LeAP-SSN: A semismooth Newton method with global convergence rates

Amal Alphonse, Pavel Dvurechensky, Ioannis P. A. Papadopoulos,

Clemens Sirotenko

Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: amal.alphonse@wias-berlin.de
        pavel.dvurechensky@wias-berlin.de
        ioannis.papadopoulos@wias-berlin.de
        clemens.sirotenko@wias-berlin.de

LeAP-SSN: A semismooth Newton method with global convergence rates

Amal Alphonse, Pavel Dvurechensky, Ioannis P. A. Papadopoulos,
Clemens Sirotenko

Abstract

We propose LeAP-SSN (Levenberg–Marquardt Adaptive Proximal Semismooth Newton method), a semismooth Newton-type method with a simple, parameter-free globalisation strategy that guarantees convergence from arbitrary starting points in nonconvex settings to stationary points, and under a Polyak–Łojasiewicz condition, to a global minimum, in Hilbert spaces. The method employs an adaptive Levenberg–Marquardt regularisation for the Newton steps, combined with backtracking, and does not require knowledge of problem-specific constants. We establish global nonasymptotic rates: $\mathcal{O}(1/k)$ for convex problems in terms of objective values, $\mathcal{O}(1/\sqrt{k})$ under nonconvexity in terms of subgradients, and linear convergence under a Polyak–Łojasiewicz condition. The algorithm achieves superlinear convergence under mild semismoothness and Dennis–Moré or partial smoothness conditions, even for non-isolated minimisers. By combining strong global guarantees with superlinear local rates in a fully parameter-agnostic framework, LeAP-SSN bridges the gap between globally convergent algorithms and the fast asymptotics of Newton's method. The practical efficiency of the method is illustrated on representative problems from imaging, contact mechanics, and machine learning.

1 Introduction

In this paper, we consider the composite minimisation problem

$$\min_{x \in \mathcal{H}} \{F(x) \coloneqq f(x) + \psi(x)\}, \tag{P}$$

where $\mathcal{H}$ is a Hilbert space, $\psi \colon \mathcal{H} \to \overline{\mathbb{R}}$ is a potentially nonsmooth proper convex lower semicontinuous function, and $f \in C^{1,1}(\mathcal{H}; \mathbb{R})$ is a (possibly nonconvex) function such that the Fréchet derivative f' is not necessarily smooth. In particular, we are interested in the setting when f' possesses some semismoothness-like properties and our main goal is to propose a variant of the semismooth Newton (SSN) method that enjoys both global convergence rate guarantees and local superlinear convergence. Note that when $\psi \equiv 0$ and f' is semismooth, the optimality condition for (P) is given by the nonlinear and nonsmooth equation $f'(x) = 0$ with the semismooth operator $f' \colon \mathcal{H} \to \mathcal{H}^*$.

The SSN method is a generalisation of the classical Newton method used for solving (systems of) nonsmooth equations and for the minimisation of functions with non-Lipschitz second derivatives, see for example the pioneering works [52, 22, 85, 86, 95, 49]. SSN methods are especially prominent when solving infinite-dimensional problems, e.g., in PDE-constrained optimisation, due to their mesh-independent and fast local convergence properties. There are recent theoretical results for the method in the context of optimisation problems; these include local superlinear convergence under (local) strong convexity or non-degeneracy and global asymptotic convergence for its variants with a Levenberg–Marquardt regularisation and/or Armijo-type linesearch strategy [84, 83, 55, 73, 98]. At the

same time, global convergence rates remain unknown for the SSN method and its variants, even in the simpler setting where $\psi \equiv 0$. In this paper, we close this theoretical gap and propose a novel variant of the globalised SSN method which we abbreviate as LeAP-SSN (Levenberg–Marquardt Adaptive Proximal Semismooth Newton, Algorithm 1) that combines the best of both worlds. Namely, it allows us to prove global convergence rates in several settings, including the convex setting and a nonconvex setting (with better rates under a Polyak–Łojasiewicz (PL) condition), and to prove superlinear convergence under semismoothness and a Dennis–Moré-type condition. In addition, in finite dimensions, we also prove that our algorithm identifies a C^2-manifold in which eventually all iterates lie, and use this and the theory of partial smoothness and active manifolds to prove superlinear convergence without requiring a Dennis–Moré assumption. We conclude the paper with some numerical experiments.

Our work, in a sense, lies in the interface of research questions studied in the (PDE-constrained) nonsmooth optimisation and (smooth) convex optimisation communities, and we hope that this paper can start an avenue for other global convergence rate results for SSN and its variants.

1.1 Contributions

On a high level, for the problem (P), we propose a SSN-type algorithm LeAP-SSN (see Algorithm 1) with guaranteed global convergence rates and local superlinear convergence. Our approach employs proximal steps and a Levenberg–Marquardt globalisation strategy with backtracking inspired by [31], that avoids relying on second-order semismoothness assumptions as required in [84, 83, 98]. Unlike classical trust-region methods, our globalisation mechanism ensures a practical and simple implementation while retaining theoretical guarantees. In each iteration, our algorithm constructs a model $\widetilde{F}$ of the objective F combining a quadratic model of f, the functional ψ, and a quadratic regularisation. Our main assumption on these subproblems, which holds even in some cases when F is nonconvex, is that it is possible to find a stationary point of $\widetilde{F}$ that does not increase $\widetilde{F}$, see Assumption 5 for a formal statement. We also introduce a mild assumption on f', see Assumption 2, which is key to global convergence rates, and which holds for a large class of functions and settings. Under Assumption 2, we obtain a series of global and local nonasymptotic convergence rates:

(1) In the nonconvex case, we prove that the sequence $(x_k)_{k \in \mathbb{N}}$ generated by our algorithm from arbitrary $x_0 \in \mathrm{dom}\, F$ satisfies[1] $\min_{0 \leq i \leq k-1} \|F'(x_{i+1})\|_* = \mathcal{O}(1/\sqrt{k})$, see Theorem 4.1.

(2) If additionally F satisfies a Polyak–Łojasiewicz (PL) condition (see Assumption 3), we prove that x_k converges to a global solution x^* of (P) with a linear rate from arbitrary $x_0 \in \mathrm{dom}\, F$. Moreover, $F(x_k) - F(x^*)$ converges to 0 linearly, see Theorem 4.4.

(3) If F is convex, we prove that for arbitrary $x_0 \in \mathrm{dom}\, F$ our algorithm guarantees $F(x_k) - F^* = \mathcal{O}(1/k)$, where F^* is the global minimal value in (P), see Theorem 4.9.

(4) If for some isolated local minimum x^* in (P), Assumption 2 holds only locally around x^*, the PL and quadratic growth (QG) conditions (see (53)) also hold locally around x^*, and the model $\widetilde{F}$ is strongly convex, we prove that $x_k \to x^*$ and $F(x_k) - F(x^*) \to 0$ linearly provided that x_0 is sufficiently close to x^*, see Theorem 5.1. In other words, we prove that our algorithm has local linear convergence.

When additionally f' has higher regularity and F is locally convex, we show that the above bounds

---

[1]Here and below $a_k = \mathcal{O}(b_k)$ means that there exists an explicit constant c such that $a_k \leq c b_k$ for all $k \geq 0$.

asymptotically improve. To be precise, the linear convergence in the above results improves to super-linear in each of the following cases:

(5) If $\mathcal{H}$ is finite-dimensional and f is C^2 (i.e., twice continuously differentiable) around x^*, see Theorem 6.5.

(6) If f' is semismooth at x^*, and a Dennis–Moré-type condition is satisfied, see Theorem 6.10 and Theorem 6.12.

(7) If f is C^2 and convex, see Corollary 6.11.

(8) If $\mathcal{H}$ is finite-dimensional, f' is semismooth around x^*, F is partially smooth at x^* relative to a C^2 manifold, and a non-degeneracy (or strict complementary) condition is satisfied, see Theorem 6.19.

Thus, in particular, we obtain local superlinear convergence of our algorithm as for classical SSN, see Theorem 6.12. The critical ingredient for these superlinear convergence results is to show that the regularisation parameter adaptively chosen in the algorithm asymptotically goes to zero, meaning that asymptotically, our method becomes classical SSN. As a result, we show that semismoothness allows for obtaining stronger local and global asymptotic convergence properties. Our results are interesting also in finite dimensions since they establish faster asymptotic convergence under higher regularity. The above global results are summarised for convenience in Table 1. We highlight that, crucially, our

Assumption	Global convergence rate
-	$\mathcal{O}(1/\sqrt{k})$ for $\min_{i \leq k}\|F'(x_i)\|_*$
$f \in C^2, \dim(\mathcal{H}) < \infty, x_k \to x^*$	$o(1/\sqrt{k})$ for $\min_{i \leq k}\|F'(x_i)\|_*$
f convex	$\mathcal{O}(1/k)$ for $F(x_k)$
f convex, $f \in C^2, \dim(\mathcal{H}) < \infty, x_k \to x^*$	$o(1/k)$ for $F(x_k)$
PL	Linear for x_k and $F(x_k)$
PL, $f \in C^2, \dim(\mathcal{H}) < \infty, x_k \to x^*$	Superlinear
PL, $f \in C^2, f$ convex	Superlinear
PL, f semismooth, DM-type condition, F locally convex, $H \succeq 0$	Superlinear
PL, f semsimooth, partial smoothness, F locally convex, $H \succeq 0$	Superlinear

Table 1: High-level summary of our main results under Assumptions 1, 2 and 5.

algorithm achieves

- **superlinear convergence under the PL condition to *non-isolated*[2] minima**, whereas most existing works guarantee superlinear convergence only under strong convexity at the limit point. Convergence to non-isolated solutions is particularly relevant in applications such as inverse problems, where non-isolated stationary points naturally arise (for instance, when the forward operator is non-injective, as is common in e.g. medical image reconstruction). Local superlinear convergence in such settings is an active area of research; see, e.g., [88] for recent results in the smooth case. To the best of our knowledge, no existing work establishes *superlinear* convergence *to non-isolated solutions* in the setting we consider.

---

[2]By a *non-isolated* minimiser, we mean a minimiser that we do not assume to be isolated.

- **automatic strong convergence of the iterates in Hilbert spaces under a PL condition**, without requiring additional assumptions commonly found in the literature, such as compactness or weak-to-strong continuity of f'. While such results are expected based on prior works on the Kurdyka–Łojasiewicz (KL) inequality (e.g., [7, 42]), these typically rely on non-vanishing step sizes, i.e., they exclude the case $\lambda_k \to 0$.

- **active manifold identification in finite time**, that is, in finite dimensions, all iterates of our algorithm eventually lie on a C^2-manifold where we can harness partial smoothness and active manifold theory to obtain *automatic satisfaction of a Dennis–Moré-type (DM-type) condition in an entirely nonsmooth setting* and deduce superlinear convergence.

- **the obtainment of the above results without assuming Lipschitz Hessians**, which is a standard assumption made in the literature on global rates for second-order and quasi-Newton methods (see the literature review in the next subsection).

- **adaptivity to parameters and function classes**, in the sense that our method is agnostic to knowledge about convexity, validity of the PL condition, semismoothness, etc., for the specific problem being tackled.

We close this subsection by noting that we implement and apply LeAP-SSN to total variation image restoration, support vector machine classification, and Signorini contact problems in Section 7. The implementation is publicly available [3].

1.2 Related literature

Related work on globalisation of SSN.   The globalisation of semismooth Newton methods is not a new endeavour. Numerous works have focused on globalising both classical and nonsmooth Newton-type methods, however, the majority of these approaches are restricted to the finite-dimensional setting and typically rely on linesearch procedures to ensure global convergence without proving rates. Though we are not aware of any existing approaches that establish global convergence rates for semismooth Newton methods in our setting, let us briefly review the works most closely related to ours.

(1) [84] (and the followup [83] with inexactness) proposes a similar regularised approach; however, the analysis there relies on the notion of second-order semismoothness (which our work does not need) and essentially requires F to be strongly convex to establish fast local convergence, and does not establish global convergence rates.

(2) [2] includes an inexact globalised semismooth Newton method to tackle fixed-point equations in Banach spaces with an emphasis on equations arising from the obstacle problem; the globalisation there is via the Banach fixed point theorem. [98] addresses a strongly convex optimal control problem and employs an Armijo-type line search strategy, and also uses second-order semismoothness for its local superlinear result. Armijo linesearch methods have the advantage of requiring only a single linear system solve per iteration, even during backtracking. Moreover, it can be shown that full Newton steps are eventually accepted, leading to superlinear convergence under semismoothness without requiring additional assumptions such as a Dennis–Moré condition. Note that neither work establishes global rates.

(3) The work in [80] (see also [39, 27]) was among the first to globalise a semismooth Newton-type method in a setting closely related to ours. They consider finite-dimensional problems

of the form $\min_{x \in X} f(x)$, where X is a polyhedron and ∇f is semismooth (also called an SC^1 function). Similar to our approach, their method employs a Levenberg–Marquardt–type regularization of the (Newton) Hessian of the form $\nabla_N^2 f(x_k) + \mu_k I$, involving the Newton derivative $\nabla_N^2 f(x_k)$. The choice of μ_k is not specified beyond the assumption $\mu_k \to 0$, and they establish superlinear convergence under suitable regularity conditions.

(4) [55] and the recent book [73] present results on minimizing $C^{1,1}$ functions f in finite dimensions, closely related to the present work. They use a Levenberg–Marquardt–type scheme which, instead of solving a regularised semismooth Newton system, tackles the inclusion $-\nabla f(x_k) \in (\partial^2 f(x_k) + \mu_k I)[d_k]$ to find the Newton direction d_k, where $\partial^2 f$ is the second-order subdifferential and $\mu_k = c\|\nabla f(x_k)\|$. This inclusion can sometimes be easier to solve than the semismooth Newton equation [55]. Combined with an Armijo-type linesearch, they prove superlinear convergence under semismoothness at the limit point, without Dennis–Moré–type conditions. A notable difference is that $\mu_k \to 0$ follows automatically from their algorithm, while our method needs extra smoothness to ensure this.

None of the above works obtain global convergence rates for their versions of SSN where applicable. Moreover, they either consider the finite-dimensional or (strongly) convex setting, or utilise stronger assumptions like second-order semismoothness.

Related work on global rates in the smooth, finite-dimensional setting. There is a vast literature establishing global convergence rates for regularised Newton methods for finite-dimensional convex and nonconvex optimisation problems, with the main assumption being that the Hessian is Lipschitz. For the nonconvex setting the classical results are [78, 20], where global rates are obtained for convex and nonconvex settings based on the cubic regularisation of the Newton method. The works [76, 72, 43, 19, 56] further improve the rates by proposing accelerated versions in the convex setting. The Lipschitz Hessian condition was relaxed to a Hölder condition in [46, 47], where universal regularised Newton methods (including accelerated versions) were proposed. The above methods mainly use a regularisation of power greater than 2 in auxiliary subproblems in each iteration, which makes this subproblem problematic to solve in practice. The most recent advances are based on the use of quadratic regularisation akin to Levenberg–Marquardt, but with the regularisation parameter proportional to some power of the gradient norm. In this avenue, global convergence rates are obtained in the convex case for the Levenberg–Marquardt method in [71, 32] under a Lipschitz Hessian assumption and in [31] under either Hölder Hessians or a third-order derivative assumption. All the above papers focus on global rates in finite dimensions and either in the convex setting with Hölder Hessians or in the nonconvex setting, but with Lipschitz Hessians. They also do not consider the behaviour of their methods under any form of semismoothness. Despite the fact that these approaches are not directly applicable in our setting of semismooth derivatives and infinite dimensions, our inspiration largely comes from the paper [31] which introduces a universal Newton method that adapts to the Hölder exponent of various smoothness classes automatically. Note that assuming Hölder or Lipschitz Hessians is out of the question in our setting because it would imply, by a simple argument, that f' is not only semismooth but differentiable.

Related work on Newton methods for non-strongly convex functions. Most works in the paragraph above rely on local strong convexity to obtain superlinear convergence, but recent research has focused on globalising Newton-type methods for nonconvex composite problems. In this setting, globalisation typically relies on Hessian regularisation, as in the present work. A central trend is replacing strong convexity with the *Kurdyka–Łojasiewicz (KL)* inequality, a generalisation of the PL condition

[82], which enables fast convergence even to non-isolated stationary points. Foundational results by Attouch et al. [6, 8, 7] established KL-based convergence for descent methods. The first Newton-type extension [42] achieved linear convergence. In the convex case, Mordukhovich et al. [75] recently obtained superlinear convergence under metric subregularity, closely related to PL-type inequalities. Inspired by this line of work, Liu et al. [69, 68] analyse a related composite setting, which is nonconvex, and establish superlinear convergence under the KL inequality and a local Hölder error bound, thereby allowing for non-isolated minima. Kanzow et al. [28] extends [69] to a linesearch-free variant with a similar analysis. To our knowledge, [99, 68] is the first to prove fast local convergence to *non-isolated* stationary points in the PL/KL framework. Importantly, *all of the above results are in finite dimensions* and *rely critically on the Lipschitz continuity of the Hessian*.

Related work on globalisation via trust regions. Trust-region methods are classical and highly effective for smooth (C^2) minimisation [25, 21]. In the setting of functions admitting a *semismooth Newton derivative*, the majority of results concern the solution of semismooth equations coming from the corresponding first order optimality system [96, 39], with recent advances providing guarantees under the KL condition [79]. For minimisation of functions with Newton differentiable derivatives the literature is still emerging with early work by Qi [87] and recent extensions to minimisation on manifolds by Zhang et al. [103]. Overall, trust-region schemes are robust to negative curvature, naturally accommodate inexact subproblem solves, and obtain global convergence and local superlinear convergence under different regularity assumptions, but are typically designed for finite-dimensional settings, demand stronger assumptions and more elaborate implementations than our regularised Newton-type globalisation. Moreover, the question of global rates remains open. Incorporating trust-region mechanisms in the spirit of [10] may further enhance our method; we leave this to future work.

Summary. To conclude, to the best of our knowledge, *no parameter-free globalisation strategy for the minimisation of functions with Newton differentiable derivatives on Hilbert spaces exists that achieves global sublinear rates together with a transition to superlinear convergence in the PL setting.* Our work closes this gap.

1.3 Organisation of the paper

The paper is organised as follows. In Section 2 we introduce basic notation and definitions, our main assumptions and preliminary results. In Section 3, we formulate our algorithm LeAP-SSN and detail some inequalities that come in use in the rest of the paper. We prove the above-mentioned global convergence results, local linear convergence, and superlinear convergence results in Section 4, Section 5, and Section 6, respectively. Finally, we conclude with some numerical experiments in Section 7.

2 Technical preliminaries

We start with some notation and necessary definitions. The dual to the space $\mathcal{H}$ is denoted by $\mathcal{H}^*$. We use $\|\cdot\|$ to denote the norm on $\mathcal{H}$ and $\|\cdot\|_*$ to denote the norm on $\mathcal{H}^*$. We write $\langle g, x \rangle$ to denote the duality pairing of $g \in \mathcal{H}^*$ and $x \in \mathcal{H}$. By $\mathcal{R} \colon \mathcal{H} \to \mathcal{H}^*$ we mean the Riesz map. For a point $x \in \mathcal{H}$ and a closed set $A \subseteq \mathcal{H}$, we denote $\mathrm{dist}(x, A) := \inf_{u \in A} \|x - u\|_{\mathcal{H}}$. Given $x \in \mathcal{H}$, we set $B_R(x) := \{u \in \mathcal{H} : \|x - u\| \leq R\}$ to be the closed ball of radius R around x. For two points $x, y \in \mathcal{H}$, we denote $[x, y] := \{tx + (1-t)y : t \in [0, 1]\}$. For a function $G \colon \mathcal{H} \to \overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$,

we denote its effective domain by $\operatorname{dom} G := \{x \in \mathcal{H} : G(x) < +\infty\}$. For the function F in (P) and $x_0 \in \operatorname{dom} F$, we denote the sublevel set of F as

$$\mathcal{F}_0 := \{x \in \operatorname{dom} F : F(x) \le F(x_0)\}. \tag{1}$$

For a proper convex function $\psi \colon \mathcal{H} \to \overline{\mathbb{R}}$, we denote by $\partial \psi$ its convex subdifferential, i.e., for any $x \in \operatorname{dom} \psi$,

$$\partial \psi(x) := \{h \in \mathcal{H}^* : \psi(x) - \psi(y) \le \langle h, x - y \rangle, \quad \forall y \in \mathcal{H}\}.$$

We also use $\psi'(x)$ to denote an element of $\partial \psi(x)$. For a function $F \colon \mathcal{H} \to \overline{\mathbb{R}}$, we denote by $\hat{\partial} F$ the Fréchet subdifferential

$$\hat{\partial} F(x) := \left\{ g \in \mathcal{H}^* : \liminf_{u \to x} \frac{F(u) - F(x) - \langle g, u - x \rangle}{\|u - x\|} \ge 0 \right\}.$$

The limiting (Mordukhovich) subdifferential $\partial_L F(x)$ at a point $x \in \operatorname{dom}(F)$ is defined as the set of elements $p \in \mathcal{H}^*$ for which there exist sequences $(x_k)_{k \in \mathbb{N}} \subset \mathcal{H}$ and $(p_k)_{k \in \mathbb{N}} \subset \mathcal{H}^*$ such that $p_k \in \hat{\partial} F(x_k)$ for all $k \in \mathbb{N}$ and the following limits hold:

$$x_k \to x, \qquad p_k \overset{*}{\rightharpoonup} p, \qquad F(x_k) \to F(x) \qquad \text{as } k \to \infty.$$

Note that if F is convex, we have $\partial_L F(x) = \hat{\partial} F(x) = \partial F(x)$ [57, Proposition 1.2].

For a function $f \colon \mathcal{H} \to \overline{\mathbb{R}}$ that is C^1 on $\operatorname{dom} f$, i.e., continuously Fréchet differentiable, we denote by $f'(x) \in \mathcal{H}^*$ its Fréchet derivative at $x \in \operatorname{dom} f$.

Remark 2.1 (On the setting of (P))**.** *If $F = f + \psi$ with f continuously Fréchet differentiable and ψ convex, the sum rule [74, Proposition 1.107 (ii)] yields, for all $x \in \operatorname{dom}(F)$,*

$$\partial_L F(x) = f'(x) + \partial_L \psi(x) = f'(x) + \hat{\partial} \psi(x) = \hat{\partial} F(x).$$

In other words, the limiting subdifferential and the Fréchet subdifferential coincide in this setting. Thus, throughout the paper, we write $\partial F(x)$ for the limiting subdifferential, which coincides with the Fréchet subdifferential in this setting. We also use $F'(x)$ to denote an element of $\partial F(x)$.

Note that, by Fermat's rule [74, Proposition 1.114], we have at a local minimum $\bar{x} \in \operatorname{dom} \psi$ that

$$0 \in \partial(f + \psi)(\bar{x}) = f'(\bar{x}) + \partial \psi(\bar{x}).$$

For a function $f \in C^{1,1}(\operatorname{dom} f, \overline{\mathbb{R}})$, i.e., $f \in C^1$ and has Lipschitz Fréchet derivative, we say that f' is *Newton differentiable* (or generalised differentiable) at a point $x \in \operatorname{dom} f$ if there exists an open subset $U \subset \operatorname{dom} f$ with $x \in U$ and a family of maps $H \colon U \to \mathcal{L}(\mathcal{H}, \mathcal{H}^*)$ such that

$$\lim_{h \in \mathcal{H}, h \to 0} \frac{\|f'(x + h) - f'(x) - H(x + h)h\|_*}{\|h\|} = 0. \tag{2}$$

(Note the difference to the Fréchet derivative, where $H(x + h)$ above would instead be $H(x)$.) We also say that f' is Newton differentiable at x with respect to H. If f' is Newton differentiable for every $x \in U$, we say that it is Newton differentiable on U. We slightly abuse notation and use $H(x)$ to denote the corresponding linear operator from $\mathcal{H}$ to $\mathcal{H}^*$ which we call the Newton or generalised derivative of f' at x. Finally, we adopt—by slight abuse of notation, but consistent with the convention in finite-dimensional optimisation—the shorthand

$$H(x) \succeq \mu I \quad \underset{\text{def}}{\Longleftrightarrow} \quad \langle (H(x) - \mu \mathcal{R})d, d \rangle \ge 0 \qquad \forall d \in \mathcal{H}, \tag{3}$$

where $\mu \in \mathbb{R}$ and, as before, $\mathcal{R} \colon \mathcal{H} \to \mathcal{H}^*$ denotes the Riesz map.

