

In this section, we report numerical experiments for our implementation `RPQN` of Algorithm 1 on several instances of (P). We evaluate algorithmic variants with three limited-memory quasi-Newton updates (BFGS, SR1 and Kleinmichel) and (non)monotone globalization mechanism.

`RPQN` is compared against three other solvers designed to tackle (P). `SPG` implements a proximal-gradient method with spectral (or Barzilai–Borwein) stepsize [8, 17, 12], which often performs better than proximal-gradient methods with simple backtracking. `PANOCplus` is a proximal-gradient method that incorporates quasi-Newton steps to speed up convergence [37, 13]. Finally, `R2N` is a proximal quasi-Newton method that combines a second-order model with an inexact inner solve and an adaptive quadratic regularization [14]. Our implementation of solvers `SPG`, `PANOCplus` and `R2N` closely follows the original descriptions in [12], [13] and [14, 15], respectively, with minimal modifications to ensure comparability across methods. In particular, all solvers have a common termination criterion and (non)monotone globalization strategy, as presented in Section 6. `PANOCplus` and `R2N` use limited-memory BFGS with memory $m = 5$. A summary of all solver configurations is given in Table 1.

`SPG`	proximal-gradient method with spectral stepsize
`PANOCplus`	PANOC$^+$ with limited-memory BFGS
`R2N`	proximal quasi-Newton method with limited-memory BFGS
`RPQN-lkm`	RPQN with limited-memory Kleinmichel
`RPQN-lsr1`	RPQN with limited-memory SR1
`RPQN-lbfgs`	RPQN with limited-memory BFGS
`...-nm`	variant with averaged nonmonotone globalization

Table 1: Solver configurations considered for comparison in our numerical experiments.

For the sake of comparison, all solvers return based on the same termination criteria: residual r_k smaller than a tolerance (`tol` > 0) or exceeded time limit (300 wall-clock seconds).[1] Our tests consider both low (`tol` $= 10^{-3}$) and high (`tol` $= 10^{-5}$) accuracy.

All `RPQN` solvers run using an identical set of parameters: $c_1 = 10^{-4}$, $c_2 = 9/10$, $\sigma_1 = 1/2$, $\sigma_2 = 4$; see Section 6.4. For all limited-memory updates, we choose a memory of $m = 10$. For the solution of subproblems (explained in Section 6.2), we choose a tolerance of 10^{-9} as a stopping criterion. For the nonmonotone variants, we set $\eta_{\text{nm}} = 1/10$.

We compared also against [26, Algorithm 5.1], which requires convex φ and differs from Algorithm 1 only in an additional condition required to accept the update at Algorithm 1. Regardless of the sufficient decrease condition $\text{ared}_k \geq c_1 \text{pred}_k$, the candidate $\hat{x}^k$ is rejected if

$$\text{pred}_k \leq p_{\min} \|d^k\| \min\left\{\|G(x^k)\|, \|G(x^k)\|^\kappa\right\}$$

holds, for some user-specified $p_{\min} \in (0, 1/2)$, $\kappa > 1$ and where $G(x) := \text{prox}_\varphi(x - \nabla f(x)) - x$. We paired each `RPQN` variant with a counterpart that incorporates this condition, using the default values $p_{\min} = 10^{-8}$ and $\kappa = 2$ suggested in [26]. We expected and observed negligible differences in the practical performance of these `RPQN` variants.

Our experiments first consider convex logistic regression models with either convex or non-convex regularization (Section 7.1). We then turn to problems with the nonconvex Student's

[1] All computations were performed sequentially in MATLAB R2025b on a 64-bit Linux laptop. The hardware configuration was an Intel Core i7 processor, 16 GB RAM. Timings are intended for relative comparison only.

t-regression model, again with different regularizers (Section 7.2). Finally, we look at subproblems arising from an augmented Lagrangian method applied to a discretized obstacle problem (Section 7.3). We analyze the impact of different quasi-Newton updates, of the nonmonotone globalization, and of accuracy requirements.