Remark 2.2. *A few words are in order regarding our definition of semismoothness.*

- *To be as general as possible, we do not specify which particular generalised derivatives $H(x)$ are used in the above definition. For example, choosing a single-valued selection of Clarke's Jacobian as generalised derivative $H(x)$, we obtain the classical definition of semismooth functions. Moreover, if f' is semismooth, every single-valued selection of Clarke's Jacobian works as $H(x)$ in the above definition. Following a tradition [49, 52], we still use the word "semismoothness"to denote that f' satisfies the above definition. Yet, we underline that our paper covers a much more general setting of Newton differentiability that extends the notion of differentiability even further than semismooth functions.*

- *In some communities, Newton differentiability is often defined in terms of set-valued derivatives. Indeed, a map $T\colon X \to Y$ is said to be Newton differentiable with Newton derivative $G\colon X \rightrightarrows \mathcal{L}(X, Y)$ if*

$$\lim_{h \to 0} \sup_{M \in G(x+h)} \frac{\|T(x + h) - T(x) - Mh\|}{\|h\|} = 0.$$

  The definition of Newton derivative we use in our work amounts to taking a selection from the set-valued G at each point. Thus, our Newton derivative can be viewed as an abuse of notation where we pick a realisation of one of the elements of the set G at each point. In this sense, our algorithm and results can handle set-valued derivatives.

We refer to [89] for a survey relating different notions of generalised derivatives.

2.1 Main assumptions

In this subsection, we present and discuss the main assumptions on the regularity of F in (P) that are used to obtain our theoretical results. We emphasise that we do not need all assumptions in all results: we will refer specifically to the assumptions that are needed in each result. However for convenience, we introduce the important assumptions here.

The following structural assumption about problem (P) is used for all the results of this paper and thus one should bear in mind that it is in effect throughout.

Assumption 1.   *(i) The function $\psi\colon \mathcal{H} \to \overline{\mathbb{R}}$ is convex, proper, and lower semicontinuous.*

  (ii) The function $f\colon \mathcal{H} \to \overline{\mathbb{R}}$ is continuously Fréchet differentiable on $\operatorname{dom} f$.

  (iii) The function $F := f + \psi$ is bounded from below by the global minimal value F^* and has a nonempty solution set $S := \{x^* \in \mathcal{H} : F(x^*) = F^*\}$.

Note that, in general, we do not assume f and F to be convex. The next assumption is fundamental and is used to prove global convergence rates of our algorithm.

Assumption 2. *There exists a constant $L \geq 0$ such that at any point $x \in \operatorname{dom} f$ it is possible to evaluate a symmetric operator $H(x) \in \mathcal{L}(\mathcal{H}, \mathcal{H}^*)$ such that f' satisfies*

$$\|f'(y) - f'(x) - H(x)(y - x)\|_* \leq L\|x - y\|, \quad \forall x, y \in \operatorname{dom} f. \tag{4}$$

We assumed symmetricity of H for simplicity, but the theory below would still work with relevant modifications (for example, $H(x)$ in (9) would need to be replaced with $\frac{1}{2}(H(x) + H(x)^*)$). Let us give some typical scenarios to illustrate the connection between H and f:

- **If f is C^2**: the smooth case when f is twice continuously differentiable offers the natural choice $H = f''$.

- **If f' is semismooth**: if f' is semismooth or Newton differentiable, we can use its Newton derivative as H.

- **The quasi-Newton setting**: here an approximation H of the Hessian (or Newton derivative) is used for implementation purposes. Typical assumptions in the literature include f'' being Lipschitz continuous and constructing H positive semi-definite.

We remark that an assumption similar to Assumption 2 is made in [101, §4] in the C^2 setting. Moreover, it can be seen as a nonasymptotic counterpart of 0th-order semismoothness [50, Definition 2.6]. To further motivate Assumption 2, note that the literature on the analysis of SSN and its variants often requires (see e.g. [84, §1], [98, Assumption 1], [100, Assumption 3.6], [2, Assumption 2.2]) that the operator $H(x)$ is uniformly bounded, i.e., for some $M > 0$,

$$\|H(x)\|_{\text{op}} := \sup_{h \in \mathcal{H}, h \neq 0} \frac{\|H(x)h\|_*}{\|h\|} \leq M. \tag{5}$$

As the next lemma shows, Assumption 2 follows from (5), and thus the former can be seen as a relaxation of the latter.

Lemma 2.3. *Let* (5) *hold.*

(i) *If f' is Lipschitz with constant L_1, then, Assumption 2 holds with constant $M + L_1$.*

(ii) *If f' is Newton differentiable with respect to H, then, Assumption 2 holds with constant $2M$.*

Proof. In case (ii), by [17, Proposition 2.3], it holds that f' is Lipschitz with constant $L_1 = M$. The rest of the proof follows by the triangle inequality:

$$\|f'(y) - f'(x) - H(x)(y - x)\|_* \leq \|f'(y) - f'(x)\|_* + \|H(x)(y - x)\|_* \leq (L_1 + M)\|x - y\|.$$

$\square$

Finally, we also remark that many semismooth functions appearing in applications have bounded Newton derivative and, hence, satisfy Assumption 2.

For some of the results, we use the following relaxation of strong convexity known as the Polyak–Łojasiewicz (PL) condition/inequality, which does not even require F to be convex. We use it in two contexts: sometimes globally for obtaining certain global convergence rates and locally for obtaining local convergence rates. The assumption for our global results is formulated as follows.

Assumption 3. *F satisfies the Polyak–Łojasiewicz (PL) condition on $\operatorname{dom} F$, i.e., for some $\mu > 0$,*

$$\frac{1}{2\mu} \operatorname{dist}(0, \partial F(x))^2 \geq F(x) - \min_{x \in \operatorname{dom} F} F(x), \quad \forall x \in \operatorname{dom} F. \tag{6}$$

Originally introduced in [82], the PL inequality has since been widely employed to analyse the linear convergence of first-order methods (see, e.g., [54]) and was later extended to nonsmooth and non-convex settings under the broader framework of the Kurdyka–Łojasiewicz (KL) inequality, see [6, 8, 7] and the discussion in [13] for a historical account. Another relaxation of strong convexity that does

not require F to be convex that we will use for one of our results on local convergence rates is the quadratic growth (QG) condition $\frac{\mu}{2} \operatorname{dist}(x, S)^2 \leq F(x) - \min_{x \in U} F(x)$.

Previous works on local and global convergence of SSN and its variants either use a strong convexity assumption [84, 98] or that the inverse of the Newton derivative is bounded from above [50, 96], which is a similar assumption to strong convexity since in that case the minimum eigenvalue of the Newton derivative is positive. For results on the relationship between strong convexity (SC), PL and QG conditions, we refer to [67] where in finite dimensions it is shown that SC implies PL which in turn implies QG for weakly convex functions and if additionally F is convex, PL and QG are equivalent but still weaker than SC; to [5], where the same is shown in infinite dimensions but with $\psi \equiv 0$; to [15] where the equivalence between PL and QG is established in infinite dimensions, but for convex functions; to [53] where for nonconvex functions and infinite dimensions it is shown that PL implies QG. Thus, our results are stronger than previously known since we use the weaker assumptions of PL and QG instead of SC as in other works.

Our next assumption is on the semismoothness of f' that is used exclusively to prove superlinear convergence of our algorithm.

Assumption 4. *The Fréchet derivative f' is semismooth (i.e., Newton differentiable) at a local solution x^* to problem* (P) *with respect to H (see* (2)*).*

We would like to underline that, despite Assumptions 2 and 3 being stated above as global assumptions, it is possible to prove some of our results by making these assumptions only locally on some ball $B_R(x^*)$ around a local solution x^*. These local counterparts of Assumptions 2 and 3 are simply obtained via changing $\operatorname{dom} f$ and $\operatorname{dom} F$ to $B_R(x^*)$.

2.1.1 Further results related to Assumption 2

Let us briefly discuss some basic results related to Assumption 2. We first see how in some cases, Assumption 2 and (5) are equivalent, cf. Lemma 2.3.

Lemma 2.4. *Let $\operatorname{dom} f = \mathcal{H}$, and f' be Lipschitz with constant L_1. If Assumption 2 holds, then,* (5) *holds with constant $L + L_1$. In fact, in this case, Assumption 2 and* (5) *are equivalent.*

Proof. Using Assumption 2 and the triangle inequality, we obtain $\|H(x)(y - x)\|_* \leq L\|x - y\| + \|f'(y) - f'(x)\|_*$. Taking $y = x + tz$ for arbitrary $z, x \in \mathcal{H}$ and $t > 0$, we further obtain

$$\|H(x)z\|_* \leq L\|z\| + \frac{\|f'(x + tz) - f'(x)\|_*}{t}.$$

Taking $t = 1$ and using the Lipschitz continuity of f', we directly obtain $\|H(x)z\|_* \leq (L + L_1)\|z\|$. Hence (5) is satisfied with $M = L + L_1$. The equivalence follows from Lemma 2.3, which gives the reverse direction and that Assumption 2 holds with constant $L = M + L_1$. $\qquad\square$

We will use the following basic lemma, which holds not only under Assumption 2, but also when the inequality (4) from the assumption holds only on line segments. Below, recall the notation $[x, y] := \{tx + (1 - t)y : t \in [0, 1]\}$.

Lemma 2.5. *Let $x, y \in \operatorname{dom} f$ and suppose that*

$$\|f'(\tilde{y}) - f'(\tilde{x}) - H(\tilde{x})(\tilde{y} - \tilde{x})\|_* \leq L\|\tilde{x} - \tilde{y}\|, \quad \forall \tilde{x}, \tilde{y} \in [x, y]. \tag{7}$$

Then, we have

$$\left| f(y) - [f(x) + \langle f'(x), y - x \rangle + \frac{1}{2} \langle H(x)(y-x), y-x \rangle] \right| \leq \frac{L}{2} \|x - y\|^2. \tag{8}$$

Proof. Indeed,

$$\left| f(y) - [f(x) + \langle f'(x), y - x \rangle + \frac{1}{2} \langle H(x)(y-x), y-x \rangle] \right|$$

$$= \left| \int_0^1 \langle f'(x + t(y-x)) - f'(x) - tH(x)(y-x), y-x \rangle \, \mathrm{d}t \right|$$

$$\leq \int_0^1 \|f'(x + t(y-x)) - f'(x) - tH(x)(y-x)\|_* \|y - x\| \, \mathrm{d}t$$

$$\overset{(7)}{\leq} \int_0^1 tL\|y - x\|^2 \, \mathrm{d}t = \frac{L}{2} \|x - y\|^2.$$

$\square$

2.2 Regularised Newton step

Our algorithm relies on the proximal Levenberg–Marquardt Newton step defined for some $x \in \mathrm{dom} F$ and constant $\lambda \geq 0$ as

$$x_+(\lambda, x) := \underset{y \in \mathcal{H}}{\arg\min} \, \widetilde{F}(y; x, \lambda),$$

$$\widetilde{F}(y; x, \lambda) := f(x) + \langle f'(x), y - x \rangle + \frac{1}{2} \langle H(x)(y-x), y-x \rangle + \frac{\lambda}{2} \|y - x\|^2 + \psi(y). \tag{P_N}$$

For brevity, if it is clear from the context, we use x_+ for $x_+(\lambda, x)$. Our assumption about the properties of a solution x_+ to the above auxiliary problem is as follows. This assumption is in effect throughout the rest of the paper.

Assumption 5. *For all $\lambda \geq L$ (where L is from Assumption 2) and all $x \in \mathrm{dom} F$ we can compute a stationary point $x_+ = x_+(\lambda; x) \in \mathcal{H}$ of the model function $\widetilde{F}(y; x, \lambda)$ in (P_N) that satisfies:*

$$-f'(x) - H(x)(x_+ - x) - \lambda \mathcal{R}(x_+ - x) \in \partial\psi(x_+) \qquad \text{(first-order condition)}, \tag{9}$$

$$\widetilde{F}(x_+; x, \lambda) \leq \widetilde{F}(x; x, \lambda) = F(x) \qquad \text{(model nonincrease)}. \tag{10}$$

Note that we do not require finding a global solution to (P_N). Also, it should be possible to relax the exact first-order condition (9) to allow for some inexactness, but we defer this to future research.

Remark 2.6. *Let us briefly discuss some cases in which Assumption 5 holds.*

■ *If f is convex and $H(x) \succeq 0$ for all $x \in \mathrm{dom} F$, then the minimisation problem in (P_N) is strongly convex for any $\lambda > 0$ and, hence, always admits a global unique minimiser x_+ that satisfies (9)–(10), independently of the constant L in Assumption 2. The global minimum can either be found explicitly if ψ is simple or by some auxiliary algorithm for convex optimisation.*

- *If f is μ–weakly convex, i.e., $f + (\mu/2)\|\cdot\|^2$ is convex, and $H(x) \succeq -\mu I$ for some $\mu \geq 0$ and all $x \in \mathrm{dom}\, F$, then Assumption 5 still holds provided the constant $L > 0$ in Assumption 2 satisfies $L \geq \mu$ since in this case the model $\widetilde{F}(\cdot; x, \lambda)$ is convex for all $x \in \mathrm{dom}\, F$ and $\lambda \geq L \geq \mu$. Thus, again $(\mathrm{P_N})$ admits a global minimiser x_+ that satisfies (9)–(10). The condition $\mu \leq L$ can always be satisfied since inequality (4) holds also for any $\tilde{L} \geq L$.*

- *If the term $\frac{1}{2}\langle H(x)(y - x),\, y - x\rangle$ in $(\mathrm{P_N})$ is nonconvex (and there is no such $\mu > 0$ that $H(x) \succeq -\mu I$), but the non-regularised model $\widetilde{F}(y; x, 0)$ is bounded from below for all $x \in \mathrm{dom}\, F$, then $\widetilde{F}(y; x, \lambda)$ is bounded from below for any $\lambda \geq 0$ and $x \in \mathrm{dom}\, F$. In this case, one can run the proximal gradient method (cf. [12]) starting from x. This method is monotone with respect to $\widetilde{F}$ and guarantees finding an (approximate) stationary point, so we may expect that both (9) and (10) also hold in this setting. The model $\widetilde{F}(y; x, 0)$ is bounded from below, for example, if $\mathrm{dom}\,\psi$ is compact or if ψ grows faster than quadratically as $\|x\| \to \infty$.*

- *Both conditions (9) and (10) are satisfied if $\widetilde{F}(y; x, \lambda)$ satisfies the PL condition.*

Note that this list is not exhaustive and there may be other situations where Assumption 5 is satisfied.

To reiterate, not only does Assumption 5 cover the convex setting often considered in other works, but also many nonconvex settings. We note also that Assumption 5 is actually needed only on the trajectory of the algorithm.

We next move to several technical results that characterise the properties of the step $(\mathrm{P_N})$. The first-order condition (9) means that there exists $\psi'(x_+) \in \partial\psi(x_+)$ such that

$$\psi'(x_+) = -f'(x) - H(x)(x_+ - x) - \lambda\mathcal{R}(x_+ - x). \tag{11}$$

Thus, defining

$$F'(x_+) := f'(x_+) + \psi'(x_+) = f'(x_+) - f'(x) - H(x)(x_+ - x) - \lambda\mathcal{R}(x_+ - x), \tag{12}$$

we immediately see that $F'(x_+) \in \partial F(x_+)$.

The following lemma gives us control over the length of the step via the subgradient norm at a point x. It is essentially [31, Lemma 1], but we give a proof for completeness.

Lemma 2.7. *Let $x_+ = x_+(x, \lambda)$ be a stationary point of $(\mathrm{P_N})$, where $H(x) \succeq \mu I$ for some $\mu \in \mathbb{R}$. Let also $\lambda + \mu > 0$. Then,*

$$\|x_+ - x\| \leq \frac{\|F'(x)\|_*}{\lambda + \mu}. \tag{13}$$

Proof. Rearranging the first-order optimality condition (11) and multiplying it by $x_+ - x$, we have, setting $r = \|x_+ - x\|$,

$$\langle H(x)(x_+ - x),\, x_+ - x\rangle = \langle -\psi'(x_+) - f'(x),\, x_+ - x\rangle - \lambda r^2.$$

Taking $\psi'(x) \in \partial\psi(x)$, we have by the assumption of the lemma and the monotonicity of $\partial\psi$,

$$\mu r^2 = \mu\|x_+ - x\|^2 \leq \langle H(x)(x_+ - x),\, x_+ - x\rangle \leq \langle -\psi'(x) - f'(x),\, x_+ - x\rangle - \lambda r^2$$
$$\leq \|\psi'(x) + f'(x)\|_*\, r - \lambda r^2.$$

Rearranging, dividing by r and $\lambda + \mu > 0$, we get (13). $\qquad\square$

In the next result, we establish a sensitivity estimate of how the point $x_+(\lambda, x)$ varies for fixed x when λ changes. This result will be fundamental for opening the path to show superlinear convergence.

Lemma 2.8. *Let $x \in \operatorname{dom} f$ be such that $H(x) \succeq 0$. Then, the following inequality holds for all $0 < \lambda \le \lambda'$:*

$$\|x_+(\lambda, x) - x_+(\lambda', x)\| \le \frac{\lambda' - \lambda}{\lambda'} \|x - x_+(\lambda, x)\|. \tag{14}$$

Proof. We denote for brevity $x_+(\lambda) := x_+(\lambda, x)$ and $\tilde{f}(y) := f(x) + \langle f'(x), y - x \rangle + \frac{1}{2} \langle H(x)(y - x), y - x \rangle$. Setting $y_1 = x_+(\lambda)$ and $y_2 = x_+(\lambda')$, using (11), we have that

$$\psi'(y_1) = -\tilde{f}'(y_1) - \lambda \mathcal{R}(y_1 - x), \quad \psi'(y_2) = -\tilde{f}'(y_2) - \lambda' \mathcal{R}(y_2 - x),$$

where $\psi'(y_1) \in \partial\psi(y_1)$ and $\psi'(y_2) \in \partial\psi(y_2)$. Multiplying the first equality by $y_1 - y_2$, the second by $y_2 - y_1$, and summing, we arrive at

$$\begin{aligned}
\langle \psi'(y_1) - \psi'(y_2), y_1 - y_2 \rangle &= \langle -\tilde{f}'(y_1) - \lambda \mathcal{R}(y_1 - x), y_1 - y_2 \rangle \\
&\quad + \langle \tilde{f}'(y_2) + \lambda' \mathcal{R}(y_2 - x), y_1 - y_2 \rangle \\
&= \langle \tilde{f}'(y_2) + \lambda' \mathcal{R}(y_2 - x) - \tilde{f}'(y_1) - \lambda \mathcal{R}(y_1 - x), y_1 - y_2 \rangle \\
&= \langle H(x)(y_2 - y_1) + \lambda' \mathcal{R}(y_2 - x) - \lambda \mathcal{R}(y_1 - x), y_1 - y_2 \rangle,
\end{aligned}$$

where in the last equality we used that $\tilde{f}'(y_2) - \tilde{f}'(y_1) = H(x)(y_2 - y_1)$. Rearranging, using $H(x) \succeq 0$ and the monotonicity of $\partial\psi(x)$, we obtain

$$\begin{aligned}
0 &\le \langle \psi'(y_1) - \psi'(y_2), y_1 - y_2 \rangle + \langle H(x)(y_1 - y_2), y_1 - y_2 \rangle \\
&= \langle \lambda' \mathcal{R}(y_2 - x) - \lambda \mathcal{R}(y_1 - x), y_1 - y_2 \rangle \\
&= \langle \lambda' \mathcal{R}(y_2 - x) - \lambda' \mathcal{R}(y_1 - x) + \lambda' \mathcal{R}(y_1 - x) - \lambda \mathcal{R}(y_1 - x), y_1 - y_2 \rangle \\
&= -\lambda' \|y_1 - y_2\|^2 + (\lambda' - \lambda) \langle \mathcal{R}(y_1 - x), y_1 - y_2 \rangle \\
&\le -\lambda' \|y_1 - y_2\|^2 + (\lambda' - \lambda) \|y_1 - x\| \, \|y_1 - y_2\|,
\end{aligned}$$

where for the last inequality we used Cauchy–Schwarz. Rearranging, we obtain (14) as desired. □

The final result of this subsection will allow us to establish per-iteration decrease of F by our algorithm.

Lemma 2.9. *Consider the update step* $(\mathrm{P_N})$ *and assume that there exists a function $\varphi : \mathbb{R} \to \mathbb{R}$ such that*

$$\|f'(x_+) - f'(x) - H(x)(x_+ - x)\|_* \le \varphi(\|x - x_+\|). \tag{15}$$

Then,

$$\langle F'(x_+), x - x_+ \rangle \ge \frac{1}{2\lambda} \|F'(x_+)\|_*^2 + \frac{\lambda}{2} \|x - x_+\|^2 - \frac{\varphi(\|x - x_+\|)^2}{2\lambda}. \tag{16}$$

Proof. We have that

$$\begin{aligned}
\varphi(\|x - x_+\|)^2 &\overset{(15)}{\ge} \|f'(x_+) - f'(x) - H(x)(x_+ - x)\|_*^2 \overset{(12)}{=} \|F'(x_+) + \lambda \mathcal{R}(x_+ - x)\|_*^2 \\
&= \|F'(x_+)\|_*^2 + 2\lambda \langle F'(x_+), x_+ - x \rangle + \lambda^2 \|x - x_+\|^2.
\end{aligned}$$

Dividing by 2λ and rearranging, we obtain the desired inequality (16). □

3 The proposed algorithm

In this section, we present LeAP-SSN — our Levenberg–Marquardt Adaptive Proximal Semismooth Newton method. The algorithm is based on the regularised proximal step $(\mathrm{P_N})$ and in each iteration we use a simple linesearch to adaptively find the regularisation parameter λ that guarantees sufficient progress of the algorithm. The resulting algorithm is listed as Algorithm 1 below.

Algorithm 1 LeAP-SSN (Levenberg–Marquardt Adaptive Proximal Semismooth Newton method)

1: **input:** Parameters $\alpha \in (0, 1/2]$, $\beta \in (0, (m-1)/(2m)]$ for some $m \geq 1$ (standard choice: $\alpha = 1/2$, $\beta = 1/4$, $m = 2$), $x_0 \in \operatorname{dom} F$, $\psi'(x_0) \in \partial\psi(x_0)$, and $\Lambda_0 > 0$.

2: **for** $k = 0, 1, \ldots$, **do**

3:     **for** $j_k = 0, 1, \ldots$, **do**

4:         Set $\lambda = 2^{j_k}\Lambda_k$.

5:         Attempt to compute

$$
x_+ = \arg\min_{y \in \mathcal{H}} \left\{ f(x_k) + \langle f'(x_k), y - x_k \rangle \right.
$$
$$
\left. + \frac{1}{2}\langle H(x_k)(y - x_k), y - x_k \rangle + \frac{\lambda}{2}\|y - x_k\|^2 + \psi(y) \right\} \tag{17}
$$

        in the sense that x_+ satisfies Assumption 5. If not possible, go to next iteration of loop (Line 3).

6:         Set $\psi'(x_+) = -f'(x_k) - H(x_k)(x_+ - x_k) - \lambda\mathcal{R}(x_+ - x_k)$.

7:         Set $F'(x_+) = f'(x_+) + \psi'(x_+)$.

8:         **if** x_+ satisfies the acceptance conditions

$$
\langle F'(x_+), x_k - x_+ \rangle \geq \frac{\alpha}{\lambda}\|F'(x_+)\|_*^2 \quad \text{and} \quad F(x_k) - F(x_+) \geq \beta\lambda\|x_+ - x_k\|^2 \tag{18}
$$

9:         Set $x_{k+1} = x_+$, $\lambda_k = \lambda = 2^{j_k}\Lambda_k$, $\Lambda_{k+1} = \lambda_k/2 = 2^{j_k}\Lambda_k/2$.

10:        Set $\psi'(x_{k+1}) = \psi'(x_+)$, $F'(x_{k+1}) = f'(x_{k+1}) + \psi'(x_{k+1})$.

11:        **break** and go to next iteration of the outer loop (Line 2).

12:        **end if**

13:     **end for**

14: **end for**

Remark 3.1 (Solving (17))**.** *Note that when* $\psi \equiv 0$, *(17) simplifies to computing*

$$
x_+ = x_k - (H(x_k) + \lambda\mathcal{R})^{-1} f'(x_k).
$$

In other words, potentially after a discretisation, one requires a single linear system solve. This x_+ *satisfies model nonincrease (10) under the various conditions discussed in Remark 2.6.*

When $\psi \neq 0$, x_+ *can be expressed via the proximal operator of* ψ *which in many cases is known explicitly [23]. Alternatively, one may need to run a proximal gradient descent or another auxiliary optimisation method to compute* x_+.

Remark 3.2 (The Riesz map and choice of norm)**.** *For many problems considered in the context of finite-dimensional optimisation (like the example in Section 7.3), the Riesz map* $\mathcal{R}$ *is the identity matrix* $\mathcal{R} = I$ *and* $\|\cdot\| = \|\cdot\|_* = \|\cdot\|_{\ell^2}$ *are the Euclidean norm. However, in discretisations of infinite-dimensional problems, the Riesz map often takes the form of a matrix* A *which is a combination of*

appropriate problem-dependent discretised operators. See Sections 7.1 and 7.2 for examples. In this case one defines $\|\cdot\| = \|\cdot\|_A$ and $\|\cdot\|_* = \|\cdot\|_{A^{-1}}$. This ensures the correct scaling of the solution coefficient vectors to their respective representations as functions and often vastly improves solver performance [93].