Numerical results are reported as median wall-clock runtimes in tables, where we highlight the method with the best performance for each problem class and accuracy level. Solvers are also compared in terms of runtime by means of profiles. Given a set P of problem instances and a set S of solvers, let $\tau_{s,p}$ denote the runtime of solver $s \in S$ for problem $p \in P$. When solver s fails on problem p, we set $\tau_{s,p} = \infty$. Given a baseline solver $\bar{s} \in S$, we adopt *relative profiles* to compare the performance of different solver variants. Each (relative) profile depicts the empirical distribution $\varrho_{s,\bar{s}} \colon [0,\infty) \mapsto [0,1]$ of the ratio $\tau_{s,\cdot}/\tau_{\bar{s},\cdot}$ over P, namely

$$\forall \kappa \in [0,\infty) \colon \quad \varrho_{s,\bar{s}}(\kappa) := \frac{|\{p \in P \,|\, \tau_{s,p} \leq \kappa \tau_{\bar{s},p}\}|}{|P|},$$

where the sample size $|P|$ is the cardinality of set P. As such, a relative profile displays the fraction of problems $\varrho_{s,\bar{s}}(\kappa)$ solved by s within a factor κ compared to the baseline solver $\bar{s}$.

7.1 Regularized logistic regression

A common model for solving sparse binary classification problems is regularized logistic regression. Given feature vectors $a_i \in \mathbb{R}^{n_\mathrm{f}}$ and labels $b_i \in \{-1, 1\}$ for $i = 1, \ldots, n_\mathrm{s}$, we aim to

$$\underset{y,v}{\text{minimize}} \quad \frac{1}{n_\mathrm{s}} \sum_{i=1}^{n_\mathrm{s}} \log\Big(1 + \exp\big(-b_i(a_i^\top y + v)\big)\Big) + \varphi(y) \tag{37}$$

with respect to $y \in \mathbb{R}^{n_\mathrm{f}}$ and $v \in \mathbb{R}$. In common test instances there are more samples than features ($n_\mathrm{s} \gg n_\mathrm{f}$). A sparse representation is encouraged by including a regularizer φ. First, we let $\varphi(y) := \lambda\|y\|_1$ for some given regularization parameter $\lambda > 0$. Then, we consider the (nonconvex) capped ℓ_1 regularizer, a piecewise linear approximation of ℓ_0 given by $\varphi(y) := \lambda \sum_{i=1}^{n_\mathrm{f}} g_1(y_i)$ where $g_1(t) := \min\{|t|, 1\}$.[2]

Setup We generate sparse instances with $n_\mathrm{f} = 10^4$ and $n_\mathrm{s} = 10^5$, and vary the feature sparsity parameter $s \in \{10, 100\}$. For each i, the vector a_i has approximately s nonzero components; the locations are chosen at random and the nonzero entries are drawn independently from a standard normal distribution $\mathcal{N}(0,1)$. We sample a ground-truth vector $y^\mathrm{true} \in \mathbb{R}^{n_\mathrm{f}}$ with $10s$ nonzero entries, drawn again from $\mathcal{N}(0,1)$, and a bias $v^\mathrm{true} \sim \mathcal{N}(0,1)$. Labels are then assigned via

$$b_i = \operatorname{sign}\left(a_i^\top y^\mathrm{true} + v^\mathrm{true} + \xi_i\right),$$

where $\xi_i \sim \mathcal{N}(0, 1/10)$ independently for $i = 1, \ldots, n_\mathrm{s}$.

We parametrize λ as $\lambda := c_\lambda \lambda_\mathrm{max}$ with $c_\lambda \in \{10^{-1}, 10^{-2}, 10^{-3}\}$. Following [25], we denote by n_s^+ and n_s^- the number of indices with $b_i = 1$ and $b_i = -1$, respectively, and choose

$$\lambda_\mathrm{max} = \frac{1}{n_\mathrm{s}} \left\| \frac{n_\mathrm{s}^-}{n_\mathrm{s}} \sum_{i:\, b_i = 1} a_i + \frac{n_\mathrm{s}^+}{n_\mathrm{s}} \sum_{i:\, b_i = -1} a_i \right\|,$$

which is the smallest value for which the trivial solution $y^* = 0$ (with some v^*) is optimal.

[2] The proximal mapping of these regularizers can be found online at proximity-operator.net.

For each combination of (s, c_λ), we generate 10 independent data sets and run each solver variant, starting from an all-zero vector $x^0 \in \mathbb{R}^{n_f+1}$.

Results Results for the nonmonotone variants are reported in Figure 1 and Table 2. We omit the monotone variants as they performed similarly or worse, as illustrated below. The `RPQN-nm` configurations exhibit comparable practical performance, across regularizers, accuracy levels and limited-memory updates. `RPQN` consistently outperforms the other solvers, with a larger gap for high accuracy. In this setting, Figure 1 indicates that `RPQN-nm` solvers are faster than `R2N-nm` in 95% of the instances, and about 2x faster in 80% of the instances, while `SPG-nm` and `PANOCplus-nm` lag behind.

regularizer	ℓ_1		capped ℓ_1	
accuracy	low	high	low	high
`SPG-nm`	1.195	3.197	2.378	63.689
`PANOCplus-nm`	2.043	45.856	2.190	162.584
`R2N-nm`	0.612	1.861	0.694	2.903
`RPQN-lkm-nm`	0.272	0.565	**0.355**	**0.565**
`RPQN-lsr1-nm`	0.333	0.500	0.367	0.652
`RPQN-lbfgs-nm`	**0.263**	**0.471**	0.364	0.579

Table 2: Comparison on regularized logistic regression, in terms of median runtimes [s]. Sample size: 60 problems for each regularizer and accuracy level.

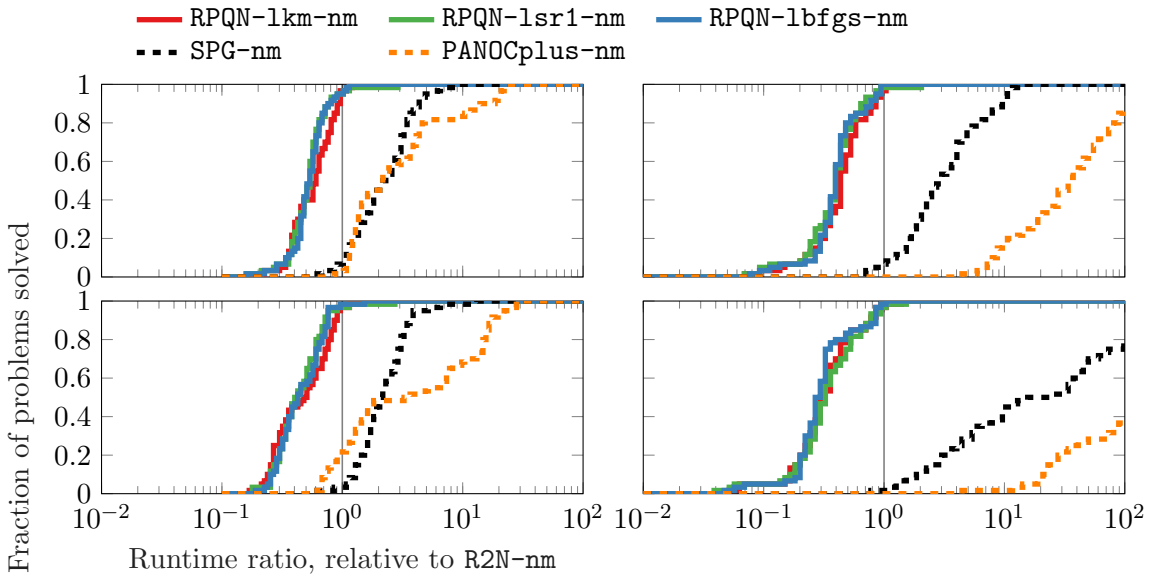

Figure 1: Comparison on regularized logistic regression problems, in terms of relative runtimes. Accuracy level: low (left) and high (right). Regularization: ℓ_1 (top) and capped-ℓ_1 (bottom).

The effect of adopting a nonmonotone globalization is illustrated with pair-wise relative runtime profiles in Figure 2, aggregating all logistic regression instances. All methods appear to benefit from the nonmonotone mechanism, for both low and high accuracy, but especially in the latter setting. `RPQN-lbfgs` shows little improvement, whereas `RPQN-lkm-nm` is at least 3x faster

than `RPQN-lkm` in 60% of the instances. Since similar results were also obtained for the problems considered below, we do not report further on the performance of the monotone variants.