As we will show later, when proving superlinear convergence under semismoothness, we obtain $\lambda_k \to 0$, which means that our algorithm in a sense automatically interpolates between regularised proximal SSN and pure proximal SSN, and becomes closer to pure proximal SSN when $k \to \infty$.

Our next goal is to show that the inner cycle to find an appropriate λ_k is finite . To that end, recall that we use $x_+(\lambda, x_k)$ to denote a solution to $(\mathrm{P_N})$ with $x = x_k$ and that solution satisfies (9)–(10) of Assumption 5.

Lemma 3.3 (Acceptance and termination of the inner loop). *Consider iteration $k \geq 0$ of Algorithm 1 applied to* (P). *Let Assumptions 1 and 5 hold and suppose that the following inequality holds:*

$$\|f'(\tilde{y}) - f'(\tilde{x}) - H(\tilde{x})(\tilde{y} - \tilde{x})\|_* \leq L\|\tilde{x} - \tilde{y}\|, \quad \forall \tilde{x}, \tilde{y} \in [x_k, x_+(\lambda, x_k)], \quad \forall \lambda \geq 0. \quad (19)$$

Pick $m \geq 1$, $0 < \alpha \leq 1/2$, and $0 < \beta \leq (m-1)/(2m)$. Then, if $\lambda_k \geq mL$, both inequalities in (18) *hold, i.e., the acceptance conditions of the inner loop of Algorithm 1 holds and the inner loop ends after a finite number of trials. Furthermore, we have*

$$\lambda_k \leq \overline{\lambda} := \max\{2mL, \Lambda_0\}, \quad \Lambda_{k+1} \leq \overline{\lambda}/2, \quad \forall k \geq 0. \quad (20)$$

Before we start the proof, let us remark that (19) is of course implied by Assumption 2. We formulated the result with the localised version (19) because we will need it for our local convergence results in a later section.

Proof of Lemma 3.3. By Assumption 5 (which is applicable since we assumed $\lambda_k \geq mL \geq L$), the subproblem $(\mathrm{P_N})$ is well defined in the sense that it admits a stationary point satisfying (9)–(10). By the assumption (19) we have that in particular, the inequality (15) holds with $x_+(\lambda, x_k)$ and x_k instead of x_+ and x respectively, and $\varphi(r) = Lr$. Thus, applying Lemma 2.9, we obtain, for any $\lambda \geq mL \geq L$,

$$\langle F'(x_+(\lambda, x_k)), x_k - x_+(\lambda, x_k) \rangle \overset{(16)}{\geq} \frac{1}{2\lambda}\|F'(x_+(\lambda, x_k))\|_*^2$$
$$+ \frac{\lambda}{2}\|x_k - x_+(\lambda, x_k)\|^2 - \frac{L^2\|x_k - x_+(\lambda, x_k)\|^2}{2\lambda}$$
$$\geq \frac{1}{2\lambda}\|F'(x_+(\lambda, x_k))\|_*^2,$$

where in the last inequality we used that $\lambda \geq mL \geq L$. Thus, since $\lambda_k = 2^{j_k}\Lambda_k$ is increasing when j_k is increasing, we obtain that the first inequality in (18) holds after a finite number of trials when $\lambda_k \geq mL$ since $\alpha \leq 1/2$ by the assumptions of the lemma.

We now move to the second inequality in (18). By the assumption (19), we can apply Lemma 2.5 and

obtain, for any $\lambda \geq mL$,

$$
\begin{aligned}
F(x_+(\lambda, x_k)) &= f(x_+(\lambda, x_k)) + \psi(x_+(\lambda, x_k)) \\
&\overset{(8)}{\leq} f(x_k) + \langle f'(x_k), x_+(\lambda, x_k) - x_k \rangle \\
&\quad + \frac{1}{2} \langle H(x_k)(x_+(\lambda, x_k) - x_k), x_+(\lambda, x_k) - x_k \rangle \\
&\quad + \frac{L}{2} \|x_+(\lambda, x_k) - x_k\|^2 + \psi(x_+(\lambda, x_k)) \\
&\overset{(\mathrm{P_N})}{=} \widetilde{F}(x_+(\lambda, x_k); x_k, \lambda) - \frac{\lambda}{2} \|x_+(\lambda, x_k) - x_k\|^2 + \frac{L}{2} \|x_+(\lambda, x_k) - x_k\|^2 \\
&\overset{(10)}{\leq} F(x_k) + \frac{L - \lambda}{2} \|x_+(\lambda, x_k) - x_k\|^2 \\
&\overset{L \leq \lambda/m}{\leq} F(x_k) - \frac{m-1}{2m} \lambda \|x_+(\lambda, x_k) - x_k\|^2 .
\end{aligned}
$$

Thus, since $\lambda_k = 2^{j_k} \Lambda_k$ is increasing when j_k is increasing, by rearranging the last inequality, we obtain that the second inequality in (18) holds after a finite number of trials when $\lambda_k \geq mL$ since $\beta \leq \frac{m-1}{2m}$ by the assumptions of the lemma. Thus, both inequalities in (18) hold after a finite number of trials, and the step of Algorithm 1 is well defined.

We now move to the proof of the first inequality in (20) by induction. Consider the base case, i.e., iteration $k = 0$. If $j_0 = 0$, i.e., the acceptance condition holds for $\lambda_0 = \Lambda_0$, we have that $\lambda_0 \leq \max\{2mL, \Lambda_0\} = \overline{\lambda}$. If $j_0 > 0$, we in any case have that $\lambda_0 = \Lambda_0 2^{j_0} \leq 2mL$ since otherwise $\lambda_0/2 > mL$ would have already been accepted by the arguments in the first part of this lemma. Thus, again $\lambda_0 \leq \max\{2mL, \Lambda_0\} = \overline{\lambda}$. Consider now iteration $k > 0$ and proceed in the same way. If $j_k = 0$, i.e., the acceptance condition holds for $\lambda_k = \Lambda_k$, we have by the induction hypothesis that $\lambda_k = \Lambda_k = \lambda_{k-1}/2 \leq \overline{\lambda}/2 \leq \overline{\lambda}$. If $j_k > 0$, we in any case have that $\lambda_k \leq 2mL$ since otherwise the previous iterate already satisfies $\lambda_k/2 > mL$, which is sufficient for the acceptance conditions to hold by the first part of this lemma. Thus, again $\lambda_k \leq \overline{\lambda}$. The second inequality in (20) just follows by definition, since for any $k \in \mathbb{N}$, $\Lambda_{k+1} = \lambda_k/2 \leq \overline{\lambda}/2$. $\qquad \square$

3.1 Useful notation and inequalities

Before we move to the main results, we introduce some technical preliminaries which are used throughout the rest of the paper. To simplify the derivations, we introduce the following notation for $k \geq 0$:

$$
F_k := F(x_k) - F^*, \qquad r_k := \|x_k - x_{k+1}\|, \qquad g_k := \|F'(x_k)\|_* . \tag{21}
$$

Then, using the definition of the point x_{k+1} in Line 9 of Algorithm 1, the acceptance conditions (18) can be written as

$$
\langle F'(x_{k+1}), x_k - x_{k+1} \rangle \geq \frac{\alpha}{\lambda_k} g_{k+1}^2, \tag{22}
$$

$$
F_k - F_{k+1} = F(x_k) - F(x_{k+1}) \geq \beta \lambda_k r_k^2. \tag{23}
$$

We will use the following implications of the acceptance conditions (18).

Lemma 3.4. *Let, at iteration $k \geq 0$ of Algorithm 1, acceptance criteria* (18) *hold. Then, we have*

$$F^* \leq F(x_{k+1}) \leq F(x_k) \qquad \textit{or equivalently} \qquad 0 \leq F_{k+1} \leq F_k, \tag{24}$$

$$\|F'(x_{k+1})\|_* \leq \frac{\lambda_k}{\alpha}\|x_k - x_{k+1}\| \qquad \textit{or equivalently} \qquad g_{k+1} \leq \frac{\lambda_k r_k}{\alpha}, \tag{25}$$

$$F(x_k) - F(x_{k+1}) \geq \frac{\beta\alpha^2}{\lambda_k}\|F'(x_{k+1})\|_*^2 \qquad \textit{or equivalently}$$

$$F_k - F_{k+1} = F(x_k) - F(x_{k+1}) \geq \frac{\beta\alpha^2}{\lambda_k}g_{k+1}^2. \tag{26}$$

Proof. The inequalities in (24) follow from (23), (21), and Assumption 1. Applying Cauchy–Schwarz on the first inequality in (18) and using the definition of x_{k+1} in Line 9 of Algorithm 1 gives

$$\frac{\alpha}{\lambda_k}\|F'(x_{k+1})\|_* \leq \|x_k - x_{k+1}\|,$$

which together with (21) clearly implies (25). Further, after plugging the previous inequality into the second inequality in (18) and using again the definition of x_{k+1} in Line 9 of Algorithm 1, we get

$$F(x_k) - F(x_{k+1}) \geq \beta\lambda_k\|x_k - x_{k+1}\|^2 \geq \frac{\beta\alpha^2}{\lambda_k}\|F'(x_{k+1})\|_*^2,$$

and (26) follows by (21). □

Lemma 3.5. *Given a set $U \subseteq \mathcal{H}$, suppose there exists $F_U^* \in \mathbb{R}$ such that $F(x) \geq F_U^*$ for all $x \in U$.*

(i) *At iteration $k \geq 0$ of Algorithm 1, let the PL inequality hold at $x_k \in U$, i.e.,*

$$\frac{1}{2\mu}\mathrm{dist}(0, \partial F(x_k))^2 \geq F(x_k) - F_U^*. \tag{27}$$

Then, we have

$$g_k^2 := \|F'(x_k)\|_*^2 \geq 2\mu(F(x_k) - F_U^*). \tag{28}$$

(ii) *At iteration $k + 1 \geq 1$ of Algorithm 1, let the assumption of the previous item hold and the acceptance criteria* (18) *hold. Let in addition $x_{k+2} \in U$. Then, we have*

$$r_{k+1}^2 \leq \frac{\lambda_k r_k^2}{\alpha^2 \beta \mu}, \tag{29}$$

$$g_{k+2}^2 \leq \frac{\lambda_{k+1} g_{k+1}^2}{2\alpha^2 \beta \mu}. \tag{30}$$

Proof. Since by construction $F'(x_k) \in \partial F(x_k)$, we have that

$$g_k^2 \stackrel{(21)}{:=} \|F'(x_k)\|_*^2 \geq \mathrm{dist}(0, \partial F(x_k))^2 \stackrel{(27)}{\geq} 2\mu(F(x_k) - F_U^*),$$

which is (28). By (23) at iteration $k+1$, inequality $F(x_{k+2}) - F_U^* \geq 0$ (which follows from $x_{k+2} \in U$), inequality (28) at iteration $k + 1$, and inequality (25), we observe that

$$\beta\lambda_{k+1}r_{k+1}^2 \stackrel{(23)}{\leq} (F(x_{k+1}) - F_U^*) - (F(x_{k+2}) - F_U^*) \leq F(x_{k+1}) - F_U^* \stackrel{(28)}{\leq} \frac{g_{k+1}^2}{2\mu} \stackrel{(25)}{\leq} \frac{\lambda_k^2 r_k^2}{2\alpha^2\mu}. \tag{31}$$

Using that $\lambda_k/2 = \Lambda_{k+1} \leq \lambda_{k+1}$ by Line 9 of Algorithm 1, we obtain (29). To obtain (30), we use (25) at iteration $k + 1$ as well as the first and the penultimate inequality in (31):

$$g_{k+2}^2 \stackrel{(25)}{\leq} \frac{\lambda_{k+1}^2}{\alpha^2}r_{k+1}^2 \stackrel{(31)}{\leq} \frac{\lambda_{k+1}^2}{\alpha^2}\frac{g_{k+1}^2}{2\beta\mu\lambda_{k+1}} = \frac{\lambda_{k+1}g_{k+1}^2}{2\alpha^2\beta\mu}.$$

□

4 Global convergence rates

In this section, we obtain our main results on global nonasymptotic convergence rate results for Algorithm 1 in various settings including under a PL condition and a convex setting. The main tool for these results is Assumption 2.

4.1 Global $\mathcal{O}(1/\sqrt{k})$ convergence rate for gradient norm

In this subsection, we focus on the setting where F in (P) is not necessarily convex (but Assumption 5 is satisfied). We show that under the global Assumption 2, Algorithm 1 has a global nonasymptotic convergence rate $\mathcal{O}(1/\sqrt{k})$ in terms of the minimal subgradient norm on the trajectory. The formal result is as follows.

Theorem 4.1 (Global convergence rate). *Let, for problem* (P)*, Assumptions 1, 2 and 5 hold. Let also* $(x_k)_{k\in\mathbb{N}}$ *be the iterates of Algorithm 1 with arbitrary starting point* $x_0 \in \operatorname{dom} F$*. Then, the sequence* $(x_k)_{k\in\mathbb{N}}$ *is well defined,* $(x_k)_{k\in\mathbb{N}} \subset \mathcal{F}_0$*, i.e., the iterates of the algorithm stay in the sublevel set* $\mathcal{F}_0$ *(see* (1)*), and*

$$\|F'(x_k)\|_* \to 0 \quad \text{as } k \to \infty.$$

Moreover, the following global nonasymptotic convergence rate holds:

$$\min_{0\le i\le k-1}\|F'(x_{i+1})\|_* \le \frac{\sqrt{\max\{2mL, \Lambda_0\}(F(x_0) - F^*)}}{\sqrt{\beta\alpha^2 k}}, \quad k \ge 1. \tag{32}$$

If in addition $\lambda_k \to 0$ *as* $k \to \infty$*, we obtain that* $\min_{0\le i\le k-1}\|F'(x_{i+1})\|_* = o(1/\sqrt{k})$*. Finally, the number of Newton steps* (17) *up to the end of iteration* k *does not exceed*

$$k + 1 + \log_2 \max\{1/2, mL/\Lambda_0\}.$$

Proof. Since Assumption 2 holds, we have by Lemma 3.3 that both inequalities in (18) hold in each iteration $k \ge 0$, i.e., the acceptance conditions of the inner loop of Algorithm 1 hold and the inner loop ends after a finite number of trials. Thus, the sequence $(x_k)_{k\in\mathbb{N}}$ is well defined. From (24), it is clear by induction that for all $k \ge 0$, $F(x_k) \le F(x_0)$ and, thus, $(x_k)_{k\in\mathbb{N}} \subset \mathcal{F}_0$, where $\mathcal{F}_0$ is defined in (1).

Summing (26) from 0 to $k - 1$ and telescoping, using that F is bounded from below by F^* by Assumption 1, Lemma 3.3 and the notation $g_{i+1} \coloneqq \|F'(x_{i+1})\|_*$ in (21), we obtain

$$F(x_0) - F^* \ge F(x_0) - F(x_k) \overset{(26)}{\ge} \sum_{i=0}^{k-1} \frac{\beta\alpha^2}{\lambda_i} g_{i+1}^2 \overset{(20)}{\ge} \sum_{i=0}^{k-1} \frac{\beta\alpha^2}{\overline{\lambda}} g_{i+1}^2 \ge \frac{k\beta\alpha^2}{\overline{\lambda}} \min_{0\le i\le k-1}\|F'(x_{i+1})\|_*^2.$$

First, observe that this chain of inequalities implies that for all $k \ge 0$, $\sum_{i=0}^{k-1} g_{i+1}^2$ is bounded from above. Thus, $\|F'(x_i)\|_* = g_i \to 0$ as $i \to \infty$. Second, recalling from (20) that $\overline{\lambda} = \max\{2mL, \Lambda_0\}$, we immediately get (32). Now, suppose that $\lambda_i \to 0$ as $i \to \infty$. We also obtain from the above inequalities, by keeping λ_i, that

$$F(x_0) - F^* \ge \beta\alpha^2 \min_{0\le i\le k-1} g_{i+1}^2 \sum_{i=0}^{k-1} \frac{1}{\lambda_i}$$

and hence

$$\min_{0 \le i \le k-1} g_{i+1}^2 \le \frac{F(x_0) - F^*}{\beta \alpha^2} \frac{1}{\sum_{i=0}^{k-1} \frac{1}{\lambda_i}}.$$

The inequality between arithmetic and geometric means (AM-GM) yields

$$\sum_{i=0}^{k-1} \frac{1}{\lambda_i} \ge k \left(\prod_{i=0}^{k-1} \frac{1}{\lambda_i} \right)^{1/k},$$

whence

$$\min_{0 \le i \le k-1} g_{i+1}^2 \le \frac{F(x_0) - F^*}{\beta \alpha^2} \frac{1}{k \left(\prod_{i=0}^{k-1} \frac{1}{\lambda_i} \right)^{1/k}} = \frac{F(x_0) - F^*}{\beta \alpha^2} \frac{\left(\prod_{i=0}^{k-1} \lambda_i \right)^{1/k}}{k},$$

and the right-hand side is $\mathscr{o}(1/k)$ because the geometric mean converges to zero. By taking the square root, we deduce the $\mathscr{o}(1/\sqrt{k})$ rate for $\min_{0 \le i \le k-1} \|F'(x_{i+1})\|_*$.

It remains to estimate the number of Newton steps (17). By Line 9 of Algorithm 1, we have for all $k \ge 0$ that $\Lambda_{k+1} = 2^{j_k} \Lambda_k / 2$, or equivalently $j_k = 1 + \log_2 \frac{\Lambda_{k+1}}{\Lambda_k}$. Summing these equalities from 0 to k, we obtain that the total number of Newton steps up to the end of iteration k is

$$N_k = \sum_{i=0}^k j_i = \sum_{i=0}^k \left(1 + \log_2 \frac{\Lambda_{i+1}}{\Lambda_i} \right) = k + 1 + \log_2 \frac{\Lambda_{k+1}}{\Lambda_0}$$

$$\overset{(20)}{\le} k + 1 + \log_2 \frac{\overline{\lambda}}{2\Lambda_0} \overset{(20)}{=} k + 1 + \log_2 \max\{1/2, mL/\Lambda_0\}.$$

$\square$

Remark 4.2 (Choice of the parameters m, α, β). *As one can see, the rate in* (32) *is proportional to* $\sqrt{m/(\beta \alpha^2)}$ *and the smaller this number, the better the rate. This means that* α, β *should be chosen as large as possible. Recall from Lemma 3.3 that* $0 < \alpha \le 1/2$, *and* $0 < \beta \le (m-1)/(2m)$. *Thus, the optimal choice is* $\alpha = 1/2, \beta = (m-1)/(2m)$. *This leads to* $\sqrt{m/(\beta \alpha^2)} = \sqrt{8m^2/(m-1)}$, *which is minimised at* $m^* = 2$. *This, in turn, implies* $\beta = 1/4$. *This is the standard choice outlined in the input of Algorithm 1.*

We finish this subsection with a result establishing the stationarity of strong accumulation points of the sequence $(x_k)_{k \in \mathbb{N}}$ of the iterates of Algorithm 1. By Theorem 4.1, we have $(x_k)_{k \in \mathbb{N}} \subset \mathcal{F}_0$. If the sublevel set $\mathcal{F}_0$ is bounded and the space $\mathcal{H}$ is finite-dimensional, then there exists a strong accumulation point of $(x_k)_{k \in \mathbb{N}}$. Unfortunately, if $\mathcal{H}$ is infinite-dimensional, only weak accumulation points are guaranteed to exist. Following existing literature on nonconvex minimisation in Hilbert spaces [44, 84], we restrict ourselves here to strong accumulation points. The study of additional assumptions on f, e.g., strong compactness of sublevel sets, that allow establishing stationarity of weak accumulation points is left for future work.

Proposition 4.3 (Stationarity of strong accumulation points). *Let, for problem* (P), *Assumptions 1, 2 and 5 hold. Then, all strong accumulation points of the sequence* $(x_k)_{k \in \mathbb{N}}$ *of the iterates of Algorithm 1 are stationary points of problem* (P).

Proof. Pick an accumulation point $\bar{x} \in \mathcal{H}$ and a subsequence such that $x_{k_l} \to \bar{x}$ as $l \to \infty$. From Theorem 4.1, we know that $F'(x_k) \to 0$ as $k \to \infty$. Thus, $F'(x_{k_l}) = f'(x_{k_l}) + \psi'(x_{k_l}) \to 0$ as $l \to \infty$. Since f is continuously Fréchet differentiable, we have that $f'(x_{k_l}) \to f'(\bar{x})$. Consequently $\psi'(x_{k_l}) = (\psi'(x_{k_l}) + f'(x_{k_l})) - f'(x_{k_l}) \to -f'(\bar{x})$. Since ψ is convex lsc, $x \mapsto \partial\psi(x)$ has a closed graph and we deduce that $-f'(\bar{x}) = \lim_{l\to\infty} \psi'(x_{k_l}) \in \partial\psi(\bar{x})$. Thus, equivalently, we have that $0 \in \partial F(\bar{x}) = f'(\bar{x}) + \partial\psi(\bar{x})$ and so $\bar{x}$ is a stationary point.   □

4.2   Global linear convergence rate for objective and iterates under PL

In this subsection, we still assume that F in (P) may be nonconvex and Assumptions 2 and 5 hold. Yet, we additionally make Assumption 3 that F satisfies the PL condition. This allows us to obtain a global nonasymptotic *linear* convergence rate for Algorithm 1 for the objective functional values and for the iterates. The formal result is as follows.

Theorem 4.4 (Global convergence rate, PL condition)**.** *Let, for problem* (P)*, Assumptions 1 to 3 and 5 hold. Let also $(x_k)_{k\in\mathbb{N}}$ be the iterates of Algorithm 1 with arbitrary starting point $x_0 \in \operatorname{dom} F$. Then, the following statements hold.*

   (i) *The objective values have the following global nonasymptotic linear convergence rate:*

$$F(x_k) - F^* \leq \exp\left(-\frac{2\beta\alpha^2\mu}{2\beta\alpha^2\mu + \max\{2mL, \Lambda_0\}} \cdot k\right) \cdot (F(x_0) - F^*), \quad k \geq 0. \quad (33)$$

   If in addition $\lambda_k \to 0$ as $k \to \infty$, then $F(x_k) - F^*$ converges to 0 superlinearly.

  (ii) *x_k converges strongly to a global minimum x^* and the convergence is global, nonasymptotic, and linear:*

$$\|x_k - x^*\| \leq \exp\left(-\frac{2\beta\alpha^2\mu}{4\beta\alpha^2\mu + 2\max\{2mL, \Lambda_0\}} \cdot (k-2)\right) \cdot \frac{5\sqrt{F(x_0) - F^*}}{\alpha\beta\mu^{1/2}}, \quad k \geq 2. \tag{34}$$

   If in addition $\lambda_k \to 0$ as $k \to \infty$, then x_k converges to x^* superlinearly.

 (iii) *The number of Newton steps* (17) *up to the end of iteration k does not exceed*

$$k + 1 + \log_2 \max\{1/2, mL/\Lambda_0\}.$$

Proof. Recall from (21) the notation for $k \geq 0$

$$F_k := F(x_k) - F^*, \qquad r_k := \|x_k - x_{k+1}\|, \qquad g_k := \|F'(x_k)\|_*.$$

(i) From Theorem 4.1, we know that the sequence $(x_k)_{k\in\mathbb{N}} \subset \mathcal{F}_0$ is well defined. We can apply the PL condition Assumption 3 (and use that F^* is the global minimal value by Assumption 1) to get

$$F_k - F_{k+1} \overset{(26)}{\geq} \frac{\beta\alpha^2}{\lambda_k} g_{k+1}^2 \overset{(6)}{\geq} \frac{\beta\alpha^2}{\lambda_k} \cdot 2\mu F_{k+1}.$$

Rearranging, we obtain

$$F_{k+1} \leq \frac{1}{1 + \frac{2\beta\alpha^2\mu}{\lambda_k}} F_k = \frac{\lambda_k}{2\beta\alpha^2\mu + \lambda_k} F_k = \left(1 - \frac{2\beta\alpha^2\mu}{2\beta\alpha^2\mu + \lambda_k}\right) F_k. \tag{35}$$

We immediately see that if $\lambda_k \to 0$ as $k \to \infty$, then $1 - \frac{2\beta\alpha^2\mu}{2\beta\alpha^2\mu+\lambda_k} \to 0$ and we obtain superlinear convergence. Using the upper bound $\lambda_k \leq \overline{\lambda}$ from Lemma 3.3 (see (20)), we obtain

$$\left(1 - \frac{2\beta\alpha^2\mu}{2\beta\alpha^2\mu+\lambda_k}\right) \leq \left(1 - \frac{2\beta\alpha^2\mu}{2\beta\alpha^2\mu+\overline{\lambda}}\right) \leq \exp\left(-\frac{2\beta\alpha^2\mu}{2\beta\alpha^2\mu+\overline{\lambda}}\right).$$

This, by induction gives, for any $k \geq 0$,

$$F(x_k) - F^* = F_k \leq \exp\left(-\frac{2\beta\alpha^2\mu}{2\beta\alpha^2\mu+\overline{\lambda}} \cdot k\right)(F(x_0) - F^*), \tag{36}$$

which is (33) since $\overline{\lambda} = \max\{2mL, \Lambda_0\}$ from (20).