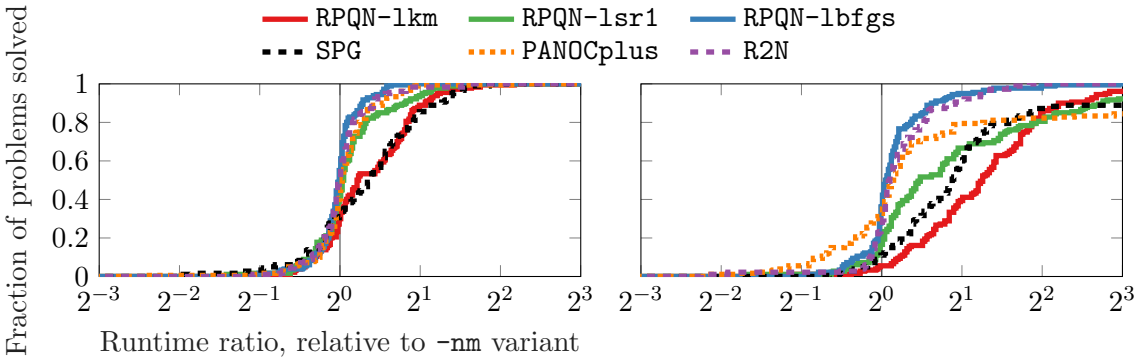

Figure 2: Comparison of solver variants with monotone and nonmonotone globalization, in terms of relative runtimes. Results aggregated for all logistic regression problems. Accuracy level: low (left) and high (right). Sample size: 120 problems for each accuracy level.

7.2 Regularized Student's t-regression

We consider the regularized Student's t-regression problem, which seeks to

$$\underset{x}{\text{minimize}} \quad \frac{1}{m}\sum_{i=1}^{m} \log\left(1 + \frac{(Ax-b)_i^2}{\nu}\right) + \varphi(x), \tag{38}$$

with respect to $x \in \mathbb{R}^n$, where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, and $\nu > 0$. The loss term is smooth, but generally nonconvex, while th regularizer φ is added to promote sparsity. As before for the logistic regression, we let $\varphi(x) := \lambda\|x\|_1$ and then $\varphi(x) := \lambda\sum_{i=1}^n g_1(x_i)$ for some weight $\lambda > 0$.

Setup The test instances are generated as in [30, 41]. We set $m := 1024$, $n := 8m$, and $\nu := 1/4$. Matrix A is defined via subsampled discrete cosine measurements: let $D \in \mathbb{R}^{n\times n}$ denote the discrete cosine transform (DCT) matrix, pick an index set $J \subset \{1,\dots,n\}$ uniformly at random with $|J| = m$, and define $Ax := (Dx)_J$. We generate a sparse ground-truth signal $x^{\text{true}} \in \mathbb{R}^n$ with $s := \lfloor \frac{n}{40} \rfloor$ nonzero entries at uniformly random locations. The nonzero amplitudes are chosen according to $x_i^{\text{true}} := \eta_{1,i}10^{\eta_{2,i}d}$, where $\eta_{1,i} \in \{-1,1\}$ is a random sign and $\eta_{2,i}$ is uniformly distributed on $[0,1]$. The parameter $d \in \{2,3,4\}$ controls the dynamic range. The observation vector is then given by $b := Ax^{\text{true}} + \xi/10$, where the components of ξ are drawn independently from a Student's t-distribution with 4 degrees of freedom. We choose the regularization parameter as $\lambda := c_\lambda\|\nabla f(0)\|_\infty$ with $c_\lambda \in \{10^{-1}, 10^{-2}\}$. For each combination of d and c_λ we generate 10 independent instances and start all methods from $x^0 := A^\top b$.

Results Figure 3 and Table 3 summarize the results obtained on Student's t-regression problems with convex and nonconvex regularizer. The `RPQN-nm` configurations have similar performance for low accuracy, but `RPQN-lsr1-nm` remains behind for high accuracy. `RPQN-lbfgs-nm` and `RPQN-lkm-nm` are consistently faster than the other solvers. Figure 3 indicates that, for high accuracy, `RPQN-lkm-nm` is at least 2x faster than `R2N-nm` on 85% of the instances.