(ii) By Lemma 3.5, (23), and (36), we observe, for all $k \geq 0$,

$$r_{k+1}^2 \overset{(29)}{\leq} \frac{\lambda_k r_k^2}{\alpha^2\beta\mu} \overset{(23)}{\leq} \frac{1}{\alpha^2\beta^2\mu}(F_k - F_{k+1}) \overset{(36)}{\to} 0 \quad \text{as } k \to \infty. \tag{37}$$

Thus, we have shown that $r_k \to 0$ as $k \to \infty$.

Using the concavity of $t \mapsto t^{1/2}$, the fact that $F_k \geq F_{k+1} \geq 0$ by (24), and inequality (23), we get, for all $k \geq 0$,

$$F_k^{1/2} - F_{k+1}^{1/2} \geq \frac{1}{2F_k^{1/2}}(F_k - F_{k+1}) \overset{(23)}{\geq} \frac{\beta\lambda_k r_k^2}{2F_k^{1/2}}. \tag{38}$$

Applying again the PL condition, we get $g_k^2 \overset{(6)}{\geq} 2\mu F_k$. Further, by (25) at iteration $k-1$ we have $g_k \leq \frac{\lambda_{k-1}}{\alpha}r_{k-1}$. Combining the last two observations with (38), we obtain

$$F_k^{1/2} - F_{k+1}^{1/2} \overset{(6)}{\geq} \frac{\beta\mu^{1/2}\lambda_k r_k^2}{\sqrt{2}g_k} \overset{(25)}{\geq} \frac{\alpha\beta\mu^{1/2}\lambda_k r_k^2}{\sqrt{2}\lambda_{k-1}r_{k-1}}.$$

By Line 9 of Algorithm 1 we have $\lambda_k \geq \Lambda_k = \lambda_{k-1}/2$ and consequently

$$F_k^{1/2} - F_{k+1}^{1/2} \geq \frac{\alpha\beta\mu^{1/2}r_k^2}{2\sqrt{2}r_{k-1}}. \tag{39}$$

Rearranging (39) and using inequality $2\sqrt{ab} \leq a + b$ that holds for $a, b \geq 0$, we get

$$r_k \leq \sqrt{\frac{2\sqrt{2}r_{k-1}(F_k^{1/2} - F_{k+1}^{1/2})}{\alpha\beta\mu^{1/2}}} \leq \frac{r_{k-1}}{2} + \frac{\sqrt{2}(F_k^{1/2} - F_{k+1}^{1/2})}{\alpha\beta\mu^{1/2}}.$$

Summing from $l + 1 \geq 1$ to $n \geq l + 1$ we obtain

$$\sum_{k=l+1}^{n} r_k \leq \frac{1}{2}\sum_{k=l}^{n-1} r_k + \frac{\sqrt{2}(F_{l+1}^{1/2} - F_{n+1}^{1/2})}{\alpha\beta\mu^{1/2}}.$$

Putting the sums on one side and using that $r_k \geq 0$, we obtain

$$\frac{1}{2}\sum_{k=l+1}^{n-1} r_k \leq r_l + \frac{\sqrt{2}(F_{l+1}^{1/2} - F_{n+1}^{1/2})}{\alpha\beta\mu^{1/2}} \to 0 \quad \text{as } l, n \to \infty, \tag{40}$$

where we used that $r_k \to 0$ by (37) and $F_k \to 0$ by (36). Hence $(x_k)_{k \in \mathbb{N}}$ is a Cauchy sequence and strongly converges to a point $x^* \in \mathcal{H}$, which by Proposition 4.3, is a stationary point. Further, the functional F is a sum of continuously differentiable f and convex proper lower semicontinuous ψ. Thus, F is lower semicontinuous, which implies

$$F^* \overset{(36)}{=} \lim_{k \to \infty} F(x_k) = \liminf_{k \to \infty} F(x_k) \geq F(x^*).$$

Therefore, x^* is a global minimiser itself.

The linear convergence of $(x_k)_{k \in \mathbb{N}}$ is shown as follows. Fixing some $l \geq 0$ and arbitrary $n \geq l + 2$, we obtain using the bound in (40) and the triangle inequality, that

$$\|x_{l+1} - x_n\| \leq \sum_{k=l+1}^{n-1} r_k \leq 2r_l + \frac{2\sqrt{2}(F_{l+1}^{1/2} - F_{n+1}^{1/2})}{\alpha\beta\mu^{1/2}} \tag{41}$$

$$\overset{(37)}{\leq} 2\sqrt{\frac{F_{l-1} - F_l}{\alpha^2\beta^2\mu}} + \frac{2\sqrt{2}(F_{l+1}^{1/2} - F_{n+1}^{1/2})}{\alpha\beta\mu^{1/2}}. \tag{42}$$

Sending $n \to \infty$ in (42), we obtain

$$\|x_{l+1} - x^*\| \overset{(42)}{\leq} 2\sqrt{\frac{F_{l-1} - F_l}{\alpha^2\beta^2\mu}} + \frac{2\sqrt{2}F_{l+1}^{1/2}}{\alpha\beta\mu^{1/2}} \leq \left(2\sqrt{\frac{1}{\alpha^2\beta^2\mu}} + \frac{2\sqrt{2}}{\alpha\beta\mu^{1/2}}\right) F_{l-1}^{1/2} \leq \frac{5F_{l-1}^{1/2}}{\alpha\beta\mu^{1/2}},$$

where the penultimate inequality follows as $F_{l-1} - F_l \leq F_{l-1}$ and $0 \leq F_{l+1} \leq F_l \leq F_{l-1}$ by (24). Thus, the linear convergence of x_l (34) follows since F_l converges to 0 linearly, see (36). In the same way, superlinear convergence when $\lambda_l \to 0$ as $l \to \infty$ follows since F_l converges to 0 superlinearly in that case, see (35).

(iii) Finally, the bound on the number of Newton steps follows from Theorem 4.1. □

As we see from the above result, the additional assumption of the PL condition Assumption 3 allows us to obtain quite strong results, including global strong nonasymptotic linear convergence of the iterates to a global minimum and global nonasymptotic linear convergence of objective values. Observe also that it would have sufficed to assume Assumption 3 only on $\mathcal{F}_0$ (instead of $\mathrm{dom}\, F$) since by Theorem 4.1 we know that $(x_k)_{k \in \mathbb{N}} \subset \mathcal{F}_0$, but we formulated it as it is to simplify the presentation. We also remark that we did not assume that F is convex.

We finish this subsection by a technical corollary of Theorem 4.4 that will be crucial later for the proof of superlinear convergence of Algorithm 1. This result relates the distance to solution $\|x_k - x^*\|$ to the distance between two consecutive iterates $\|x_{k+1} - x_k\|$, which in general turns out to be important for proving fast convergence of various algorithms, see, e.g., [35, 34]. A similar result was proved in the recent work [84], but under strong convexity. In our setting, we considerably relax this assumption by assuming that the PL condition holds instead of strong convexity.

Lemma 4.5. *Let the assumptions of Theorem 4.4 hold. Then, there exist constants $c_1, c_2 > 0$ such that, for all $k \geq 0$,*

$$\|x_{k+1} - x^*\| \leq c_1 \|x_{k+1} - x_k\| \quad and \quad \|x_k - x^*\| \leq c_2 \|x_{k+1} - x_k\|. \tag{43}$$

Proof. Since the assumptions of Theorem 4.4 hold, we know that $x_k \to x^*$ as $k \to \infty$ and we can use (41) which holds for any $l \geq 0$ and $n \geq l + 2$. Whence, changing in (41) l to k and sending

$n \to \infty$, we have

$$\|x_{k+1} - x^*\| \leq 2r_k + \frac{2\sqrt{2}F_{k+1}^{1/2}}{\alpha\beta\mu^{1/2}}. \tag{44}$$

By Assumption 3 and the bound $\lambda_k \leq \overline{\lambda}$ from Lemma 3.3, we obtain

$$F_{k+1}^{1/2} \overset{(21)}{=} (F(x_{k+1}) - F^*)^{1/2} \overset{(6)}{\leq} \frac{g_{k+1}}{(2\mu)^{1/2}} \overset{(25)}{\leq} \frac{\lambda_k r_k}{\alpha(2\mu)^{1/2}} \leq \frac{\overline{\lambda}}{\alpha(2\mu)^{1/2}} r_k.$$

Combining this bound and (44) and recalling that, by (21), $r_k = \|x_k - x_{k+1}\|$, we have that the first inequality in (43) holds with $c_1 = 2 + \frac{2\overline{\lambda}}{\alpha^2\beta\mu}$. The triangle inequality then yields the second inequality in (43) with $c_2 = 1 + c_1$. $\qquad\square$

4.3 Global $\mathcal{O}(1/k)$ convergence rate for objective under convexity

In this subsection, we focus on the setting where F in (P) is convex on its sublevel set $\mathcal{F}_0$ defined in (1). This assumption allows us to improve some results of the previous subsections. In particular, we obtain for Algorithm 1 a global nonasymptotic convergence rate $\mathcal{O}(1/k)$ for the objective residual $F(x_k) - F^*$. Convexity also allows us to show that all *weak* accumulation points of the sequence generated by Algorithm 1 are global minima.

First, we can improve the bound (26) in Lemma 3.4 by using convexity.

Lemma 4.6. *Let at iteration $k \geq 0$ of Algorithm 1 the acceptance condition* (18) *hold. Let also F be convex between x_k and x_{k+1}, i.e., $F(x_k) \geq F(x_{k+1}) + \langle F'(x_{k+1}), x_k - x_{k+1}\rangle$. Then, we have that*

$$F(x_k) - F(x_{k+1}) \geq \frac{\alpha}{\lambda_k}\|F'(x_{k+1})\|_*^2 \quad \textit{or equivalently}$$

$$F_k - F_{k+1} = F(x_k) - F(x_{k+1}) \geq \frac{\alpha}{\lambda_k} g_{k+1}^2. \tag{45}$$

Proof. Using convexity, the definition of x_{k+1} in Line 9 of Algorithm 1, and the first inequality in (18) we obtain $F(x_k) - F(x_{k+1}) \geq \langle F'(x_{k+1}), x_k - x_{k+1}\rangle \overset{(18)}{\geq} \frac{\alpha}{\lambda_k}\|F'(x_{k+1})\|_*^2$. The rest follows from the notation (21). $\qquad\square$

Recall the definition (1) of the sublevel set $\mathcal{F}_0$ defined by the starting point x_0. We assume it to be bounded, i.e.,

$$D_0 := \sup_{x,y \in \mathcal{F}_0} \|x - y\| < +\infty. \tag{46}$$

By Theorem 4.1, we have that $(x_k)_{k\in\mathbb{N}} \subset \mathcal{F}_0$ is well defined. This, together with convexity of F on $\mathcal{F}_0$ and the notation $g_k := \|F'(x_k)\|_*$ in (21) imply that, for any $\bar{x} \in \mathcal{F}_0$ and $k \geq 0$, we have

$$F(x_k) - F(\bar{x}) \leq \langle F'(x_k), x_k - \bar{x}\rangle \leq g_k \sup_{x,y \in \mathcal{F}_0} \|x - y\| = g_k D_0. \tag{47}$$

As a direct consequence we obtain that *weak* accumulation points of $(x_k)_{k\in\mathbb{N}}$ are global minima.

Proposition 4.7 (Optimality of weak accumulation points)**.** *Let, for problem* (P)*, Assumptions 1, 2 and 5 hold. Let additionally F have a bounded sublevel set $\mathcal{F}_0$ and be convex on $\mathcal{F}_0$. Then, all weak accumulation points of the sequence $(x_k)_{k\in\mathbb{N}}$ of the iterates of Algorithm 1 are global minima of problem* (P)*.*

Proof. Note that since $\mathcal{F}_0$ is bounded and $(x_k)_{k \in \mathbb{N}} \subset \mathcal{F}_0$, the set of weak accumulation points is not empty. Pick an accumulation point $\bar{x} \in \mathcal{H}$ and a subsequence such that $x_{k_l} \rightharpoonup \bar{x}$ as $l \to \infty$. By (47), we conclude for any minimiser $x^* \in \mathcal{H}$,

$$F(x_{k_l}) - F(x^*) \leq g_{k_l} D_0. \tag{48}$$

By Theorem 4.1, we have that $g_k \to 0$ as $k \to \infty$ and consequently, by (48), also

$$F(x^*) = \lim_{l \to \infty} F(x_{k_l}) = \liminf_{l \to \infty} F(x_{k_l}) \geq F(\bar{x})$$

by weak lower semicontinuity of the convex function F. Therefore, $\bar{x}$ is itself a global minimiser. $\qquad \square$

Remark 4.8. *If the minimiser is unique, we deduce weak convergence of the full sequence to the unique minimiser.*

We now present the main result of this subsection on the global nonasymptotic convergence rate of Algorithm 1 when F is convex on $\mathcal{F}_0$.

Theorem 4.9 (Global convergence rate, convex case)**.** *Let, for problem* (P)*, Assumptions 1, 2 and 5 hold. Let also $(x_k)_{k \in \mathbb{N}}$ be the iterates of Algorithm 1 with arbitrary starting point $x_0 \in \operatorname{dom} F$. Let additionally F have a bounded sublevel set $\mathcal{F}_0$ and be convex on $\mathcal{F}_0$. Then, the following global nonasymptotic convergence rate holds:*

$$F(x_k) - F^* \leq \|F'(x_0)\|_* D_0 \exp\left(-\frac{k}{4}\right) + \frac{2D_0^2 \max\{2mL, \Lambda_0\}}{\alpha k}, \quad k \geq 0. \tag{49}$$

If in addition $\lambda_k \to 0$ as $k \to \infty$, we obtain that $F(x_k) - F^* = o(1/k)$. Moreover, the number of Newton steps (17) *up to the end of iteration k does not exceed*

$$k + 1 + \log_2 \max\{1/2, mL/\Lambda_0\}.$$

Proof. From Theorem 4.1, we know that $(x_k)_{k \in \mathbb{N}} \subset \mathcal{F}_0$ is well defined. Hence, we have the convexity assumption at our disposal for the sequence $(x_k)_{k \in \mathbb{N}}$, by assumption. This, in particular, means that we can use inequalities (45) and (47) for all $k \geq 0$.

The proof of the convergence rate is based on the per iteration decrease inequality (45) and follows the same lines as in [31]. Recall from (21) that we denote $g_k := \|F'(x_k)\|_*$ and $F_k := F(x_k) - F^* \geq 0$. We start with the following chain of inequalities for any iteration $k \geq 0$

$$\frac{1}{F_{k+1}} - \frac{1}{F_k} = \frac{F_k - F_{k+1}}{F_k F_{k+1}} \overset{(45)}{\geq} \frac{\alpha}{\lambda_k F_k F_{k+1}} g_{k+1}^2 = \frac{\alpha}{\lambda_k F_k F_{k+1}} \left(\frac{g_{k+1}}{g_k}\right)^2 g_k^2$$

$$\overset{(47)}{\geq} \frac{\alpha}{D_0^2 \lambda_k F_k F_{k+1}} \left(\frac{g_{k+1}}{g_k}\right)^2 F_k^2 \geq \frac{\alpha}{D_0^2 \lambda_k} \left(\frac{g_{k+1}}{g_k}\right)^2,$$

where in the last step we used that, by (45), $F_k \geq F_{k+1}$. Summing these inequalities for j from 0 to

$k-1$ and using the inequality between arithmetic and geometric means (AM-GM), we obtain

$$
\frac{1}{F_k} - \frac{1}{F_0} \geq \frac{\alpha}{D_0^2} \sum_{j=0}^{k-1} \frac{1}{\lambda_j} \left(\frac{g_{j+1}}{g_j} \right)^2 \geq \frac{\alpha}{D_0^2} k \left(\prod_{j=0}^{k-1} \frac{1}{\lambda_j} \left(\frac{g_{j+1}}{g_j} \right)^2 \right)^{1/k} = \frac{\alpha k}{D_0^2} \left(\frac{g_k}{g_0} \right)^{2/k} \left(\prod_{j=0}^{k-1} \frac{1}{\lambda_j} \right)^{1/k}
$$

$$
\overset{(47)}{\geq} \frac{\alpha k}{D_0^2} \left(\frac{F_k}{D_0 g_0} \right)^{2/k} \left(\prod_{j=0}^{k-1} \lambda_j \right)^{-1/k} = \frac{\alpha k}{D_0^2} \left(\frac{D_0 g_0}{F_k} \right)^{-2/k} \left(\prod_{j=0}^{k-1} \lambda_j \right)^{-1/k}
$$

$$
= \frac{\alpha k}{D_0^2} \exp\left(-\frac{2}{k} \ln\left(\frac{D_0 g_0}{F_k} \right) \right) \left(\prod_{j=0}^{k-1} \lambda_j \right)^{-1/k}
$$

$$
\geq \frac{\alpha k}{D_0^2} \left(1 - \frac{2}{k} \ln\left(\frac{D_0 g_0}{F_k} \right) \right) \left(\prod_{j=0}^{k-1} \lambda_j \right)^{-1/k},
$$

where in the last step we used that $\exp(x) \geq 1 + x$. Consider two cases. If $\frac{2}{k} \ln\left(\frac{D_0 g_0}{F_k} \right) \geq \frac{1}{2}$, we obtain $F_k \leq g_0 D_0 \exp\left(-\frac{k}{4} \right)$. Otherwise, if $\frac{2}{k} \ln\left(\frac{D_0 g_0}{F_k} \right) < \frac{1}{2}$, we obtain

$$
\frac{1}{F_k} \geq \frac{\alpha k}{2 D_0^2} \left(\prod_{j=0}^{k-1} \lambda_j \right)^{-1/k} \quad \Longleftrightarrow \quad F_k \leq \frac{2 D_0^2}{\alpha k} \left(\prod_{j=0}^{k-1} \lambda_j \right)^{1/k}. \tag{50}
$$

Combining these two cases, we obtain

$$
F(x_k) - F^* = F_k \leq g_0 D_0 \exp\left(-\frac{k}{4} \right) + \frac{2 D_0^2}{\alpha k} \left(\prod_{j=0}^{k-1} \lambda_j \right)^{1/k}.
$$

Applying the bound $\lambda_k \leq \bar{\lambda} = \max\{2mL, \Lambda_0\}$ from Lemma 3.3, we obtain the convergence rate result (49) since $g_0 := \|F'(x_0)\|_*$. Further, if $\lambda_k \to 0$ as $k \to \infty$, we obtain that $\left(\prod_{j=0}^{k-1} \lambda_j \right)^{1/k} \to 0$ since $\lambda_k \geq 0$ for $k \geq 0$, and, hence, $F(x_k) - F^* = o(1/k)$. The bound on the number of Newton steps is obtained in the same way as in Theorem 4.1. $\qquad \square$

Remark 4.10 (Choice of the parameters m, α, β)**.** *First, observe that the proof of the convergence rate in (49) was based on (45), which in turn follows from the first inequality in (18) and convexity. Thus, if it is known that F is convex, the second inequality in (18) can be omitted, and the algorithm would simplify. This also means that in the convex case β can be chosen arbitrarily. Nevertheless, we keep the second inequality in (18) to make the algorithm agnostic to the knowledge of whether the problem is convex or not. Further, as one can see, the rate in (49) is proportional to m/α and the smaller that number, the better the rate. This, in particular, means that α should be chosen as large as possible and m should be as small as possible. Recall from Lemma 3.3 that $0 < \alpha \leq 1/2$, and $1 \leq m$. Thus, the optimal choice in the convex case is $\alpha = 1/2, m = 1$. Yet, the standard choice $\alpha = 1/2, m = 2$ also works.*

We can show that the convexity of F also allows for improving the global nonasymptotic convergence rate in terms of the minimal norm of the subgradient on the trajectory of Algorithm 1. Namely, in the next result, we obtain an improved $\mathcal{O}(1/k)$ rate compared to $\mathcal{O}(1/\sqrt{k})$ in Theorem 4.1.

Theorem 4.11 (Global convergence rate for subgradients, convex case). *Let assumptions of Theorem 4.9 hold. Then, the following global nonasymptotic convergence rate holds:*

$$\min_{i=k,\ldots,2k} \|F'(x_{i+1})\|_* \leq 2\sqrt{\frac{\max\{2mL, \Lambda_0\}\|F'(x_0)\|_* D_0}{\alpha k}} \exp\left(-\frac{k}{8}\right)$$
$$+ \frac{4D_0 \max\{2mL, \Lambda_0\}}{\alpha k}, k \geq 0. \tag{51}$$

Proof. The proof is an adaptation of [77] to our setting with a more complicated rate from (49). Summing (45) from k to $2k$, recalling from (21) the notation $g_{i+1} := \|F'(x_{i+1})\|_*$, using (49), and the bound $\lambda_k \leq \overline{\lambda}$ from Lemma 3.3, we obtain

$$\frac{\alpha k}{\overline{\lambda}} \min_{i=k,\ldots,2k} \|F'(x_{i+1})\|_*^2 \overset{\lambda_k \leq \overline{\lambda}, (24)}{\leq} F(x_{2k+1}) - F^* + \sum_{i=k}^{2k} \frac{\alpha g_{i+1}^2}{\lambda_i}$$

$$\overset{(45)}{\leq} F(x_{2k+1}) - F^* + \sum_{i=k}^{2k} (F(x_i) - F(x_{i+1})) = F(x_k) - F^*$$

$$\overset{(49)}{\leq} g_0 D_0 \exp\left(-\frac{k}{4}\right) + \frac{2D_0^2 \max\{2mL, \Lambda_0\}}{\alpha k}.$$

Rearranging, using the inequality $\sqrt{a^2 + b^2} \leq 2(a + b)$ for $a, b \geq 0$, and recalling from (20) that $\overline{\lambda} = \max\{2mL, \Lambda_0\}$, we obtain (51).

□

Note that if in addition $\lambda_k \to 0$ as $k \to \infty$, the rate in the above result improves to $o(1/k)$.

5 Local linear convergence

In this section, we return to the nonconvex setting and show local nonasymptotic linear convergence of Algorithm 1 to a local minimum of problem (P) under local versions of Assumptions 2 and 3, and a local QG condition. This result complements Theorem 4.4 in the sense that all the assumptions are made locally on a ball around a local minimum, whereas in Theorem 4.4 we made for example Assumption 2 globally on $\mathrm{dom}\, f$. Our motivation for this result comes from the classical results on SSN which make only local assumptions of semismoothness and non-degeneracy at a solution and prove local superlinear convergence, see, e.g., Theorem 2.11 in [50]. The main result of this section is that under similar assumptions, but in a more general setting, Algorithm 1 converges to a local minimum. In the next section, we strengthen this result by showing that under additional assumptions including semismoothness the convergence is actually superlinear, like in the classical theory.

In this section, we let x^* be an isolated local minimum in problem (P). This means that there exists some $R > 0$ such that

$$F(x) \geq F(x^*), \quad \mathrm{dist}(x, S) = \|x - x^*\|, \quad \forall x \in B_R(x^*). \tag{52}$$

Theorem 5.1 (Local linear convergence). *Let, for problem* (P)*, Assumptions 1 and 5 hold. Let x^* be an isolated local minimum in* (P) *with the local minimal value $F(x^*)$ and let R be such that (52) holds, Assumption 2 holds locally on $B_R(x^*)$ (instead of $\mathrm{dom}\, f$), the PL condition Assumption 3 holds*

locally on $B_R(x^*)$ (instead of $\operatorname{dom} F$), $H(x) \succeq \mu I$ for some $\mu > 0$ and for all $x \in B_R(x^*)$, and the quadratic growth (QG) condition holds locally on $B_R(x^*)$, i.e.,

$$\frac{\mu}{2} \operatorname{dist}(x, S)^2 \leq F(x) - F(x^*), \quad \forall x \in B_R(x^*). \tag{53}$$

Let also $(x_k)_{k \in \mathbb{N}}$ be the iterates of Algorithm 1 with starting point $x_0 \in \operatorname{dom} F$. Then, there exists Δ such that for any $x_0 \in B_R(x^*)$ with $\|F'(x_0)\|_* \leq \Delta$, x_k converges to x^* linearly and $F(x_k)$ converges to $F(x^*)$ linearly (explicit rates are given in the proof). If in addition $\lambda_k \to 0$ as $k \to \infty$, the convergence is superlinear.

Proof. Define constants $\varkappa := \frac{\overline{\lambda}}{\overline{\lambda} + 2\mu\beta\alpha^2} < 1$, $c := 1 + \sqrt{\frac{\overline{\lambda}}{\mu\beta\alpha^2}}$, where $\overline{\lambda}$ is defined in Lemma 3.3 (see (20)) and choose

$$0 < \Delta \leq \frac{\mu R}{1 + c}. \tag{54}$$

Recall from (21) that $g_k := \|F'(x_k)\|_*$. We slightly abuse notation for the purpose of this proof and define $F_k := F(x_k) - F(x^*)$, where $F(x^*)$ is the minimal value of F on $B_R(x^*)$, cf. with global minimal value F^* in (21). We also note that the local PL condition on $B_R(x^*)$ implies that whenever $x_k \in B_R(x^*)$, we have that

$$F(x_k) - F(x^*) \leq \frac{1}{2\mu} \|F'(x_k)\|_*^2. \tag{55}$$

Let $x_0 \in B_R(x^*)$ be such that $\|F'(x_0)\|_* \leq \Delta$. We proceed by induction and prove that for $k \geq 0$,

$$\|F'(x_k)\|_* \leq c\Delta, \tag{56}$$

$$F_k = F(x_k) - F(x^*) \leq \varkappa^k \frac{\Delta^2}{2\mu}, \tag{57}$$

$$\|x_k - x^*\| \leq \varkappa^{k/2} \frac{\Delta}{\mu}. \tag{58}$$

Since $x_0 \in B_R(x^*)$, we can apply the assumptions of the theorem to show that the base case holds. Indeed, since $c \geq 1$, we have that $\|F'(x_0)\|_* \leq \Delta \leq c\Delta$. Since x^* is an isolated local minimum and the QG condition holds, we have

$$\frac{\mu}{2} \|x_0 - x^*\|^2 \overset{(52),(53)}{\leq} F(x_0) - F(x^*) = F_0 \overset{(55)}{\leq} \frac{1}{2\mu} \|F'(x_0)\|_*^2 \leq \frac{\Delta^2}{2\mu} \Rightarrow \|x_0 - x^*\| \leq \frac{\Delta}{\mu}. \tag{59}$$

Thus, (56), (57), (58) hold for $k = 0$.

We now proceed with the induction step. Let (56), (57), (58) hold. Note that (58), (54), and $\varkappa < 1$ imply that $\|x_k - x^*\| \leq \varkappa^{k/2} \frac{\Delta}{\mu} \leq R$ and, hence, $x_k \in B_R(x^*)$. Thus, by the assumption that $H(x) \succeq \mu I$ on $B_R(x^*)$ with $\mu > 0$, recalling the notation $x_+(\lambda, x)$ from $(\mathrm{P_N})$, we obtain from Lemma 2.7 that

$$\|x_+(\lambda, x_k) - x_k\| \overset{(13)}{\leq} \frac{\|F'(x_k)\|_*}{\lambda + \mu} \overset{(56)}{\leq} \frac{c\Delta}{\mu}, \quad \forall \lambda \geq 0.$$

This in combination with (58) and $\varkappa < 1$ gives

$$\|x_+(\lambda, x_k) - x^*\| \leq \|x_+(\lambda, x_k) - x_k\| + \|x_k - x^*\| \leq \frac{c\Delta}{\mu} + \frac{\Delta}{\mu} \overset{(54)}{\leq} R, \quad \forall \lambda \geq 0.$$

Hence, for all $\lambda \geq 0$, we have $x_+(\lambda, x_k) \in B_R(x^*)$ and $[x_k, x_+(\lambda, x_k)] \subset B_R(x^*)$. By the assumption that Assumption 2 holds on $B_R(x^*)$, we observe that it holds also on $[x_k, x_+(\lambda, x_k)]$ for all

$\lambda \geq 0$. Thus, Lemma 3.3 applies. Whence, the next iterate x_{k+1} is well defined, we have $\lambda_k \leq \overline{\lambda}$ by inequality (20), and

$$F_k - F_{k+1} = F(x_k) - F(x_{k+1}) \overset{(26)}{\geq} \frac{\beta\alpha^2}{\lambda_k}\|F'(x_{k+1})\|_*^2 \overset{\lambda_k \leq \overline{\lambda}}{\geq} \frac{\beta\alpha^2}{\overline{\lambda}}\|F'(x_{k+1})\|_*^2 \geq 0. \tag{60}$$

Since $x_{k+1} = x_+(\lambda_k, x_k) \in B_R(x^*)$, we can use the localised PL condition Assumption 3 at x_{k+1} and obtain

$$F_k - F_{k+1} \overset{(60)}{\geq} \frac{\beta\alpha^2}{\lambda_k}\|F'(x_{k+1})\|_*^2 \overset{(55)}{\geq} \frac{2\mu\beta\alpha^2}{\lambda_k}(F(x_{k+1}) - F(x^*)) = \frac{2\mu\beta\alpha^2}{\lambda_k}F_{k+1}. \tag{61}$$

Rearranging and using the QG condition, we obtain

$$\frac{\mu}{2}\|x_{k+1} - x^*\|^2 \overset{(52),(53)}{\leq} F(x_{k+1}) - F(x^*) = F_{k+1} \overset{(61)}{\leq} \frac{F_k}{1 + \frac{2\mu\beta\alpha^2}{\lambda_k}} = \frac{\lambda_k F_k}{\lambda_k + 2\mu\beta\alpha^2}$$

$$\overset{\lambda_k \leq \overline{\lambda},(57)}{\leq} \frac{\overline{\lambda}}{\overline{\lambda} + 2\mu\beta\alpha^2}\varkappa^k\frac{\Delta^2}{2\mu} = \varkappa^{k+1}\frac{\Delta^2}{2\mu} \Rightarrow \|x_{k+1} - x^*\| \leq \varkappa^{(k+1)/2}\frac{\Delta}{\mu}. \tag{62}$$

Thus, we obtain (57) and (58) for $k+1$. Finally, we show (56) for $k+1$ by contradiction. Assume the opposite, i.e.,

$$g_{k+1} \overset{(21)}{:=} \|F'(x_{k+1})\|_* > c\Delta. \tag{63}$$

Since $x_k, x_{k+1} \in B_R(x^*)$, by the first inequality in (52), definition of F_k, F_{k+1}, and by induction from (60), we have $0 \leq F_{k+1} \leq F_k \leq F_0$. Hence,

$$0 \leq F_{k+1} \overset{(60)}{\leq} F_k - \frac{\beta\alpha^2}{\overline{\lambda}}g_{k+1}^2 \overset{F_k \leq F_0}{\leq} F_0 - \frac{\beta\alpha^2}{\overline{\lambda}}g_{k+1}^2 \overset{(59),(63)}{<} \frac{\Delta^2}{2\mu} - \frac{\beta\alpha^2}{\overline{\lambda}}c^2\Delta^2$$

$$< \frac{\Delta^2}{2\mu} - \frac{\beta\alpha^2}{\overline{\lambda}}\Delta^2\left(\sqrt{\frac{\overline{\lambda}}{\mu\beta\alpha^2}}\right)^2 < 0,$$

which is a contradiction. Thus, we have (56) for $k+1$.

We finish the proof by observing that (58) implies that $x_k \to x^*$ linearly:

$$\|x_k - x^*\| \leq \left(1 - \frac{2\mu\beta\alpha^2}{\overline{\lambda} + 2\mu\beta\alpha^2}\right)^{k/2}\frac{R}{1+c} \overset{(20)}{\leq} \exp\left(-\frac{2\beta\alpha^2\mu}{4\beta\alpha^2\mu + 2\max\{2mL, \Lambda_0\}}\cdot k\right)\cdot\frac{R}{1+c},$$

and that (57) implies $F(x_k)$ converges to $F(x^*)$ linearly:

$$F(x_k) - F(x^*) \leq \left(1 - \frac{2\mu\beta\alpha^2}{\overline{\lambda} + 2\mu\beta\alpha^2}\right)^k\frac{\mu R^2}{2(1+c)^2}$$

$$\overset{(20)}{\leq} \exp\left(-\frac{2\beta\alpha^2\mu}{\beta\alpha^2\mu + \max\{2mL, \Lambda_0\}}\cdot k\right)\cdot\frac{\mu R^2}{2(1+c)^2}.$$

Moreover, from the first two inequalities and first two equalities in (62), we see that if $\lambda_k \to 0$ as $k \to \infty$, the convergence of both x_k and $F(x_k)$ is superlinear. $\square$

We remark that assuming QG condition (53), $H(x) \succ 0$ on $B_R(x^*)$ and that x^* is an isolated minimum may be considered as a price for making Assumption 2 only locally, cf. Theorem 4.4, where

it is made globally. Indeed, we use those assumptions to guarantee that the iterates of the algorithm stay in the ball $B_R(x^*)$ where we can apply local Assumption 2. In future works, it would be interesting to avoid these assumptions and prove local convergence based only on the PL condition in the spirit of [42]. In particular, we hope that this will allow us to omit the assumption that x^* is isolated, whose main purpose currently is to establish the equality in (52) that allows for using the QG condition. In fact, we expect that the QG condition follows from the assumed PL condition [53], however, we were not able to find a proof of this statement needed for our precise setting. We also note that the assumption of x^* being an isolated minimum is, in a sense, standard since many existing papers use (local) strong convexity, see e.g., [31, 32], or other assumptions [38] that imply that the convergence is established to an isolated (local) minimum.

As already mentioned, our inspiration for the above result comes from the classical results on the local convergence of SSN (with $\psi \equiv 0$), see, e.g., Theorem 2.11 in [50], where all the assumptions are made locally. In particular, it is assumed that the generalised derivative $H(x)$ is non-singular at x^* and f' is semismooth at x^*. Since $H(x)$ is non-singular at x^* it is non-singular in some $B_R(x^*)$ and there is a $\mu > 0$ that lower bounds the smallest eigenvalue of $H(x)$. This justifies the assumption $H(x) \succ 0$ on $B_R(x^*)$. Moreover, this means that f behaves like a strongly convex functional locally, which implies both local PL and QG conditions. Thus, our setting is close to the classical one, yet we allow for some nonconvexity.

We conclude this subsection with the following technical result, which is used in the proofs of superlinear convergence in the next section. This result provides a relation of the distance to solution $\|x_k - x^*\|$ to the distance between two consecutive iterates $\|x_{k+1} - x_k\|$, analogous to Lemma 4.5, but under local assumptions.

Lemma 5.2. *Let assumptions of Theorem 5.1 hold. Then, there exist constants $c_1, c_2 > 0$ such that, for all $k \geq 0$,*

$$\|x_{k+1} - x^*\| \leq c_1 \|x_k - x_{k+1}\| \quad \text{and} \quad \|x_k - x^*\| \leq c_2 \|x_k - x_{k+1}\|. \tag{64}$$

Proof. As shown in the proof of Theorem 5.1, we have that $x_k \in B_R(x^*)$ for all $k \geq 0$. Thus,

$$\frac{\mu}{2}\|x_{k+1} - x^*\|^2 \overset{(52),(53)}{\leq} F(x_{k+1}) - F(x^*) \overset{(55)}{\leq} \frac{1}{2\mu}\|F'(x_{k+1})\|_*^2.$$

By (25) and the bound $\lambda_k \leq \overline{\lambda}$ from Lemma 3.3, we have

$$\|F'(x_{k+1})\|_* \overset{(25)}{\leq} \frac{\lambda_k}{\alpha}\|x_k - x_{k+1}\| \leq \frac{\overline{\lambda}}{\alpha}\|x_k - x_{k+1}\|,$$

which in combination with the previous inequality gives the first inequality in (64) with the constant $c_1 = \frac{\overline{\lambda}}{\alpha\mu}$. The second inequality with $c_2 = 1 + c_1$ follows from the triangle inequality. $\qquad\square$

6 Superlinear convergence under higher regularity

In this section, based on the additional assumptions on the regularity of f', we further refine the convergence rates established in previous sections. We start with several technical lemmas that are used in what follows. Then, we show that if $\mathcal{H}$ is finite dimensional and f is twice continuously differentiable, i.e., $f \in C^2$, then in Algorithm 1 $\lambda_k \to 0$ and the convergence rates proved in previous sections asymptotically improve. After that, we move on to the main setting of the paper and consider

the behaviour of Algorithm 1 under a semismoothness assumption, which allows us to obtain its superlinear convergence. Finally, based on the active manifold identification framework, we show that superlinear convergence can also occur when f is not locally C^2, but instead satisfies a combination of partial smoothness and semismoothness of the derivative. To simplify the proofs, we focus on the setting when F is locally convex, but this assumption can be slightly relaxed; we defer the details to future work.

6.1 Technical preliminaries

We start with several technical results that are crucial for establishing superlinear convergence of Algorithm 1. The first one establishes several useful inequalities for $x_+(\lambda, x_k)$.

Lemma 6.1. *Consider iteration $k \geq 0$ of Algorithm 1 and let $0 < \lambda < \lambda_k$. Let $H(x_k) \succeq 0$ and, for some constant $c > 0$, the following inequality hold:*

$$\|x_{k+1} - x^*\| \leq c\|x_k - x_{k+1}\|. \tag{65}$$

Then, we have

$$\|x_+(\lambda, x_k) - x^*\| \leq \frac{(2c+1)\lambda_k - (1+c)\lambda}{\lambda_k}\|x_+(\lambda, x_k) - x_k\|, \tag{66}$$

$$\|x_k - x^*\| \leq \frac{(2c+2)\lambda_k - (1+c)\lambda}{\lambda_k}\|x_+(\lambda, x_k) - x_k\|. \tag{67}$$

Before we start the proof, for the reader's convenience we note that when λ is chosen as $\lambda_k/2$, the constants in (66) and (67) are $\frac{1+3c}{2}$ and $\frac{3+3c}{2}$ respectively.

Proof of Lemma 6.1. Applying two times the triangle inequality and one time Lemma 2.8, recalling from Line 9 of Algorithm 1 and $(\mathrm{P_N})$ that $x_{k+1} = x_+(\lambda_k, x_k)$, we obtain

$$
\begin{aligned}
\|x_+(\lambda, x_k) - x^*\| &\leq \|x_+(\lambda, x_k) - x_{k+1}\| + \|x_{k+1} - x^*\| \\
&\overset{(65)}{\leq} \|x_+(\lambda, x_k) - x_{k+1}\| + c\|x_{k+1} - x_k\| \\
&\leq \|x_+(\lambda, x_k) - x_{k+1}\| + c(\|x_+(\lambda, x_k) - x_{k+1}\| + \|x_+(\lambda, x_k) - x_k\|) \\
&= (1+c)\|x_+(\lambda, x_k) - x_{k+1}\| + c\|x_+(\lambda, x_k) - x_k\|) \\
&\overset{(14)}{\leq} (1+c)\frac{\lambda_k - \lambda}{\lambda_k}\|x_+(\lambda, x_k) - x_k\| + c\|x_+(\lambda, x_k) - x_k\|,
\end{aligned}
$$

which is (66). Applying once again the triangle inequality, we obtain (67):

$$
\begin{aligned}
\|x_k - x^*\| &\leq \|x_k - x_+(\lambda, x_k)\| + \|x_+(\lambda, x_k) - x^*\| \\
&\overset{(66)}{\leq} \|x_k - x_+(\lambda, x_k)\| + \frac{(2c+1)\lambda_k - (1+c)\lambda}{\lambda_k}\|x_+(\lambda, x_k) - x_k\|.
\end{aligned}
$$

$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

Note that the technical condition (65) is guaranteed by Lemma 4.5 or Lemma 5.2. The next result establishes the convergence of non-accepted trial points $x_+(\lambda_k/2, x_k)$ to x^* when accepted points x_k converge to x^*.

Lemma 6.2. *Let $(x_k)_{k\in\mathbb{N}}$ be the iterates of Algorithm 1 and $x_k \to x^*$ as $k \to \infty$. Let also $H(x_k) \succeq 0$ for sufficiently large k. Then, $x_+(\lambda_k/2, x_k) \to x^*$ as $k \to \infty$.*

Proof. For any sufficiently large k and $0 < \lambda \le \lambda_k$, by the triangle inequality, the definition $x_{k+1} = x_+(\lambda_k, x_k)$ in Line 9 of Algorithm 1 and $(\mathrm{P_N})$, and Lemma 2.8, we have

$$
\|x_k - x_+(\lambda, x_k)\| \le \|x_{k+1} - x_+(\lambda, x_k)\| + \|x_{k+1} - x_k\|
$$
$$
\overset{(14)}{\le} \frac{\lambda_k - \lambda}{\lambda_k}\|x_k - x_+(\lambda, x_k)\| + \|x_{k+1} - x_k\|.
$$

Rearranging, we obtain

$$
\|x_k - x_+(\lambda, x_k)\| \le \frac{\lambda_k}{\lambda}\|x_{k+1} - x_k\|.
$$

Now, utilising again the triangle inequality, we calculate

$$
\|x_+(\lambda, x_k) - x^*\| \le \|x_+(\lambda, x_k) - x_k\| + \|x_k - x^*\| \le \frac{\lambda_k}{\lambda}\|x_{k+1} - x_k\| + \|x_k - x^*\|.
$$

The claim follows by taking $\lambda = \lambda_k/2$, using the upper bound $\lambda_k \le \overline{\lambda}$ from Lemma 3.3, and taking the limit. $\qquad\square$

The following result will be crucial to show that in Algorithm 1 $\lambda_k \to 0$ when f has additional regularity. On a high level, it can be interpreted as $\lambda_k \to 0$ if the constant L in Assumption 2 goes to 0 around the trajectory of Algorithm 1. It is, in some sense, natural to expect this from (20), and we now prove it in rigour.

Lemma 6.3. *Let F be locally convex around x^*. Let $(x_k)_{k\in\mathbb{N}}$ be the iterates of Algorithm 1 and $x_k \to x^*$ as $k \to \infty$. Let also $H(x_k) \succeq 0$ for sufficiently large k. If there exists a sequence $(\varepsilon_k)_{k\in\mathbb{N}} \subset \mathbb{R}_+$ such that $\varepsilon_k \to 0$ and, for sufficiently large k,*

$$
\begin{aligned}
\|f'(x_+(\lambda_k/2, x_k)) &- f'(x_k) - H(x_k)(x_+(\lambda_k/2, x_k) - x_k)\|_* \\
&\le \varepsilon_k\|x_k - x_+(\lambda_k/2, x_k)\|,
\end{aligned}
\tag{68}
$$

then, we have $\lambda_k \to 0$ as $k \to \infty$.

Proof. For brevity, given an index $k \ge 0$, we denote by $x_{k,+} := x_+(\lambda_k/2, x_k)$ the solution of the auxiliary problem $(\mathrm{P_N})$ given current iterate x_k and accepted regularisation parameter $\lambda_k > 0$. Moreover, we set $r_{k,+} := \|x_k - x_{k,+}\|$. By assumption, there exists K_1 such that (68) holds for all $k \ge K_1$. Thus, for any $k \ge K_1$, using (68), Lemma 2.9 with $x = x_k$, $x_+ = x_{k,+}$, $\varphi(r) = \varepsilon_k r$, and $\lambda = \lambda_k/2$, and that $\alpha \le 1/2$, we obtain

$$
\begin{aligned}
\langle F'(x_{k,+}), x_k - x_{k,+}\rangle &\overset{(16),(68)}{\ge} \frac{1}{\lambda_k}\|F'(x_{k,+})\|_*^2 + \left(\frac{\lambda_k^2 - 4\varepsilon_k^2}{4\lambda_k}\right)r_{k,+}^2 \\
&\ge \frac{2\alpha}{\lambda_k}\|F'(x_{k,+})\|_*^2 + \left(\frac{\lambda_k^2 - 4\varepsilon_k^2}{4\lambda_k}\right)r_{k,+}^2.
\end{aligned}
\tag{69}
$$

By the local convexity of F around x^*, there exists a ball $B_R(x^*)$ such that F is convex on that ball. Since by the assumptions of the lemma $x_k \to x^*$ and, for sufficiently large k, $H(x_k) \succeq 0$, we have by Lemma 6.2 that $x_{k,+} \to x^*$ as well. Thus, there exists K_2 s.t. for all $k \ge K_2$, we

have $x_k, x_{k,+} \in B_R(x^*)$. Applying convexity of F on $B_R(x^*)$, using that the first term on the right-hand side in (69) is nonnegative, and adding and subtracting $\beta \lambda_k r_{k,+}^2 / 2$, we obtain from (69) for all $k \geq \max\{K_1, K_2\}$

$$
F(x_k) - F(x_{k,+}) \geq \langle F'(x_{k,+}), x_k - x_{k,+} \rangle \geq \frac{\beta \lambda_k r_{k,+}^2}{2} + \left(\frac{\lambda_k^2 - 4\varepsilon_k^2 - 2\beta\lambda_k^2}{4\lambda_k} \right) r_{k,+}^2
$$
$$
= \frac{\beta \lambda_k r_{k,+}^2}{2} + \left(\frac{(1 - 2\beta)\lambda_k^2 - 4\varepsilon_k^2}{4\lambda_k} \right) r_{k,+}^2. \quad (70)
$$

From now on, we assume $k \geq \max\{K_1, K_2\}$. From Line 9 of Algorithm 1, we know that $\lambda_k = 2^{j_k} \Lambda_k$ for some $j_k \geq 0$. Consider the two cases: when $j_k = 0$ and when $j_k > 0$. In the first case, i.e., when $j_k = 0$, acceptance occurs immediately in the first round of the inner loop. By Line 9 of Algorithm 1, $\Lambda_k = \lambda_{k-1}/2$, and hence in this case $\lambda_k = 2^{j_k} \Lambda_k = 2^0 \Lambda_k = \Lambda_k = \lambda_{k-1}/2$. In the second case, i.e., when $j_k > 0$, $\lambda_k/2$ must have been rejected. This implies that the candidate point $x_+ (\lambda_k/2, x_k)$ and candidate regularisation parameter $\lambda_k/2$ did not satisfy the acceptance conditions (18). Thus, we have

$$
\langle F'(x_{k,+}), x_k - x_{k,+} \rangle < \frac{2\alpha}{\lambda_k} \|F'(x_{k,+})\|_*^2 \quad \text{or} \quad F(x_k) - F(x_{k,+}) < \frac{\beta \lambda_k r_{k,+}^2}{2}. \quad (71)
$$

If the first inequality in (71) holds, from (69) we must have $\lambda_k \leq 2\varepsilon_k$ since otherwise, we are led to a contradiction. If the second inequality in (71) holds, then in the same way, (70) and the fact $\beta \leq (m-1)/(2m) < 1/2$ (see Line 1 of Algorithm 1) imply that $\lambda_k \leq 2\varepsilon_k/\sqrt{1-2\beta}$. Thus, in both cases we have $\lambda_k \leq \max\{2, 2/\sqrt{1-2\beta}\}\varepsilon_k$. Combining this with the case $j_k = 0$ and denoting $\gamma := \max\{2, 2/\sqrt{1-2\beta}\}$, we have, for $k \geq \max\{K_1, K_2\}$,

$$
\lambda_k \leq \begin{cases} \frac{\lambda_{k-1}}{2}, & j_k = 0, \\ \gamma \varepsilon_k, & j_k > 0. \end{cases}
$$

Now let us show that this implies $\lambda_k \to 0$. Define $\mathcal{I} := \{k \geq 0 : j_k > 0\} = \{k_1 < k_2 < \dots\}$. Thus if $k \in \mathcal{I}$, $\lambda_k \leq \gamma \varepsilon_k$, whereas if $k \notin \mathcal{I}$, we have $\lambda_k \leq \lambda_{k-1}/2$. Now, if $|\mathcal{I}| < \infty$, then for sufficiently large k, we have $\lambda_k \leq \lambda_{k-1}/2$ and hence $\lambda_k \to 0$. So let us suppose that $|\mathcal{I}| = \infty$. Fix $\delta > 0$ and choose i_0 such that $\gamma \varepsilon_k \leq \delta$ for all $k \geq k_{i_0}$. Take $k \geq k_{i_0}$. Then either k belongs to $\mathcal{I}$ or not. If $k \notin \mathcal{I}$, there exists $i \geq i_0$ such that $k_i < k < k_{i+1}$ with $k_i, k_{i+1} \in \mathcal{I}$ (recall $\{k_i\}$ from the definition of $\mathcal{I}$). Thus, we have

$$
\lambda_k \leq \frac{\lambda_{k_i}}{2^{k-k_i}} \leq \frac{\gamma \varepsilon_{k_i}}{2^{k-k_i}} \leq \gamma \varepsilon_{k_i} \leq \delta,
$$

where we used that $k - k_i \geq 0$ and hence $2^{k-k_i} \geq 1$. On the other hand, if $k \in \mathcal{I}$, we have that $k = k_i$ for some $i \geq i_0$, and so $\lambda_k \leq \gamma \varepsilon_k \leq \delta$. In total, we infer $\lambda_k \leq \delta$ for all $k \geq k_{i_0}$. Since $\delta > 0$ was chosen arbitrarily, it follows that $\lambda_k \to 0$ as $k \to \infty$. $\qquad \square$

Remark 6.4. *When F is globally convex, $\dim(\mathcal{H}) < \infty$, and the minimiser x^* is unique, then $x_k \to x^*$ holds automatically in the above lemma.*

6.2 Faster asymptotic rates in the finite-dimensional C^2 setting

In this subsection, we present our first result on improved asymptotic rates of Algorithm 1 when $\mathcal{H}$ is finite-dimensional and f possesses additional regularity. Namely, we assume that $f \in C^2$, i.e., f is twice continuously differentiable. The main idea is to prove that if $x_k \to x^*$ as $k \to \infty$, then we also have that $\lambda_k \to 0$ as $k \to \infty$. According to the results of the previous sections, the latter implies faster asymptotic rates, including superlinear convergence.