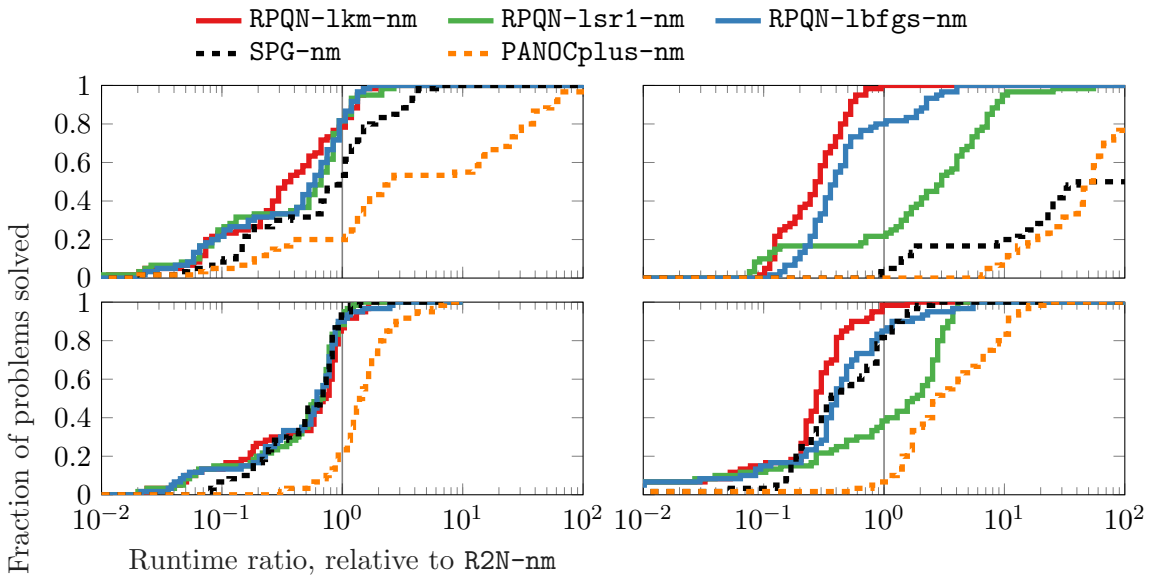

Figure 3: Comparison on regularized Student's t-regression problems, in terms of relative runtimes. Accuracy level: low (left) and high (right). Regularization: ℓ_1 (top) and capped-ℓ_1 (bottom).

regularizer	ℓ_1		capped ℓ_1	
accuracy	low	high	low	high
SPG-nm	0.004	47.992	**0.002**	0.138
PANOCplus-nm	0.006	24.737	0.003	1.529
R2N-nm	0.035	0.671	0.003	0.371
RPQN-lkm-nm	0.003	**0.122**	**0.002**	**0.096**
RPQN-lsr1-nm	**0.002**	1.340	**0.002**	0.851
RPQN-lbfgs-nm	**0.002**	0.199	**0.002**	0.146

Table 3: Comparison on regularized Student's t-regression, in terms of median runtimes [s]. Sample size: 60 problems for each regularizer and accuracy level.

7.3 Obstacle problem

We consider the optimal control of a discretized obstacle problem as investigated in [17, Example 6.2]. The problem is to

$$\begin{aligned} \underset{u,v,z\in\mathbb{R}^N}{\text{minimize}} \quad & \frac{1}{2}\|u\|^2 + \frac{1}{2}\|v\|^2 - \sum_{i=1}^{N} v && \text{(39a)} \\ \text{subject to} \quad & u + \mathcal{A}v - z = 0 && \text{(39b)} \\ & u \geq 0, \quad v \geq 0, \quad z \geq 0, \quad v^\top z = 0, && \text{(39c)} \end{aligned}$$

where $\mathcal{A} \in \mathbb{R}^{N\times N}$ is a tridiagonal matrix that arises from a discretization of the negative Laplace operator in one dimension, i.e., $a_{ii} = 2$ for all i and $a_{ij} = -1$ for all $i = j \pm 1$. The nonnegativity and complementarity constraints in (39c) can be represented as the inclusion of (u, v, z) in a set $X \subset \mathbb{R}^{3N}$, which happens to have an easy-to-evaluate projection operator. Problem (39) has the unique solution $u^* = v^* = z^* = 0$, whose degeneracy makes it hard to find.

Using $x := (u, v, z)$, problem (39) is equivalently rewritten as

$$\underset{x \in \mathbb{R}^{3N}}{\text{minimize}} \quad f_0(x) + \varphi(x) \quad \text{subject to} \quad Ax = 0 \tag{40}$$

where f_0 denotes the quadratic cost in (39a), $A \in \mathbb{R}^{N \times 3N}$ stems from the linear constraint in (39b), and setting φ as the indicator of set X encodes the inclusion $x \in X$ for the constraints in (39c). Following [17], the augmented Lagrangian function for (40) reads

$$L_\mu(x, y) := f_0(x) + \frac{1}{2\mu}\|Ax + \mu y\|^2 + \varphi(x) \tag{41}$$

and has to be minimized for some given $\mu > 0$ and $y \in \mathbb{R}^N$. This is clearly of the form (P), with convex quadratic f and nonsmooth nonconvex φ.