Theorem 6.5 (Improved asymptotic rates, finite dimensions, C^2 case)**.** *Let, for problem* (P)*, Assumptions 1 and 5 hold. Let* $\dim(\mathcal{H}) < \infty$*, f be locally twice continuously differentiable and locally convex around some point x^*, and $H(x) := \nabla^2 f(x)$. Let $(x_k)_{k \in \mathbb{N}}$ be the iterates of Algorithm 1. Then, the following hold.*

   (i) If $x_k \to x^*$ as $k \to \infty$, then we have that $\lambda_k \to 0$ as $k \to \infty$.

   (ii) If the assumptions of Theorem 4.1 for the nonconvex setting hold and $x_k \to x^*$, then
   $\min_{0 \leq i \leq k-1} \|F'(x_{i+1})\|_* = o(1/\sqrt{k})$.

   (iii) If the assumptions of Theorem 4.4 for the PL setting hold, then $F(x_k) - F^*$ converges to 0 superlinearly and x_k converges to x^* superlinearly, where x^* is a global solution to (P)*.*

   (iv) If the assumptions of Theorem 4.9 for the convex setting hold and $x_k \to x^*$, then $F(x_k) - F^* = o(1/k)$.

   (v) If the assumptions of Theorem 5.1 on local convergence hold, then x_k converges to x^* locally superlinearly and $F(x_k)$ converges to $F(x^*)$ locally superlinearly, where x^* is a local solution to (P)*.*

Proof. (i) Again, given an index $k \geq 0$, we denote by $x_{k,+} := x_+(\lambda_k/2, x_k)$ (see ($\mathrm{P_N}$)). Since we are in the finite dimensional setting, we write $f' \equiv \nabla f$. We first show that

$$\omega_k := \frac{\|\nabla f(x_{k,+}) - \nabla f(x_k) - \nabla^2 f(x_k)(x_{k,+} - x_k)\|_*}{\|x_{k,+} - x_k\|} \to 0 \quad \text{as } k \to \infty. \tag{72}$$

For this purpose, fix an arbitrary $\varepsilon > 0$. By assumption, there exists a ball $B_R(x^*)$ on which $\nabla^2 f$ exists and is continuous on $B_R(x^*)$. Moreover, by local convexity and twice differentiability of f we obtain $H(x) := \nabla^2 f(x) \succeq 0$ for all $x \in B_R(x^*)$. In addition, using compactness of $B_R(x^*)$ and the Heine–Cantor theorem, we get that $\nabla^2 f$ is even uniformly continuous on $B_R(x^*)$. The latter means that there exists $\delta > 0$ such that for every $x, y \in B_R(x^*)$ with $\|x - y\| \leq \delta$, we have

$$\|\nabla^2 f(x) - \nabla^2 f(y)\|_{\mathrm{op}} \leq \varepsilon.$$

As $x_k \to x^*$, there exist K_1 such that $k \geq K_1$ implies $x_k \in B_R(x^*)$. This in turn implies that, for $k \geq K_1$, $H(x_k) \succeq 0$. We conclude by Lemma 6.2 that $x_{k,+} \to x^*$ as well. We further infer that there is an index $K_2 \geq K_1$ such that $x_k, x_{k,+} \in B_R(x^*)$ and $\|x_{k,+} - x_k\| \leq \delta$ for all $k \geq K_2$. Consequently, we deduce from the fundamental theorem of calculus that, for all $k \geq K_2$,

$$\left\|\nabla f(x_{k,+}) - \nabla f(x_k) - \nabla^2 f(x_k)(x_{k,+} - x_k)\right\|_*$$
$$= \left\|\int_0^1 (\nabla^2 f(x_k + t(x_{k,+} - x_k)) - \nabla^2 f(x_k))(x_{k,+} - x_k)\, \mathrm{d}t\right\|_* \leq \varepsilon\|x_{k,+} - x_k\|$$

as, clearly, for $x_t := x_k + t(x_{k,+} - x_k)$, we have $x_t \in B_R(x^*)$ and $\|x_t - x_k\| \leq \|x_{k,+} - x_k\| \leq \delta$ for any $t \in (0, 1)$. Since the choice of $\varepsilon > 0$ was arbitrary, we infer (72). Since F is locally convex around x^*, recalling that $x_{k,+} = x_+(\lambda_k/2, x_k)$ and applying Lemma 6.3 with $H(x_k) = \nabla^2 f(x_k)$ and $\varepsilon_k = \omega_k$, we get $\lambda_k \to 0$.

(ii) The claim follows by item (i) above and Theorem 4.1.

(iii) The claim follows by item (i) above and Theorem 4.4, which guarantees also that $x_k \to x^*$, where x^* is a global solution to (P).

(iv) The claim follows by item (i) above and Theorem 4.9.

(v) The claim follows by item (i) above and Theorem 5.1, which guarantees also that $x_k \to x^*$, where x^* is a local solution to (P). □

6.3 Superlinear convergence under semismoothness

In this subsection, we present one of our main results on superlinear convergence of Algorithm 1 in the semismooth setting, i.e., when Assumption 4 holds. This setting is less restrictive than the C^2 setting in the previous subsection. Yet, it still allows us to obtain superlinear convergence in the same spirit as for classical SSN. Note that we again consider infinite-dimensional spaces $\mathcal{H}$.

We start with the following result that establishes superlinear convergence if convergence itself is already established. Recall that Theorem 4.4 for the PL setting guarantees global convergence and Theorem 5.1 guarantees local convergence under PL and QG conditions. Thus, these results in combination with the following proposition establish respectively global superlinear and local superlinear convergence under semismoothness. As already mentioned, for simplicity we focus on the setting when F is locally convex.

Proposition 6.6 (Superlinear convergence under semismoothness). *Let, for problem* (P)*, Assumptions 1 and 5 hold. Let x^* be a local minimum in problem* (P)*, F be locally convex around x^*, and the semismoothness at x^* Assumption 4 hold. Let $(x_k)_{k\in\mathbb{N}}$ be the iterates of Algorithm 1 and $x_k \to x^*$ as $k \to \infty$. Assume also that, for sufficiently large k, $H(x_k) \succeq 0$, the PL condition* (27) *holds at x_k with $F_U^* = F(x^*)$, and there exists constants $c_1, c_2 > 0$ such that*

$$\|x_{k+1} - x^*\| \leq c_1\|x_k - x_{k+1}\| \quad \text{and} \quad \|x_k - x^*\| \leq c_2\|x_k - x_{k+1}\|. \tag{73}$$

Consider the following statements.

(i) The following Dennis–Moré-type (DM-type) condition holds (as $k \to \infty$):

$$\|(H(x_+(\lambda_k/2, x_k)) - H(x_k))(x_+(\lambda_k/2, x_k) - x^*)\|_* = \mathcal{o}(\|x_+(\lambda_k/2, x_k) - x_k\|). \tag{74}$$

(ii) The following condition holds:

$$\frac{\|f'(x_+(\lambda_k/2, x_k)) - f'(x_k) - H(x_k)(x_+(\lambda_k/2, x_k) - x_k)\|_*}{\|x_+(\lambda_k/2, x_k) - x_k\|} \to 0 \quad \text{as } k \to \infty. \tag{75}$$

(iii) The sequence of regularisation parameters goes to zero, i.e., $\lambda_k \to 0$ as $k \to \infty$.

(iv) The convergence $x_k \to x^*$ is superlinear, and in fact

$$\|x_{k+1} - x^*\| \leq \frac{2c_2\lambda_k^{1/2}}{\sqrt{\alpha^2\beta\mu}}\|x_k - x^*\| = \mathcal{o}(\|x_k - x^*\|) \quad \text{as } k \to \infty.$$

(v) The subgradients converge superlinearly to zero, and in fact

$$\|F'(x_{k+1})\|_* \leq \frac{\lambda_k^{1/2}}{\sqrt{2\alpha\mu}}\|F'(x_k)\|_* = \mathcal{o}(\|F'(x_k)\|_*) \quad \text{as } k \to \infty.$$

Then, we have the following relationship:

$$(i) \iff (ii) \implies (iii) \implies \begin{cases} (iv), \\ (v). \end{cases}$$

Proof. By the assumptions of the theorem, there exist a ball $B_R(x^*)$ such that

(1) $F(x) \geq F(x^*)$ for all $x \in B_R(x^*)$;

(2) F is convex on $B_R(x^*)$;

(3) By Assumption 4 of semismoothness at x^* it holds that

$$\|f'(x) - f'(x^*) - H(x)(x - x^*)\|_* = \mathcal{O}(\|x - x^*\|), \quad \text{as } x \to x^*. \tag{76}$$

For clarity, given an index $k \geq 0$, we again denote $x_{k,+} := x_+(\lambda_k/2, x_k)$ (see $(\mathrm{P_N})$). By the assumptions, we also have the following.

(4) There exists K_0 such that for all $k \geq K_0$, we have $H(x_k) \succeq 0$.

(5) Combining the previous item with $x_k \to x^*$, we have by Lemma 6.2 that $x_{k,+} \to x^*$ as well. Thus, there exists K_1 s.t. for all $k \geq K_1$, we have $x_k, x_{k,+} \in B_R(x^*)$ and, hence, we can use local optimality, convexity, and semismoothness.

(6) There exists K_2 such that for all $k \geq K_2$, the PL condition holds at x_k, i.e., (27) holds with $F_U^* = F(x^*)$. By the previous item, $x_k \in B_R(x^*)$ for all $k \geq K_1$. Also, we have $F(x) \geq F(x^*)$ for all $x \in B_R(x^*)$. Taking $U = B_R(x^*)$, $F_U^* = F(x^*)$ in Lemma 3.5 and combining with these observations, we can apply Lemma 3.5 for any $k \geq \max\{K_1, K_2\}$.

(7) There exists K_3 such that for all $k \geq K_3$ the inequalities (73) hold.

From now on, we consider arbitrary $k \geq \max\{K_0, K_1, K_2, K_3\}$ so that we can apply all the above seven items.

(i) $\implies$ (ii): Thanks to the first inequality in (73) and $H(x_k) \succeq 0$, we can use Lemma 6.1 with $\lambda = \lambda_k/2$ to obtain, recalling that $x_{k,+} = x_+(\lambda_k/2, x_k)$,

$$\|x_{k,+} - x^*\| \leq a\|x_{k,+} - x_k\|, \quad \|x_k - x^*\| \leq b\|x_{k,+} - x_k\|, \tag{77}$$

where $a, b > 0$ are constants. Using these inequalities, by the triangle inequality, semismoothness at x^* (76), and the Dennis–Moré-type condition (74), we have

$$\begin{aligned}
&\|f'(x_{k,+}) - f'(x_k) - H(x_k)(x_{k,+} - x_k)\|_* \\
&\leq \|f'(x_{k,+}) - f'(x^*) - H(x_{k,+})(x_{k,+} - x^*)\|_* \\
&\quad + \|f'(x_k) - f'(x^*) - H(x_k)(x_k - x^*)\|_* \\
&\quad + \|(H(x_{k,+}) - H(x_k))(x_{k,+} - x^*)\|_* \\
&\overset{(76),(74)}{=} \mathcal{O}(\|x_{k,+} - x^*\|) + \mathcal{O}(\|x_k - x^*\|) + \mathcal{O}(\|x_{k,+} - x_k\|) \\
&\overset{(77)}{=} \mathcal{O}(\|x_{k,+} - x_k\|),
\end{aligned}$$

which is (75).

(ii) $\implies$ (i): Analogous to the decomposition above, we have

$$\begin{aligned}
\|(H(x_{k,+}) - H(x_k))(x_{k,+} - x^*)\|_* &\leq \|f'(x_{k,+}) - f'(x^*) - H(x_{k,+})(x_{k,+} - x^*)\|_* \\
&\quad + \|f'(x_k) - f'(x^*) - H(x_k)(x_k - x^*)\|_* \\
&\quad + \|f'(x_{k,+}) - f'(x_k) - H(x_k)(x_{k,+} - x_k)\|_* \\
&\overset{(76),(75)}{=} \mathcal{O}(\|x_{k,+} - x^*\|) + \mathcal{O}(\|x_k - x^*\|) + \mathcal{O}(\|x_{k,+} - x_k\|) \\
&\overset{(77)}{=} \mathcal{O}(\|x_{k,+} - x_k\|),
\end{aligned}$$

which is (74).

(ii) $\implies$ (iii): By the assumption of this item, there exists a sequence $(\varepsilon_k)_{k\in\mathbb{N}} \subset \mathbb{R}_+$ such that $\varepsilon_k \to 0$ and

$$\|f'(x_+(\lambda_k/2, x_k)) - f'(x_k) - H(x_k)(x_+(\lambda_k/2, x_k) - x_k)\|_* \leq \varepsilon_k \|x_k - x_+(\lambda_k/2, x_k)\| \,.$$

Thanks to local convexity of F and $H(x_k) \succeq 0$, we can use Lemma 6.3, which gives us that $\lambda_k \to 0$.

(iii) $\implies$ (iv): Due to the assumed PL condition at x_k and the fact that $F(x^*) \leq F(x_k)$, we can apply Lemma 3.5 and obtain (29), where we recall from (21) the notation $r_k := \|x_k - x_{k+1}\|$. This together with the second inequality in (73) at iteration $k+1$ gives for some constants $c_1, c_2 > 0$

$$\|x_{k+1} - x^*\| \overset{(73)}{\leq} c_2\|x_{k+1} - x_{k+2}\| \,, \qquad \|x_{k+1} - x_{k+2}\| \overset{(29)}{\leq} \frac{1}{\sqrt{\alpha^2\beta\mu}}\lambda_k^{1/2}\|x_k - x_{k+1}\| \,.$$

Combining these two inequalities and using the triangle inequality, we deduce

$$\|x_{k+1} - x^*\| \leq \frac{c_2}{\sqrt{\alpha^2\beta\mu}}\lambda_k^{1/2}\|x_k - x_{k+1}\| \leq \frac{c_2}{\sqrt{\alpha^2\beta\mu}}\lambda_k^{1/2}(\|x_{k+1} - x^*\| + \|x_k - x^*\|).$$

Choosing $K_4 \geq \max\{K_0, K_1, K_2, K_3\}$ sufficiently large such that $k \geq K_4$ implies $\frac{c_2}{\sqrt{\alpha^2\beta\mu}}\lambda_k^{1/2} \leq \frac{1}{2}$, we have

$$\|x_{k+1} - x^*\| \leq \frac{2c_2}{\sqrt{\alpha^2\beta\mu}}\lambda_k^{1/2}\|x_k - x^*\|, \quad \forall k \geq K_4,$$

which implies the superlinear convergence of the iterates in (iv) as $\lambda_k \to 0$ by assumption in this item.

(iii) $\implies$ (v): By the local convexity, Lemma 4.6 applies and we can use (45). Due to the PL condition and the fact that $F(x^*) \leq F(x_k)$, we can also apply Lemma 3.5 and use (28). Using additionally that $F(x^*) \leq F(x_{k+1})$, we finally obtain, recalling the notation $g_k := \|F'(x_k)\|_*$ from (21),

$$g_{k+1}^2 \overset{(45)}{\leq} \frac{\lambda_k}{\alpha}(F(x_k) - F(x_{k+1})) = \frac{\lambda_k}{\alpha}(F(x_k) - F(x^*) + F(x^*) - F(x_{k+1})) \overset{(28)}{\leq} \frac{\lambda_k}{2\alpha\mu}g_k^2.$$

Since $\lambda_k \to 0$, we clearly have superlinear convergence of the subgradient norms. Note that due to the local convexity, the above bound is tighter than the one that can be obtained from (30). $\qquad\square$

Proposition 6.6 shows that the semismoothness Assumption 4 and DM condition (74) are sufficient for superlinear convergence of Algorithm 1. We consider (74) as a kind of Dennis–Moré condition, referring to the celebrated Dennis–Moré theorem [29], which essentially gives a necessary and sufficient condition for superlinear convergence of quasi-Newton methods for smooth functions (in the nonsmooth setting, we refer to, e.g., [24, 33]). An interesting open question is to understand if (74) is also necessary to obtain superlinear convergence in our setting. We also note that a similar condition was used recently in [84, 83] to show superlinear convergence in the semismooth setting.

The next result provides a sufficient condition for (74) by showing that it is satisfied when H is continuous.

Lemma 6.7. *Let H be continuous at x^*. Let $(x_k)_{k\in\mathbb{N}}$ be the iterates of Algorithm 1 and $x_k \to x^*$ as $k \to \infty$. Let, for sufficiently large k, $H(x_k) \succeq 0$ and inequality (66) hold with $\lambda = \lambda_k/2$. Then the Dennis–Moré-type condition (74) is satisfied.*

Proof. Again, given an index $k \geq 0$, we denote by $x_{k,+} := x_+(\lambda_k/2, x_k)$ (see $(\mathrm{P_N})$). Then, by Lemma 6.2, we have $x_{k,+} \to x^*$ as $k \to \infty$. This, together with triangle inequality and (66) gives us

$$
\frac{\|(H(x_{k,+}) - H(x_k))(x_{k,+} - x^*)\|_*}{\|x_{k,+} - x_k\|}
$$

$$
= \frac{\|(H(x_{k,+}) - H(x^*) + H(x^*) - H(x_k))(x_{k,+} - x^*)\|_*}{\|x_{k,+} - x_k\|}
$$

$$
\leq \frac{(\|H(x_{k,+}) - H(x^*)\|_{\mathrm{op}} + \|H(x^*) - H(x_k)\|_{\mathrm{op}})\|x_{k,+} - x^*\|}{\|x_{k,+} - x_k\|}
$$

$$
\overset{(66)}{\leq} \frac{1 + 3c}{2}(\|H(x_{k,+}) - H(x^*)\|_{\mathrm{op}} + \|H(x^*) - H(x_k)\|_{\mathrm{op}}) \to 0,
$$

where convergence holds by continuity of H at x^* since both x_k and $x_{k,+}$ converge to x^* as $k \to \infty$. $\qquad\square$

Note that the inequality (66) is guaranteed, for example by Lemma 4.5 or Lemma 5.2 since they guarantee (65).

Remark 6.8. *It is possible to use the following condition instead of* (74)*:*

$$
\|(H(x_+(\lambda_k/2, x_k)) - H(x_k))(x_+(\lambda_k/2, x_k) - x^*\|_* = \mathcal{O}(\|x_k - x^*\|) \quad \text{as } x_k \to x^*.
$$

Yet, this requires additional technical steps using (77) *to show the implication (i) $\implies$ (ii). Also, Lemma 6.7 would still go through with additional estimates, as done in the proof of [84, Proposition 8].*

As we also showed in Proposition 6.6, the condition (75) *is sufficient for superlinear convergence as well. This condition can be interpreted as a kind of uniform semismoothness along the trajectory of Algorithm 1.*

Remark 6.9. *It is possible to not assume local convexity of F in Proposition 6.6. The main place where it is used is in the proof of Lemma 6.3 to show the inequality* (70)*, which could also be shown by using the same arguments as in Lemma 3.3 based on the model nonincrease condition* (10)*. At the same time, the assumption $H(x_k) \succeq 0$ is is crucial to show inequalities* (77) *by Lemma 6.1 and is thus crucial to show superlinear convergence.*

We finish this subsection by the main results that combine the convergence rate results of previous sections with the result of Proposition 6.6. The first result is that in the PL setting local convexity, semismoothness, and DM condition imply global superlinear convergence of Algorithm 1.

Theorem 6.10 (Global superlinear convergence)**.** *Let the assumptions of Theorem 4.4 hold. Let also $(x_k)_{k \in \mathbb{N}}$ be the iterates of Algorithm 1 with arbitrary starting point $x_0 \in \operatorname{dom} F$. Let additionally F be locally convex around x^*, and the semismoothness at x^* Assumption 4 hold, where x^* is the limit point of $(x_k)_{k \in \mathbb{N}}$. Let also either the DM condition* (74) *or condition* (75) *hold. Let $H(x_k) \succeq 0$ for sufficiently large k.*

Then, x^* is a global minimum and

$x_k \to x^*$ *superlinearly,* $\quad F(x_k) - F^* \to 0$ *superlinearly,* $\quad \|F'(x_k)\|_* \to 0$ *superlinearly as $k \to \infty$.*

Proof. Since the assumptions of Theorem 4.4 hold, we obtain $x_k \to x^*$ as $k \to \infty$ where x^* is a global minimum, and, hence, is also a local minimum. From Theorem 4.4 we also inherit Assumption 5 and the PL inequality of Assumption 3 at any x_k. Finally, inequalities (73) hold by Lemma 4.5. Thus, all the assumptions of Proposition 6.6 hold and we have that $\lambda_k \to 0$, which by Theorem 4.4 and item (v) of Proposition 6.6 prove the claimed superlinear convergence. $\qquad\square$

For convenience, let us state our result in the C^2 convex case.

Corollary 6.11 (Global superlinear convergence, C^2 case, convex setting, PL condition)**.** *Let, for problem* (P)*, Assumptions 1 to 3 hold, let f be C^2 and convex, and set $H = f''$. Let also $(x_k)_{k\in\mathbb{N}}$ be the iterates of Algorithm 1 with arbitrary starting point $x_0 \in \operatorname{dom} F$.*

Then, x^* is a global minimum and

$x_k \to x^*$ superlinearly, $F(x_k) - F^* \to 0$ superlinearly, $\|F'(x_k)\|_* \to 0$ superlinearly as $k \to \infty$.

The following result is a counterpart in our setting of classical results on local superlinear convergence of SSN, see, e.g., Theorem 2.11 in [50]. In particular, if all the assumptions are fullfilled locally around a solution and the starting point of Algorithm 1 is sufficiently close to that solution, then the algorithm converges superlinearly to that solution.

Theorem 6.12 (Local superlinear convergence)**.** *Let the assumptions of Theorem 5.1 hold. Let also $(x_k)_{k\in\mathbb{N}}$ be the iterates of Algorithm 1 with a starting point $x_0 \in \operatorname{dom} F$ which is sufficiently close to an isolated local minimum x^* in problem* (P)*. Let additionally F be locally convex around x^*, and the semismoothness at x^* Assumption 4 hold. Let also either DM condition* (74) *or condition* (75) *hold.*

Then,

$x_k \to x^*$ superlinearly, $F(x_k) - F^* \to 0$ superlinearly, $\|F'(x_k)\|_* \to 0$ superlinearly as $k \to \infty$.

Proof. Since the assumptions of Theorem 5.1 hold, we obtain $x_k \to x^*$ as $k \to \infty$ with x^* being a local minimum. From Theorem 5.1 we also inherit Assumption 5, the PL inequality of Assumption 3 at any x_k, and $H(x_k) \succeq 0$ for any k. Finally, inequalities (73) hold by Lemma 5.2. Thus, all the assumptions of Proposition 6.6 hold and we have that $\lambda_k \to 0$, which by Theorem 5.1 and item (v) of Proposition 6.6 prove the claimed superlinear convergence. $\qquad\square$

Remark 6.13 (On the assumption of $H(x) \succeq 0$)**.** *Theorem 6.10 and Theorem 6.12 assume that $H(x) \succeq 0$ to show the transition to superlinear convergence. We would like to underline that this assumption, without loss of generality, covers the setting of a more flexible assumption that*

$$H(x) \succeq \mu_H I \quad \text{and} \quad \psi \text{ is } \mu_\psi\text{-strongly-convex,}$$

i.e., $\psi - \frac{\mu_\psi}{2}\|\cdot\|^2$ is convex, where $\mu = \mu_H + \mu_\psi \geq 0$. A slightly stronger assumption with $\mu > 0$ was made, e.g., in [84]. Indeed, if $\mu_H < 0$, we can consider the transformation

$$\widetilde{\psi} := \psi - \frac{\mu_\psi}{2}\|\cdot\|^2, \qquad \widetilde{H}(x) := H(x) + \mu_\psi I.$$

This clearly implies that $\widetilde{\psi}$ is convex and, since $\mu_H + \mu_\psi \geq 0$, also $\widetilde{H}(x) \succeq 0$. Moreover, as shown in [84, Lemma 1], the algorithm with steps $(\mathrm{P_N})$ produces exactly the same iterates for the reformulated problem. Thus, to improve the readability of the paper, we used the assumption $H(x) \succeq 0$ for the results on superlinear convergence and assume throughout the paper that ψ is convex, but bear in mind that these results are valid also when $H(x)$ has negative curvature and ψ is sufficiently strongly convex. Moreover, our results also cover the opposite situation when ψ is weakly convex, i.e., $\mu_\psi < 0$, but $H(x)$ has sufficient positive curvature. These two scenarios were considered in [84] and both of them are also covered by our framework thanks to the transformation above. We underline also that the assumption $H(x) \succeq 0$ is not used to prove our global convergence rate results in Section 4.

6.4 Superlinear convergence via active manifold identification in the finite-dimensional setting

In Proposition 6.6, we established that a transition to fast superlinear convergence can occur under several conditions. A consequence of Lemma 6.7 is that the assumption $f \in C^2$ implies the Dennis–Moré condition (74) of Proposition 6.6 if $H \equiv f''$, and therefore also fast superlinear convergence of the iterates. In this subsection, we go further by showing that superlinear convergence can also occur when f is not locally C^2, but instead satisfies a combination of *partial smoothness* and *semismoothness* of the gradient. Specifically, in the finite-dimensional setting, we demonstrate that our algorithm identifies an active manifold, along which the function behaves smoothly, and that this identification enables rapid local convergence even when ∇f is not differentiable at the limit point x^*. For this purpose, we utilise the above-mentioned notion of *partial smoothness*, which originates from the influential work [61], see also [60, 16, 36]. The key idea is to show that, under a partial smoothness assumption, the iterates eventually lie on a smooth manifold along which the objective function F is C^2. This phenomenon has a rich history and forms a well-established part of variational analysis. Partial smoothness has been employed in the analysis of first-order methods [66, 65], in applications areas such as inverse problems [97] and variational inequalities [64], and more recently, it has found use in the context of Newton-type algorithms [63, 11, 51]. While we do not aim to give a comprehensive treatment of partial smoothness here, we refer the reader to the classical references [16, 62, 70] and focus only on the aspects essential for our convergence analysis.

In this section we work in the finite-dimensional setting. Let $\mathcal{H} = \mathbb{R}^n$ for some $n \in \mathbb{N}$ so that $F \colon \mathbb{R}^n \to \overline{\mathbb{R}}$. We begin by recalling the definition of a C^2 manifold in the context of embedded C^2 submanifolds relevant to our analysis, see [14, Definition 2.1.1].

Definition 6.14 (C^2-submanifold). Let $d \leq n$. A subset $\mathcal{M} \subset \mathbb{R}^n$ is called a d-dimensional C^2-*submanifold* of $\mathbb{R}^n$ if for every $x \in \mathcal{M}$ there exists an open neighbourhood $U_x \subset \mathbb{R}^n$ containing x and a C^2 map $\psi_x \colon U_x \to \mathbb{R}^n$ such that $\psi_x(U_x) \subset \mathbb{R}^n$ is open, ψ is a C^2–diffeomorphism onto its image and

$$\psi_x(U_x \cap \mathcal{M}) = \psi_x(U_x) \cap E_d$$

where E_d is the d-dimensional linear subspace defined as $E_d := \{x \in \mathbb{R}^n \colon x_{d+1} = \ldots = x_n = 0\}$.

We recall the following characterisation (see [14, Theorem 2.1.2] and the proof) of a C^2–submanifold which will be important in the forthcoming. The set $\mathcal{M} \subset \mathbb{R}^n$ is a d-dimensional C^2-submanifold of $\mathbb{R}^n$ if and only if for every $x \in \mathcal{M}$ there is an open subset $\Omega_x \subset \mathbb{R}^d$ containing $0 \in \mathbb{R}^d$ and an open neighbourhood $U \subset \mathbb{R}^n$ of x in the topology of $\mathbb{R}^n$ and a C^2 map

$$\phi_x \colon \Omega_x \to \mathbb{R}^n \tag{78}$$

that satisfies the following properties:

(i) $\phi_x \colon \Omega_x \to \mathbb{R}^n$ is C^2, $\phi'(0)$ is injective and $\phi_x(0) = x$ (thus, ϕ_x is a C^2 immersion around 0),

(ii) $\phi_x \colon \Omega_x \to \phi_x(\Omega_x) = U \cap \mathcal{M}$ is a homeomorphism with respect to the canonical subspace topology on $\mathcal{M}$,

(iii) ϕ_x can be chosen (after translation to ensure that $\psi_x(x) = 0$) in the explicit form

$$\phi_x = \psi_x^{-1} \circ i_d \colon \underbrace{i_d^{-1}(\psi_x(U_x \cap \mathcal{M}))}_{:=\Omega_x} \to U_x \cap \mathcal{M}, \tag{79}$$

where $\psi_x\colon U_x \to \psi_x(U_x) \subset \mathbb{R}^n$ is the C^2–diffeomorphism in Definition 6.14 and $i_d\colon \mathbb{R}^d \to \mathbb{R}^n$ denotes the canonical embedding defined by $i_d(x) = (x,0) \in E_d \subset \mathbb{R}^n$. Note that (79) makes sense as by Definition 6.14 we have $\psi_x(U_x \cap \mathcal{M}) \subset E_d$ and therefore the inverse $i_d^{-1}\colon E_d \to \mathbb{R}^d$ is well defined in (79).

The collection $(\phi_x(\Omega_x), \phi_x^{-1})_{x\in\mathcal{M}}$ forms an atlas of $\mathcal{M}$ in the usual sense of differential geometry, cf. the proof of [14, Theorem 2.2.10.1]. We make the following assumption for the subsequent two results.

Assumption 6. *Assume that*

    (i) The set $\mathcal{M} \subset \mathbb{R}^n$ is a C^2-submanifold,

    (ii) $\nabla f|_{\mathcal{M}}$ is of class C^1,

    (iii) There exists an $R > 0$ such that ∇f is Newton differentiable on $B_R(x^*)$ with Newton derivative H and H is uniformly bounded on $B_R(x^*)$.

Assumption 6 (iii) essentially asks for Assumption 4 to hold not only at x^* but on $B_R(x^*)$, and for the Newton derivative to be bounded there in the sense of (5). Now, due to the first item of this assumption, for every $x \in \mathcal{M}$, if $\phi_x\colon \Omega_x \to \mathbb{R}^n$ is a C^2 map of the form (78) mapping Ω_x homeomorphically onto $\phi_x(\Omega_x) \subset \mathcal{M}$, then the composition

$$\nabla f \circ \phi_x\colon \Omega_x \to \mathbb{R}^n$$

is of class C^1.

The following lemma establishes a form of *uniform semismoothness*, which will be instrumental in deriving the superlinear convergence of the algorithm via Proposition 6.6.

Lemma 6.15. *Assume Assumption 6. Consider the local C^2-immersion*

$$\phi_{x^*}\colon \Omega_{x^*} \subset \mathbb{R}^d \to \mathcal{M} \subset \mathbb{R}^n$$

as defined in (79). Then, there exists a radius $r > 0$ such that $B_r(0) \subset \Omega_{x^*}$ and for every $\varepsilon > 0$, there exists $\delta > 0$ satisfying

$$\|\nabla f(\phi_{x^*}(w)) - \nabla f(\phi_{x^*}(z)) - H(\phi_{x^*}(z))(\phi_{x^*}(w) - \phi_{x^*}(z))\|_* \leq \varepsilon\|\phi_{x^*}(w) - \phi_{x^*}(z)\|$$

for all $w, z \in B_r(0)$ with $\|w - z\| \leq \delta$.

Proof. Without loss of generality, with U_{x^*} as in Definition 6.14, we may assume that

$$B_R(x^*) \subset U_{x^*},$$

otherwise R can be reduced if necessary (this is possible since U_{x^*} is open and contains x^*). To show the statement of the lemma, we fix an arbitrary $\varepsilon > 0$. Since Ω_{x^*} is an open neighbourhood of $0 \in \mathbb{R}^d$ and ϕ_{x^*} is continuous, there exists $r > 0$ such that

$$B_r(0) \subset \Omega_{x^*} \qquad \text{and} \qquad \phi_{x^*}(B_r(0)) \subset B_R(x^*) \subset U_{x^*}.$$

By Assumption 6 (ii), the gradient ∇f is of class C^1 on the manifold $\mathcal{M}$, hence we have that $\nabla f \circ \phi_{x^*}\colon \Omega_{x^*} \subset \mathbb{R}^d \to \mathbb{R}^n$ is of class C^1 with Fréchet derivative $(\nabla f \circ \phi_{x^*})'\colon \Omega_{x^*} \to \mathcal{L}(\mathbb{R}^d, \mathbb{R}^n)$

and also by the characterisation above, the mapping $\phi_{x^*} \colon \Omega_{x^*} \to \mathbb{R}^n$ is C^2 with Fréchet derivative $\phi'_{x^*} \colon \Omega_{x^*} \to \mathcal{L}(\mathbb{R}^d, \mathbb{R}^n)$. In particular the derivatives ϕ'_{x^*} and $(\nabla f \circ \phi_{x^*})'$ are continuous over $B_r(0)$ and therefore by the Heine–Cantor theorem, even uniformly continuous. Hence, as in Theorem 6.5, we find a $\delta > 0$ such that $w, z \in B_r(0)$ and $\|w - z\| \leq \delta$ implies

$$\|(\nabla f \circ \phi_{x^*})'(z) - (\nabla f \circ \phi_{x^*})'(w)\|_* \leq \varepsilon \qquad \text{and} \qquad \|\phi'_{x^*}(z) - \phi'_{x^*}(w)\| \leq \varepsilon.$$

By the fundamental theorem of calculus we conclude that

$$\|\nabla f(\phi_{x^*}(w)) - \nabla f(\phi_{x^*}(z)) - (\nabla f \circ \phi_{x^*})'(z)(w - z)\|_* \leq \varepsilon \|w - z\|, \qquad (80)$$

$$\|\phi_{x^*}(w) - \phi_{x^*}(z) - \phi'_{x^*}(z)(w - z)\| \leq \varepsilon \|w - z\|. \qquad (81)$$

Using the Newton differentiability of ∇f,

$$(\nabla f \circ \phi_{x^*})'(v) = H(\phi_{x^*}(v))\phi'_{x^*}(v) \quad \text{for all } v \in \operatorname{int}(B_r(0)),$$

by the chain rule[3] for Newton differentiable functions [96, Proposition 3.8], which is applicable, since for given $v \in \operatorname{int}(B_r(0))$ the set $\phi_{x^*}(\operatorname{int}(B_r(0)))$ is an open environment of $\phi_{x^*}(v)$ contained inside $B_R(x^*)$ where the Newton derivative $H(\cdot)$ is uniformly bounded by the constant $M > 0$, and also $\phi'_{x^*}(\cdot)$ is uniformly bounded on $\operatorname{int}(B_r(0))$ (by its continuity on the compact set $B_r(0)$) which is an open neighbourhood of $v \in \operatorname{int}(B_r(0))$ by choice. Thus we have

$$\|\nabla f(\phi_{x^*}(w)) - \nabla f(\phi_{x^*}(z)) - H(\phi_{x^*}(z))\phi'_{x^*}(z)(w - z)\|_* \leq \varepsilon \|w - z\|$$

for $w, z \in \operatorname{int}(B_r(0))$ with $\|w - z\| \leq \delta$. Consequently we deduce by (80), (81), and the triangle inequality,

$$\begin{aligned}
\|\nabla f(\phi_{x^*}(w)) &- \nabla f(\phi_{x^*}(z)) - H(\phi_{x^*}(z))(\phi_{x^*}(w) - \phi_{x^*}(z))\|_* \\
&\leq \|\nabla f(\phi_{x^*}(w)) - \nabla f(\phi_{x^*}(z)) - H(\phi_{x^*}(z))\phi'_{x^*}(z)(w - z)\|_* \\
&\quad + \|H(\phi_{x^*}(z))\phi'_{x^*}(z)(w - z) - H(\phi_{x^*}(z))(\phi_{x^*}(w) - \phi_{x^*}(z))\|_* \\
&\leq \varepsilon \|w - z\| + \|H(\phi_{x^*}(z))\|_{\operatorname{op}} \|\phi_{x^*}(w) - \phi_{x^*}(z) - \phi'_{x^*}(z)(w - z)\| \\
&\leq (1 + M)\,\varepsilon \|w - z\|.
\end{aligned} \qquad (82)$$

Recall that we used the explicit form (79), that ensures $\phi_{x^*} = \psi_{x^*}^{-1} \circ i_d \colon \Omega_{x^*} \to \mathbb{R}^n$ with the C^2–diffeomorphism ψ_{x^*} from Definition 6.14. In particular, ψ_{x^*} has a bounded first derivative on the compact set $\overline{\phi_{x^*}(\operatorname{int}(B_r(0)))} \subset U_x \subset \mathbb{R}^n$, and is therefore Lipschitz continuous. We conclude that the map $\phi_{x^*}^{-1} = i_d^{-1} \circ \psi_{x^*} \colon \phi_{x^*}(\operatorname{int}(B_r(0))) \subset \mathcal{M} \to \mathbb{R}^d$ is Lipschitz continuous, as the map $i_d^{-1} \colon E_d \subset \mathbb{R}^n \to \mathbb{R}^d$ is 1-Lipschitz (note again that $i_d^{-1} \circ \psi_{x^*}$ is well defined in this setting as $\psi_{x^*}(x) \subset E_d$ for $x \in \phi_{x^*}(\operatorname{int}(B_r(0))) \subset \mathcal{M}$). In particular we find $L > 0$ such that

$$\|w - z\| = \|\phi_{x^*}^{-1}(\phi_{x^*}(w)) - \phi_{x^*}^{-1}(\phi_{x^*}(z))\| \leq L \|\phi_{x^*}(w) - \phi_{x^*}(z)\|$$

for all $w, z \in \operatorname{int}(B_r(0))$. Eventually we deduce from the previous observations and (82), that

$$\|\nabla f(\phi_{x^*}(w)) - \nabla f(\phi_{x^*}(z)) - H(\phi_{x^*}(z))(\phi_{x^*}(w) - \phi_{x^*}(z))\|_* \leq (1 + M)\,L\varepsilon \|\phi_{x^*}(w) - \phi_{x^*}(z)\|.$$

By rescaling $\varepsilon > 0$ and shrinking r (to obtain the result on the closed ball), the assertion follows. $\quad\square$

---

[3]To be more precise, the left-hand side above is a Frechét derivative, which is equal to the Newton derivative as the former exists, then we apply the chain rule for Newton differentiable functions to get the right-hand side.

Proposition 6.16. *Let $(x_k)_{k \in \mathbb{N}}$ be the iterates of Algorithm 1 and $x_k \to x^*$ as $k \to \infty$. Let also $H(x_k) \succeq 0$ for sufficiently large k and suppose that $x_k \in \mathcal{M}$ and $x_+(\lambda_k/2, x_k) \in \mathcal{M}$ for sufficiently large k. Assume also Assumption 6. Then, we have that*

$$\frac{\|\nabla f(x_+(\lambda_k/2, x_k)) - \nabla f(x_k) - H(x_k)(x_+(\lambda_k/2, x_k) - x_k)\|_*}{\|x_+(\lambda_k/2, x_k) - x_k\|} \to 0 \quad \text{as } k \to \infty \quad (83)$$

and hence, $\lambda_k \to 0$ as $k \to \infty$.

Proof. We follow the argument in Theorem 6.5, denoting again $x_{k,+} := x_+(\lambda_k/2, x_k)$. We first show that (83) holds. For this purpose fix an arbitrary $\varepsilon > 0$ and note that, due to Lemma 6.2 the convergence of $x_k \to x^*$ also implies $x_{k,+} \to x^*$. Now consider the limit point x^* and the associated immersion $\phi_{x^*} \colon \Omega_{x^*} \to \mathbb{R}^n$ from (79). Employing Lemma 6.15, we find a $\delta > 0$ and radius $r > 0$ such that $w, z \in B_r(0)$ and $\|z - w\| \leq \delta$ implies

$$\|\nabla f(\phi_{x^*}(w)) - \nabla f(\phi_{x^*}(z)) - H(\phi_{x^*}(z))(\phi_{x^*}(w) - \phi_{x^*}(z))\|_* \leq \varepsilon \|\phi_{x^*}(w) - \phi_{x^*}(z)\|. \quad (84)$$

Recall that $\phi_{x^*} \colon \Omega_{x^*} \to \phi_{x^*}(\Omega_{x^*})$ is a homeomorphism, and consequently $\phi_{x^*}(B_r(0)) \subset \mathcal{M}$ is a neighbourhood (in the topology of $\mathcal{M}$) of $\phi_{x^*}(0) = x^*$. From this and the facts that $x_{k,+} \to x^*$ and $x_k, x_{k,+} \in \mathcal{M}$ for $k \geq k_0$ for some $k_0 \in \mathbb{N}$, we infer that there is a possibly larger index $k_1 \geq k_0$ such that $x_k, x_{k,+} \in \phi_{x^*}(B_r(0))$ for all $k \geq k_1$. For those k, by injectivity of ϕ_{x^*}, we may set $w_k := \phi_{x^*}^{-1}(x_k)$ and $w_{k,+} := \phi_{x^*}^{-1}(x_{k,+})$. Then by construction, we have $w_k, w_{k,+} \in B_r(0)$ and by continuity of $\phi_{x^*}^{-1}$ we deduce $w_k \to 0$ and $w_{k,+} \to 0$. Thus, by possibly enlarging k_1, we also have $\|w_k - w_{k,+}\| \leq \delta$ for all $k \geq k_1$. These observations combined with (84) eventually lead to

$$\begin{aligned}
&\|\nabla f(x_{k,+}) - \nabla f(x_k) - H(x_k)(x_{k,+} - x_k)\|_* \\
&= \|\nabla f(\phi_{x^*}(w_{k,+})) - \nabla f(\phi_{x^*}(w_k)) - H(\phi_{x^*}(w_k))(\phi_{x^*}(w_{k,+}) - \phi_{x^*}(w_k))\|_* \\
&\leq \varepsilon \|\phi_{x^*}(w_{k,+}) - \phi_{x^*}(w_k)\| = \varepsilon \|x_{k,+} - x_k\|.
\end{aligned}$$

for all $k \geq k_1$. As $\varepsilon > 0$ was arbitrary, we infer (83). Now we can apply Lemma 6.3 with ε_k set as the left-hand side of (83) to get $\lambda_k \to 0$ as $k \to \infty$. $\square$

To apply this result and ensure superlinear convergence, it is essential that all iterates as well as the limit point eventually lie on a C^2 manifold — commonly referred to as the *active manifold*. This is precisely where the concept of *partial smoothness* becomes central. We begin with several foundational definitions, following [60]. Let C be a non-empty convex set. The *subspace parallel to C* is defined as

$$\operatorname{par} C := \operatorname{aff} C - x \quad \text{for any } x \in C,$$

where $\operatorname{aff} C$ denotes the affine hull (or affine span) of C. The expression $\operatorname{par} C$ yields a linear subspace that is parallel to the affine set $\operatorname{aff} C$, and is independent of the particular choice of $x \in C$.

Definition 6.17 (Partial smoothness). We say that a function $g \colon \mathbb{R}^n \to \overline{\mathbb{R}}$ is C^2-*partly smooth at $\bar{x}$ relative to $\mathcal{S}$* if $\mathcal{S}$ is a C^2-smooth manifold around $\bar{x}$ and

   (i) (restricted smoothness) $g|_{\mathcal{S}}$ is C^2 around $\bar{x}$,

   (ii) (regularity) in a neighbourhood of $\bar{x}$ in $\mathcal{S}$, g is subdifferentially regular and has a subgradient,

   (iii) (normal sharpness) $N_{\mathcal{S}}(\bar{x}) = \operatorname{par} \partial g(\bar{x})$ where $N_{\mathcal{S}}(\bar{x})$ means the normal space to $\mathcal{S}$ at $\bar{x}$,

   (iv) (subgradient continuity) the map ∂g is continuous at $\bar{x}$ relative to $\mathcal{S}$.

The normal sharpness condition essentially encodes that the function g, which is C^2-smooth around $\bar{x}$ on the manifold, must change drastically along directions that are normal to $\mathcal{S}$. Indeed, another characterisation [70, Definition 1.3] of the normal sharpness condition is that the function $t \mapsto g(\bar{x} + td)$ is *not* differentiable at $t = 0$ for any $d \in N_{\mathcal{S}}(\bar{x})$. An example of a partially smooth function is $g(x) = \|x\|_{\ell^1}$, which is partially smooth at $\bar{x}$ relative to the manifold $\mathcal{S} = \{x \in \mathbb{R}^n : I_x \subseteq I_{\bar{x}}\}$ where the notation $I_x := \{i \in \mathbb{N} : x_i \neq 0\}$. For more examples we refer to e.g. [65, §5.1] and [70].

We first show that our algorithm LeAP-SSN identifies the active manifold in finite time. For this purpose and for later use, we introduce the next assumption.

Assumption 7. *Assume that*

(i) *F is C^2-partially smooth at a stationary point x^* relative to the C^2-smooth manifold $\mathcal{M} \subset \mathbb{R}^n$,*

(ii) *The following* non-degeneracy *condition holds:*

$$0 \in \operatorname{ri} \partial F(x^*),$$

where ri *means the relative interior.*

The non-degeneracy condition in Assumption 7 (ii) is also referred to as *strict complementarity*. These assumptions are standard[4] in the literature, see e.g. [45], [70, Assumption 1.5]. In addition, Proposition 6.18 suggests designing an algorithm that switches once finite identification has taken place to utilise the smooth structure, as considered for example in [59].

Proposition 6.18 (Finite time identification)**.** *Let $(x_k)_{k \in \mathbb{N}}$ be the iterates of Algorithm 1, $x_k \to x^*$ and $F(x_k) \to F(x^*)$ as $k \to \infty$, where x^* is a global minimum. Assume also Assumption 7. Then, we have that $x_k \in \mathcal{M}$ for sufficiently large k.*

If, in addition, $H(x_k) \succeq 0$ for sufficiently large k, the PL condition Assumption 3 holds, ∇f is Lipschitz on $\operatorname{int}(B_R(x^*))$ with constant L_1, and H is uniformly bounded (in the sense of (5)) on $B_R(x^*)$, we also have $x_+(\lambda_k/2, x_k) \in \mathcal{M}$ for sufficiently large k.

Proof. First observe that F is prox-regular since it is the sum of a continuously Fréchet differentiable function and a convex function, see [90, Exercise 13.35]. From Theorem 4.1, we know that $F'(x_k) \to 0$ as $k \to \infty$. This implies that $\operatorname{dist}(0, \partial F(x_k)) \leq \|F'(x_k)\|_* \to 0$, and applying [60, Theorem 4.10] we get that $x_k \in \mathcal{M}$ for large enough k.

To prove the second claim, let us for simplicity denote $x_{k,+} := x_+(\lambda_k/2, x_k)$. By Lemma 6.2 (which is applicable since $H(x_k) \succeq 0$ for sufficiently large k), we have that $x_{k,+} \to x^*$. This means that there exist K_1 such that for all $k \geq K_1$, we have $x_k, x_{k,+} \in \operatorname{int}(B_R(x^*))$. Using the Lipschitzness of ∇f on $\operatorname{int}(B_R(x^*))$, the boundedness of H on $B_R(x^*)$, and boundedness of λ_k by Lemma 3.3, we obtain

$$\operatorname{dist}(0, \partial F(x_{k,+})) \leq \|F'(x_{k,+})\|_* \overset{(12),(20)}{\leq} (L_1 + M)\|x_{k,+} - x_k\| + \frac{\bar{\lambda}}{2}\|x_{k,+} - x_k\| \to 0 \text{ as } k \to \infty.$$

Using this and the PL condition Assumption 3, we also obtain $F(x_{k,+}) \to F(x^*)$. Applying again [60, Theorem 4.10] we get that $x_+(\lambda_k/2, x_k) \in \mathcal{M}$ for large enough k. $\square$

---

[4]One could explore the possibility to weaken the strict complementarity condition, see [40], however a full examination and treatment of these topics is far beyond the scope of this work.

We are now in a position to state and prove the main theorem of this section, which is the sister result of Theorem 6.10.

Theorem 6.19 (Global superlinear convergence under partial smoothness)**.** *Let, for problem* (P)*, Assumptions 1 to 3 and 5 hold. Let also* $(x_k)_{k \in \mathbb{N}}$ *be the iterates of Algorithm 1 with arbitrary starting point* $x_0 \in \mathrm{dom}\, F$*. Let additionally Assumptions 6 and 7 hold. Assume also that* F *is locally convex around* x^* *and that* $H(x_k) \succeq 0$ *for sufficiently large* k*.*

Then, x^* *is a global minimum and*

$x_k \to x^*$ *superlinearly,* $F(x_k) - F^* \to 0$ *superlinearly,* $\|F'(x_k)\|_* \to 0$ *superlinearly as* $k \to \infty$*.*

Proof. Note that we are in the regime of Theorem 4.4. Also, thanks to Assumption 6 (iii), by [17, Proposition 2.3] we have that ∇f is Lipschitz on $\mathrm{int}(B_R(x^*))$, so the full result of Proposition 6.18 is available. The theorem now follows from collecting the implications of Theorem 4.4 (which gives $x_k \to x^*$, $F(x_k) \to F(x^*)$, and that x^* is a global minimum), and in turn Proposition 6.18 (which gives $x_k, x_+(\lambda_k/2, x_k) \in \mathcal{M}$ for sufficiently large k) and Proposition 6.16 (which gives $\lambda_k \to 0$) and applying Proposition 6.6. □

Remark 6.20. *In particular, under the conditions of the above result, we see that the DM condition* (74) *is satisfied. Indeed, we proved in Proposition 6.16 the condition* (75)*, which is equivalent to the DM condition* (74) *as shown in Proposition 6.6.*

For the result of Theorem 6.19, the fact that we need $\nabla f|_{\mathcal{M}}$ to be C^1 from Assumption 6 (ii) is inconvenient since in combination with the normal sharpness condition in Assumption 7 (i), it means that we require ψ to be non-zero and to provide the active manifold $\mathcal{M}$ on which we need ∇f to be C^1. This is because we always assume $f \in C^1$ and hence f cannot satisfy the normal sharpness condition (unless $N_{\mathcal{M}}(x^*) = \emptyset$) on a submanifold as f is differentiable everywhere on its domain. However, our algorithm identifies the manifold irrespective of whether $\nabla f|_{\mathcal{M}}$ is C^1 or not. Nevertheless, we hope future work can excise this assumption.