Setup We generate instances with $N \in \{2^7, 2^8, 2^9, 2^{10}, 2^{11}, 2^{12}\}$ and $\mu \in \{10^{-1}, 10^{-2}, 10^{-3}\}$. For each combination of N and μ we draw 10 independent samples of $y \in \mathbb{R}^N$ and $x^0 \in \mathbb{R}^{3N}$ from a normal distribution, for a total of 180 calls to each solver for each accuracy level.

Results Numerical results are reported in Figure 4 and Table 4. The `RPQN-nm` variants perform similarly across limited-memory updates and accuracy levels, and consistently better than the other solvers. For both low and high accuracy, all `RPQN-nm` variants are at least 3x faster than `R2N-nm` in 80% of the instances. Compared to the numerical experiments in [17, §6.1], `RPQN-nm` handles larger instances (from $N = 2^6$ in [17] to $N = 2^{12}$ here) with a fraction of the effort.

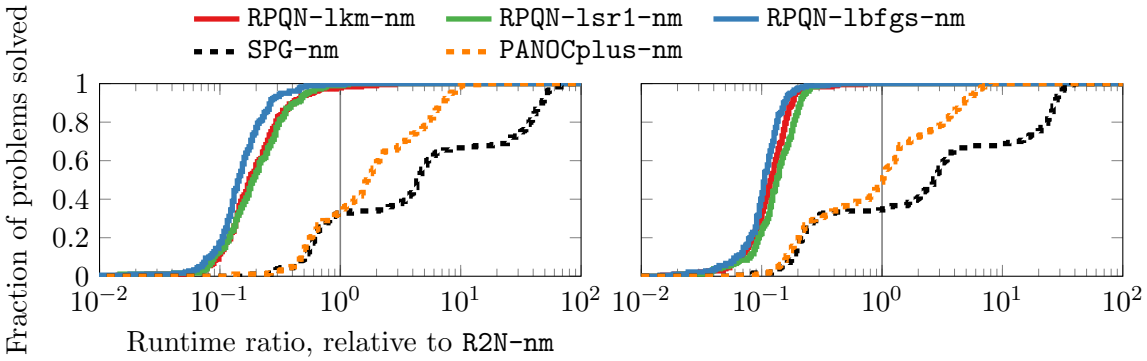

Figure 4: Comparison on discretized obstacle problems, in terms of relative runtimes. Accuracy level: low (left) and high (right).

8 Conclusion

A proximal method with adaptive quadratic regularization and limited-memory quasi-Newton models was introduced in this work. Global convergence was established under weak assumptions; in particular, no global Lipschitz continuity is required. Moreover, convergence of the iterates is shown under the Kurdyka–Łojasiewicz property and a convergence rate was derived for the subsequence of successful iterates.

We also developed a compact representation for the limited-memory Kleinmichel update—a rank-one quasi-Newton formula guaranteeing positive definiteness under suitable assumptions—making it applicable within our subproblem solver. Numerical experiments indicate that the

accuracy	low	high
`SPG-nm`	0.615	1.044
`PANOCplus-nm`	0.312	0.387
`R2N-nm`	0.136	0.332
`RPQN-lkm-nm`	0.033	0.045
`RPQN-lsr1-nm`	0.034	0.054
`RPQN-lbfgs-nm`	**0.026**	**0.044**

Table 4: Comparison on discretized obstacle problems, in terms of median runtimes [s]. Sample size: 180 problems for each accuracy level.

Kleinmichel update can be effective in practice, sometimes performing better than the well-known BFGS and SR1 updates. More in general, enabled by the solution of proximal quasi-Newton subproblems through their compact representation, our numerical approach efficiently incorporates curvature information without costly inner iterations. Comparisons against other recently developed solvers demonstrated the validity of the proposed numerical scheme, witnessing reliable and fast convergence in practice.

Our findings suggest several directions for future research. First, our analysis does not establish eventual successfulness of the iterations prior to convergence of the iterates. It would be interesting to investigate whether there are problem instances for which this property fails. Second, further study of the Kleinmichel update appears promising: it is theoretically superior to SR1, and our experiments suggest that it can also yield improved performance in practice.

Declarations

Data availability The test problems in Section 7 are based on randomly generated data. The code used for data generation and numerical experiments is archived on Zenodo and openly available at DOI: 10.5281/zenodo.20025657.

Conflict of interest The authors declare that they have no conflict of interest to this work.

References

[1] F. J. Aragón Artacho, R. M. T. Fleming, and P. T. Vuong. Accelerating the DC algorithm for smooth functions. *Mathematical Programming*, 169(1):95–118, 2018. doi:10.1007/s10107-017-1180-1.

[2] A. Y. Aravkin, R. Baraldi, and D. Orban. A proximal quasi-Newton trust-region method for nonsmooth regularized optimization. *SIAM Journal on Optimization*, 32(2):900–929, 2022. doi:10.1137/21M1409536.

[3] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality. *Mathematics of Operations Research*, 35:438–457, 2010. doi:10.1287/moor.1100.0449.

[4] S. Becker and J. Fadili. A quasi-Newton proximal splitting method. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

[5] S. Becker, J. Fadili, and P. Ochs. On quasi-Newton forward-backward splitting: Proximal calculus and convergence. *SIAM Journal on Optimization*, 29(4):2445–2481, 2019. doi:10.1137/18M1167152.

[6] W. Bian and X. Chen. Linearly constrained non-Lipschitz optimization for image restoration. *SIAM Journal on Imaging Sciences*, 8(4):2294–2322, 2015. doi:10.1137/140985639.

[7] E. G. Birgin, J. M. Martínez, and M. Raydan. Nonmonotone spectral projected gradient methods on convex sets. *SIAM Journal on Optimization*, 10(4):1196–1211, 2000. doi:10.1137/S1052623497330963.

[8] E. G. Birgin, J. M. Martínez, and M. Raydan. Spectral projected gradient methods: Review and perspectives. *Journal of Statistical Software*, 60(3):1–21, 2014. doi:10.18637/jss.v060.i03.

[9] J. Bolte, A. Daniilidis, A. Lewis, and M. Shiota. Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization*, 18:556–572, 2007. doi:10.1137/060670080.

[10] J. Bolte, T. Nguyen, J. Peypouquet, and B. Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165, 10 2016. doi:10.1007/s10107-016-1091-6.

[11] R. H. Byrd, J. Nocedal, and R. B. Schnabel. Representations of quasi-Newton matrices and their use in limited memory methods. *Mathematical Programming*, 63(1):129–156, 1994. doi:10.1007/BF01582063.

[12] A. De Marchi. Proximal gradient methods beyond monotony. *Journal of Nonsmooth Analysis and Optimization*, 4, 2023. doi:10.46298/jnsao-2023-10290.

[13] A. De Marchi and A. Themelis. Proximal gradient algorithms under local Lipschitz gradient continuity. *Journal of Optimization Theory and Applications*, 194(3):771–794, 2022. doi:10.1007/s10957-022-02048-5.

[14] Y. Diouane, M. L. Habiboullah, and D. Orban. A proximal modified quasi-Newton method for nonsmooth regularized optimization. *SIAM Journal on Optimization*, 36(2):534–563, 2026. doi:10.1137/24M169761X.

[15] M. Gollier, M. L. Habiboullah, G. Leconte, R. Baraldi, A. De Marchi, D. Orban, and Y. Diouane. RegularizedOptimization.jl: A Julia framework for regularized and nonsmooth optimization. *Journal of Open Source Software*, 11(118):9344, 2026. doi:10.21105/joss.09344.

[16] L. Grippo, F. Lampariello, and S. Lucidi. A nonmonotone line search technique for Newton's method. *SIAM Journal on Numerical Analysis*, 23(4):707–716, 1986. doi:10.1137/0723046.

[17] X. Jia, C. Kanzow, and P. 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. doi:10.1137/23M1548293.

[18] Q. Jin. *Lectures on Nonsmooth Optimization*. Springer, 2025. doi:10.1007/978-3-031-91417-1.

[19] C. Kanzow and T. Lechner. Efficient regularized proximal quasi-Newton methods for large-scale nonconvex composite optimization problems. *Pacific Journal of Optimization*, 20:537–568, 2024. doi:10.61208/pjo-2023-036.

[20] C. Kanzow and L. Lehmann. Convergence of nonmonotone proximal gradient methods under the Kurdyka-Łojasiewicz property without a global Lipschitz assumption. *Journal of Optimization Theory and Applications*, 207(1):4, 2025. doi:10.1007/s10957-025-02762-w.

[21] C. Kanzow and P. Mehlitz. Convergence properties of monotone and nonmonotone proximal gradient methods revisited. *Journal of Optimization Theory and Applications*, 195:1–23, 2022. doi:10.1007/s10957-022-02101-3.

[22] C. Kanzow and D. Steck. Regularization of limited memory quasi-Newton methods for large-scale nonconvex minimization. *Mathematical Programming Computation*, 15(3):417–444, 2023. doi:10.1007/s12532-023-00238-4.

[23] H. Kleinmichel. Quasi-Newton-Verfahren vom Rang-Eins-Typ zur Lösung unrestringierter Minimierungsprobleme, Teil 1. *Numerische Mathematik*, 38(2):219–228, 1981. doi:10.1007/BF01397091.

[24] H. Kleinmichel. Quasi-Newton-Verfahren vom Rang-Eins-Typ zur Lösung unrestringierter Minimierungsprobleme, Teil 2. *Numerische Mathematik*, 38(2):229–244, 1981. doi:10.1007/BF01397092.

[25] K. Koh, S.-J. Kim, and S. Boyd. An interior-point method for large-scale l1-regularized logistic regression. *Journal of Machine Learning Research*, 8(54):1519–1555, 2007.

[26] T. Lechner. *Proximal Methods for Nonconvex Composite Optimization Problems*. PhD Thesis, Institute of Mathematics, University of Würzburg, 2022. doi:10.25972/OPUS-28907.

[27] J. Lee, Y. Sun, and M. Saunders. Proximal Newton-type methods for minimizing composite functions. *SIAM Journal on Optimization*, 24:1420–1443, 2014. doi:10.1137/130921428.

[28] H. Li and Z. Lin. Accelerated proximal gradient methods for nonconvex programming. In *Advances in Neural Information Processing Systems 28*, volume 28, 2015.

[29] D. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45:503–528, 1989. doi:10.1007/BF01589116.

[30] R. Liu, S. Pan, Y. Wu, and X. Yang. An inexact regularized proximal Newton method for nonconvex and nonsmooth optimization. *Computational Optimization and Applications*, 88(2):603–641, 2024. doi:10.1007/s10589-024-00560-0.

[31] B. Mordukhovich. *Variational Analysis and Applications*, volume 30. Springer, 2018. doi:10.1007/978-3-319-92775-6.

[32] J. Nocedal. Updating quasi-Newton matrices with limited storage. *Mathematics of Computation*, 35(151):773–782, 1980. doi:10.1090/S0025-5718-1980-0572855-7.

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

[34] L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60:259–268, 1992. doi:10.1016/0167-2789(92)90242-F.

[35] K. Scheinberg and X. Tang. Practical inexact proximal quasi-Newton method with global complexity analysis. *Mathematical Programming*, 160(1–2):495–529, 2016. doi:10.1007/s10107-016-0997-3.

[36] P. Spellucci. A modified rank one update which converges Q-superlinearly. *Computational Optimization and Applications*, 19:273–296, 2001. doi:10.1023/A:1011259905470.

[37] L. Stella, A. Themelis, P. Sopasakis, and P. Patrinos. A simple and efficient algorithm for nonlinear model predictive control. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 1939–1944, 2017. doi:10.1109/CDC.2017.8263933.

[38] A. Themelis, L. Stella, and P. Patrinos. Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone linesearch algorithms. *SIAM Journal on Optimization*, 28(3):2274–2303, 2018. doi:10.1137/16M1080240.

[39] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. doi:10.1111/j.2517-6161.1996.tb02080.x.

[40] K. Ueda and N. Yamashita. A regularized Newton method without line search for unconstrained optimization. *Computational Optimization and Applications*, 59(1):321–351, 2014. doi:10.1007/s10589-014-9656-x.

[41] S. vom Dahl and C. Kanzow. An inexact regularized proximal Newton method without line search. *Computational Optimization and Applications*, 89(3):585–624, 2024. doi:10.1007/s10589-024-00600-9.

[42] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, 2009. doi:10.1109/TSP.2009.2016892.

[43] H. Zhang and W. W. Hager. A nonmonotone line search technique and its application to unconstrained optimization. *SIAM Journal on Optimization*, 14(4):1043–1056, 2004. doi:10.1137/S1052623403428208.