6.4.1 Example

In order to build intuition and gain a geometric understanding, we discuss the results of this subsection through a small academic example, postponing more realistic applications to future work. For this purpose we define $F \colon \mathbb{R}^2 \to \mathbb{R}$ via $F(x) = f(x) + \psi(x)$ where for $x = (x_1, x_2)$

$$f(x_1, x_2) = \|x\|^2 + \max(0, x_1)^2, \quad \psi(x_1, x_2) = |x_1|.$$

We directly observe that $f \colon \mathbb{R}^2 \to \mathbb{R}$ is globally $C^{1,1}(\mathbb{R}^2, \mathbb{R})$, with gradient

$$\nabla f(x_1, x_2) = \begin{bmatrix} 2x_1 + 2\max(0, x_1) \\ 2x_2 \end{bmatrix}.$$

Clearly, the function $\nabla f \colon \mathbb{R}^2 \to \mathbb{R}^2$ is Newton differentiable with Newton derivative $H(x) \in \mathbb{R}^{2 \times 2}$ defined by

$$H(x) = \begin{bmatrix} 2 + 2\chi_{[0,+\infty)}(x_1) & 0 \\ 0 & 2 \end{bmatrix},$$

where the notation $\chi_A \colon \mathcal{H} \to \mathbb{R}$ denotes the characteristic function on the set A, i.e., $\chi_A(x) = 1$ if $x \in A$ and $\chi_A(x) = 0$ if $x \notin A$. We have $H(x) \succeq 0$ and $H(x)$ is globally bounded in the
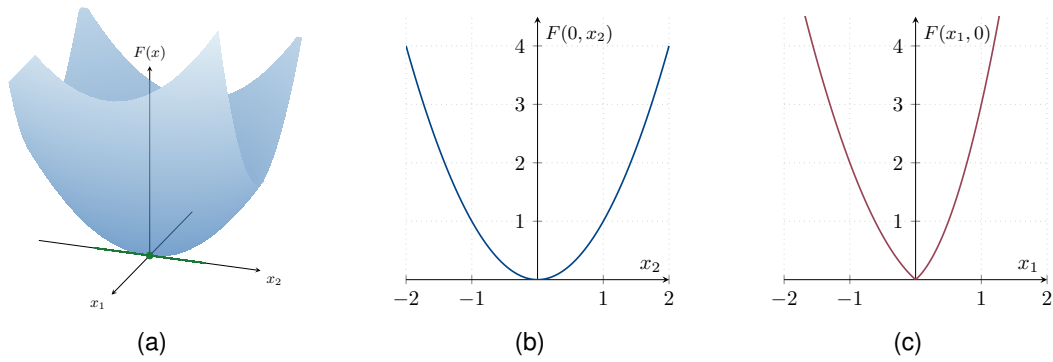
Figure 1: Depiction of the function $F : \mathbb{R}^2 \to \mathbb{R}$ as described in the example in Section 6.4.1. In panel (a), the manifold $\mathcal{M}$ is shown in green along with the minimiser $x^* = (0, 0)$. Panel (b) illustrates the C^2 behaviour of F when restricted to the manifold. Panel (c) highlights the sharpness of the function along the normal direction $N_{\mathcal{M}}(x^*)$.

sense of (5). We further observe that the function F is strongly convex as the sum of a convex and the strongly convex function $\| \cdot \|^2$ on $\mathbb{R}^2$. Strong convexity implies that it satisfies the global PL condition Assumption 3 and has the global unique minimiser $x^* = (0, 0)$. Note that it is not possible to find a ball $B_R(x^*)$ such that f is C^2 on $B_R(x^*)$ and therefore the results of Theorem 6.5 and Lemma 6.7 (which is used to verify the Dennis–Moré condition, in order to apply Proposition 6.6) are not applicable. Now consider the subdifferential of F at x^*: we directly obtain by the sum rule

$$\partial F(x^*) = [-1, 1] \times \{0\}. \tag{85}$$

Given the global solution x^*, we may set $\mathcal{M} := \{0\} \times (-1, 1)$. Then F is C^2-partly smooth at x^* relative to $\mathcal{M}$. To see this, we argue as follows. First, Definition 6.17 (i) is immediate as we see that F is C^2 and ∇f is C^1 on $\mathcal{M}$:

$$F|_{\mathcal{M}}(x) = x_2^2, \quad \text{and} \quad \nabla f|_{\mathcal{M}}(x_1, x_2) = \begin{bmatrix} 0 \\ 2x_2 \end{bmatrix}.$$

Moreover, F is convex and therefore subdifferentially regular, i.e., Definition 6.17 (ii) is satisfied. Moreover $N_{\mathcal{M}}(x^*) = \mathbb{R} \times \{0\} = \operatorname{par} \partial F(x^*)$ according to (85), giving Definition 6.17 (iii). To show continuity of ∂F at x^* relative to $\mathcal{M}$, according to [61, Note 2.9 (b)] it is enough to show that for any sequence $x_k \in \mathcal{M}$ with $x_k \to x^*$ and any $v^* \in \partial F(x^*)$ there is a sequence of subgradients $v_k \in \partial F(x_k)$ with $v_k \to v^*$ (here we note that again by the sum rule, for $x = (0, x_2)$ belonging to $\mathcal{M}$, $\partial F(x) = [-1, 1] \times \{2x_2\}$). This is easy to prove: for given $x_k = (0, x_{k,2}) \in \mathcal{M}$ and $v^* = (v_1^*, 0)$ for some $v_1^* \in [-1, 1]$ (every $v^* \in \partial F(x^*)$ is of this form by construction), we simply set $v_k = (v_1^*, 2x_{k,2}) \in \partial F(x_k)$ and deduce $v_k \to v^*$ as $x_k \to x^*$. Consequently also Definition 6.17 (iv) holds. Eventually in order to apply Theorem 6.19, we need to show non-degeneracy, i.e. $0 \in \operatorname{ri} \partial F(x^*)$. Recalling $\partial F(x^*) = [-1, 1] \times \{0\}$ and $\operatorname{aff}(\partial F(x^*)) = \mathbb{R} \times \{0\}$ and using the definition of the relative interior,

$$\operatorname{ri} \partial F(x^*) = \{v \in [-1, 1] \times \{0\} : \exists \varepsilon > 0 \text{ s.t. } \operatorname{int}(B_\varepsilon(v)) \cap \operatorname{aff}(\partial F(x^*)) \subseteq \partial F(x^*)\}$$
$$= (-1, 1) \times \{0\}.$$

Hence $0 \in \operatorname{ri} \partial F(x^*)$ and we may apply Theorem 6.19 to obtain superlinear convergence of our method on this problem.

7 Numerical Experiments

Data availability. Software to generate all the tables and figures in this section may be found at [3]. The specific version used in this paper is archived on Zenodo [4]. The implementation is written in Python. The first two examples utilise the Firedrake library [48] to implement the finite element discretisations as well as PETSc [9] for its sparse linear algebra routines. The final example adapts code found in the `super-newton` library [30].

In the first two examples, we assume $\Omega \subset \mathbb{R}^2$ to be a bounded and convex domain. As is standard, we denote the Lebesgue spaces by $L^p(\Omega)$, $p \in [1, \infty]$ and write $W^{s,p}(\Omega)$, $s \geq 0$, $p \in [1, \infty]$ to denote the Sobolev spaces [1]. We define the Sobolev Hilbert spaces as $H^s(\Omega) := W^{s,2}(\Omega)$. We denote the space of compactly supported smooth functions by $C_c^\infty(\Omega)$. For a vector-valued function $v \colon \Omega \to \mathbb{R}^2$ where each component lives in the space $H^s(\Omega)$ (or $C_c^\infty(\Omega)$), we say that $v \in H^s(\Omega; \mathbb{R}^2)$ (respectively $C_c^\infty(\Omega; \mathbb{R}^2)$).

7.1 A contact problem via a Moreau–Yosida regularisation

We consider a Signorini contact problem [41] where a partially clamped linearly elastic beam, deformed by gravity, must lie above a nontrivial obstacle. In particular we fix a rectangular domain $\Omega = (0,5) \times (0,1)$ and seek a displacement $u \colon \Omega \to \mathbb{R}^2$ that satisfies

$$\min_{u \in H^1(\Omega; \mathbb{R}^2)} \int_\Omega \mu_L |\varepsilon(u)|^2 + \frac{\lambda_L}{2} |\operatorname{div}(u)|^2 - g \cdot u \, \mathrm{d}x, \tag{86a}$$

$$u = (0,0)^\top \text{ on } \Gamma_D = \{(x,y) \in \partial\Omega : x = 0 \text{ or } x = 5\}, \tag{86b}$$

$$(u - \varphi) \cdot n_u \geq 0 \text{ on } \Gamma_S = \{(x,y) \in \partial\Omega : y = 0\}. \tag{86c}$$

Here $\mu_L = 76.92$ and $\lambda_L = 115.38$ are the Lamé coefficients, $\varepsilon(u) = (\nabla u + (\nabla u)^\top)/2$ denotes the symmetrised gradient, $g = (0, -10)^\top$ is the body force due to gravity, and the Dirichlet boundary condition, (86b), models an elastic beam that is clamped on both the left- and right-hand side. Inequality (86c) models the Signorini (non-penetration) condition at the bottom of the beam enforcing that the beam must lie above a rigid body with the shape $(0,5) \times (0, 1/4)$ under the deformation

$$\varphi(x,y) = (0, -1/2 - 2\sin(\pi x)/5)^\top. \tag{86d}$$

Finally n_u denotes the outward normal of the *deformed* elastic body as given by

$$n_u = \frac{(I + \nabla u)^{-\top} \tilde{n}}{|(I + \nabla u)^{-\top} \tilde{n}|} \text{ where } \tilde{n} = (0, -1)^\top \text{ and } I \text{ is the } 2 \times 2 \text{ identity matrix.} \tag{86e}$$

The pointwise constraint can be difficult to enforce directly. Hence, the constraint is relaxed and we include a nonsmooth Moreau–Yosida regularisation to the formulation. For a given penalisation parameter $\gamma > 0$, the relaxed problem seeks $u_\gamma \colon \Omega \to \mathbb{R}^2$ satisfying

$$u_\gamma = \underset{v \in H^1(\Omega; \mathbb{R}^2)}{\arg\min} \ F(v) \text{ subject to } v = (0,0)^\top \text{ on } \Gamma_D,$$

$$F(v) = \int_\Omega \mu_L |\varepsilon(v)|^2 + \frac{\lambda_L}{2} |\operatorname{div}(v)|^2 - g \cdot v \, \mathrm{d}x + \frac{\gamma}{2} \int_{\Gamma_S} \max(0, (v - \varphi) \cdot n_v)^2 \, \mathrm{d}s. \tag{87}$$

Thus we fix $\psi \equiv 0$ in (P). We mesh the domain into 24,000 simplices and discretise (87) with a continuous piecewise linear finite element method for u resulting in 24,682 degrees of freedom [37, Ch. 7.4]. A discretised approximate solution, when $\gamma = 10^6$, is plotted in Figure 2a.

In order to minimise the violation of the non-penetration condition, we wish to consider a large penalisation parameter γ. However, the larger the penalty, the more difficult the problem becomes to solve. In particular a Newton method, even coupled with a linesearch, will struggle to converge within 200 iterations. This effect is demonstrated in Figure 2b where we compare Algorithm 1 against a Newton solver without a linesearch, with an ℓ^2-minimising linesearch [18, Alg. 2] (ℓ^2-Newton) and with a backtracking linesearch [92, Ch. 6.3] (backtracking Newton). From a zero initial guess, we found that even for γ-values as low as 10^3, the Newton solver failed to converge. The backtracking and ℓ^2-Newton solvers were more robust, achieving convergence up to $\gamma = 10^4$ and $\gamma = 10^3$, respectively, but failing once $\gamma \geq 10^5$. By contrast, Algorithm 1 successfully converged up until $\gamma = 10^6$.

(a) Approximate solution when $\gamma = 10^6$.

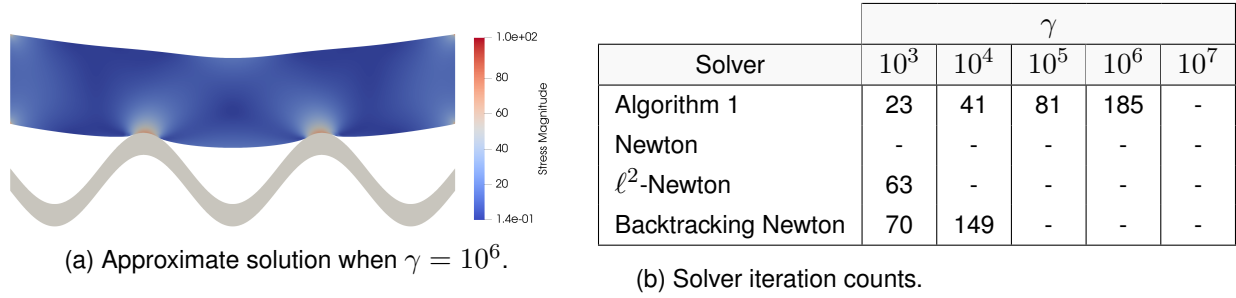|	γ				
Solver	10^3	10^4	10^5	10^6	10^7
Algorithm 1	23	41	81	185	-
Newton	-	-	-	-	-
ℓ^2-Newton	63	-	-	-	-
Backtracking Newton	70	149	-	-	-

(b) Solver iteration counts.

Figure 2: Results for the contact problem of Section 7.1. (Left) The discretised solution to the relaxed Signorini problem (87) with $\gamma = 10^6$. The colourmap indicates the magnitude of the stress in the elastic beam and the obstacle is plotted in grey. (Right) The number of linear system solves (including those leading to rejected updates in Algorithm 1) to reach a residual vector ℓ^2-norm ($\|F'\|_{\ell^2}$) of 10^{-8} for increasing values of γ. The solvers are initialised at a zero initial guess $u = (0, 0)^\top$ for each penalty parameter γ. A dashed line indicates a failure to converge within 200 linear system solves. We choose the parameters $\alpha = 1/2$, $\beta = 1/4$ and $\Lambda_0 = 1$ for Algorithm 1 and use the H^1-norm, $\|u\|^2 = \|u\|^2_{L^2(\Omega)} + \|\nabla u\|^2_{L^2(\Omega)}$, for the primal space norm and its corresponding discrete dual norm for $\|\cdot\|_*$.

In Figure 3 we plot the reduction in the objective against the iteration number k. For $\gamma = 10^6$, we include the theoretical global convergence rate predicted by (50), where $\alpha = 1/2$ and we estimate D_0 by $D_0 \approx \max_i \{\|u^* - u_i\|_{H^1(\Omega)}\} = \|u^*\|_{H^1(\Omega)}$. This gives us the estimate $F(u_k) - F^* \leq 4k^{-1}(\Pi_{j=0}^{k-1}\lambda_j)^{1/k}\|u^*\|^2_{H^1(\Omega)}$. For this example, we see that the algorithm performs orders of magnitude better than what is theoretically guaranteed. We also observe eventual local superlinear convergence for all values of γ.

7.2 Total variation image restoration

In this example we consider an image restoration problem. The domain is the unit square $\Omega = (0, 1)^2$ and we are given a polluted image represented by a function $\omega := g + \sigma\eta$ for some $g: \Omega \to [0, 1]$, $\sigma > 0$, and normally distributed noise $\eta \sim \mathcal{N}(0, 1)$. Our goal is to recover an approximation to the original image represented by g. One approach is to consider an optimisation problem where the minimiser is regularised by the total variation seminorm. In other words, we seek $u: \Omega \to \mathbb{R}$ that minimises

$$\min_{u \in \mathrm{BV}(\Omega)} \frac{1}{2}\|u - \omega\|^2_{L^2(\Omega)} + \delta|u|_{\mathrm{TV}}, \tag{88}$$
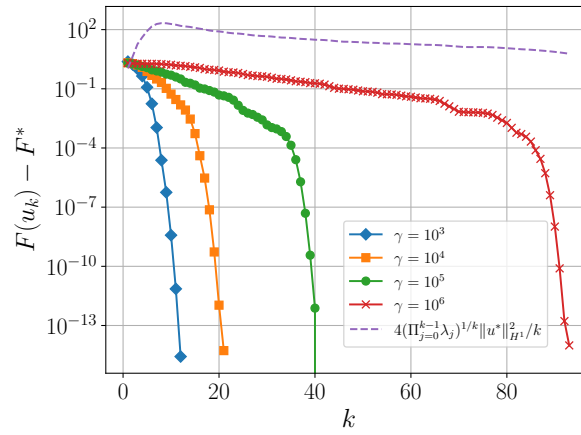
Figure 3: The reduction in the difference between the current and final objective with increasing iteration number k when using Algorithm 1 to minimise (87) for various choices of γ. We observe local superlinear convergence. The dashed line indicates the theoretical global convergence rate when $\gamma = 10^6$.

where $|u|_{\mathrm{TV}}$ denotes the total variation of u,

$$|u|_{\mathrm{TV}} \coloneqq \sup \left\{ \int_\Omega u \operatorname{div}(v) \, \mathrm{d}x : v \in C_c^\infty(\Omega; \mathbb{R}^2), \|v\|_{L^\infty(\Omega)} \leq 1 \right\},$$

and $\mathrm{BV}(\Omega)$ denotes the space of functions of bounded variation [58]. Regularising u with its total variation, as opposed to the H^1-norm, is known to better reduce blurring and better preserve edges [91, 58]. However, it is difficult to directly numerically tackle (88). Hence, we follow the approach outlined in [58] and introduce the equivalent Fenchel predual problem which seeks $p \colon \Omega \to \mathbb{R}^2$ that minimises

$$\min_{p \in H_0(\operatorname{div})} \frac{1}{2} \| \operatorname{div}(p) + \omega \|_{L^2(\Omega)}^2 \text{ subject to } -\delta \mathbf{1} \leq p \leq \delta \mathbf{1} \text{ a.e. in } \Omega, \tag{89}$$

where $\mathbf{1} = (1,1)^\top$. Here $H_0(\operatorname{div})$ denotes the space

$$H_0(\operatorname{div}) \coloneqq \{p \in L^2(\Omega; \mathbb{R}^2) : \operatorname{div}(p) \in L^2(\Omega), \ p \cdot \tilde{n} = 0 \text{ on } \partial\Omega\},$$

where $\tilde{n}$ is the outward normal to $\partial\Omega$. The minimisers u of (88) and p of (89) are then related by $u = \operatorname{div}(p) + \omega$ [58, Theorem 2.2].

To approximate the minimiser of (89), we first mesh the domain into 400×400 uniform quadrilateral cells and employ a first-order $H_0(\operatorname{div})$-conforming Raviart–Thomas (RT_1) discretisation for p [37, Chapter 14.5.2]. To enforce the pointwise box constraints, we regularise the problem with a nonsmooth Moreau–Yosida regularisation with penalisation parameter $\gamma > 0$. We also include a broken H^1-smoothing term controlled by $\epsilon > 0$. Altogether we seek $p_h \colon \Omega \to \mathbb{R}^2$ that satisfies

$$p_h = \operatorname*{arg\,min}_{q_h \in \mathrm{RT}_1} F(q_h) \text{ subject to } q_h \cdot \tilde{n} = 0 \text{ on } \partial\Omega,$$

$$\text{where } F(q) = \frac{1}{2} \| \operatorname{div}(q) + \omega \|_{L^2(\Omega)}^2$$
$$+ \frac{\gamma}{2} \| \max(\mathbf{0}, q - \delta \mathbf{1}) \|_{L^2(\Omega)}^2 + \frac{\gamma}{2} \| \min(\mathbf{0}, q + \delta \mathbf{1}) \|_{L^2(\Omega)}^2 + \epsilon |q|_{H^1(\mathcal{T}_h)}^2.$$

Here $\mathcal{T}_h$ is the triangulation of the domain and $|q|^2_{H^1(\mathcal{T}_h)} \coloneqq \sum_{K \in \mathcal{T}_h} |\nabla q|^2_{L^2(K)}$. Note that we set $\psi \equiv 0$ in (P).

We now fix the model parameters $\delta = 10^{-4}$, $\epsilon = 10^{-1}$, and $\sigma = 0.06$. We pick g to be the piecewise constant finite element projection of the Shepp–Logan phantom image [94]. We will consider various choices for γ.

We plot the original and polluted images as well as the restoration in Figure 4a. We also provide the approximate solutions' peak signal-to-noise ratio (PSNR) as defined by $\mathrm{PSNR}(u) = 10 \log_{10}(|\Omega| \cdot \|u - g\|^{-2}_{L^2(\Omega)})$. As in the previous example, in Figure 4b we compare Algorithm 1 against a Newton solver without a linesearch, with an ℓ^2-minimising linesearch [18, Algorithm 2], and a backtracking linesearch [92, Chapter 6.3] for a range of values of γ. We observe that the classical Newton method is surprisingly effective requiring the fewest linear system solves when $\gamma \leq 10^6$. However, it fails to converge once $\gamma = 10^7$. The other three solves do converge. Aside from when $\gamma = 10^4$, backtracking Newton requires the largest number of iterations. Algorithm 1 and ℓ^2-Newton have a similar performance throughout.
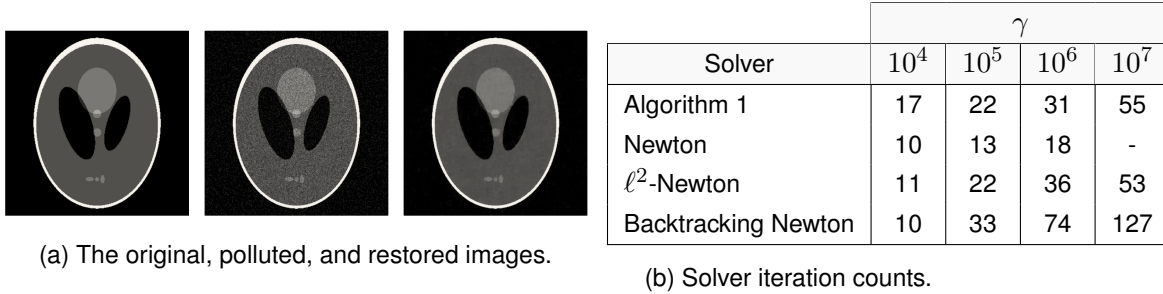

(a) The original, polluted, and restored images.

Solver	γ			
	10^4	10^5	10^6	10^7
Algorithm 1	17	22	31	55
Newton	10	13	18	-
ℓ^2-Newton	11	22	36	53
Backtracking Newton	10	33	74	127

(b) Solver iteration counts.

Figure 4: Results for the image restoration problem of Section 7.2 with parameters $\delta = 10^{-4}$, $\sigma = 0.06$, and $\epsilon = 10^{-1}$. (Left) The original (g), polluted (ω, PSNR: 27.97) and restored ($u = \mathrm{div}(p) + \omega$, PSNR: 39.10) images when $\gamma = 10^7$. (Right) The number of linear system solves (including those leading to rejected updates in Algorithm 1) to reach a residual vector ℓ^2-norm ($\|F'\|_{\ell^2}$) of 10^{-8} for increasing values of γ. We use the initial guess $p = \delta \mathbf{1}$ for each value of γ. Algorithm 1 is initialised with $\alpha = \beta = 10^{-4}$ and $\Lambda_0 = 64$. We use the norm $\|p\|^2 = \|p\|^2_{L^2(\Omega)} + \|\mathrm{div}\, p\|^2_{L^2(\Omega)} + |p|^2_{H^1(\mathcal{T}_h)}$ for the primal space norm and its corresponding discrete dual norm for $\| \cdot \|_*$. A dashed line indicates a failure to converge within 500 linear system solves.
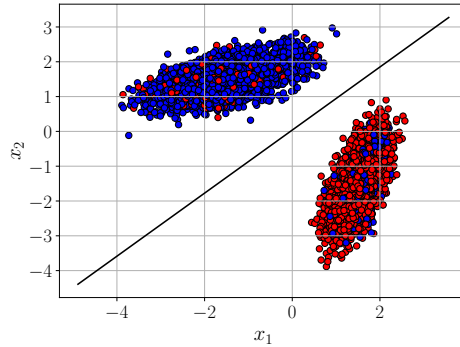
7.3 Support Vector Machine classification

Our final example is a Support Vector Machine classification problem from the field of machine learning [26]. Given a number of points $x_1, \ldots, x_\ell \in \mathbb{R}^n$ for some $\ell, n \in \mathbb{N}$ where each point x_i is classified into one of two groups denoted by $y_i \in \{-1, 1\}$, $i = 1, \ldots, \ell$, the aim is to find an $\omega \in \mathbb{R}^n$ and $b \in \mathbb{R}$ that defines a hyperplane $\omega^\top x + b = 0$ which separates the two groups. As in [102], given a $\gamma > 0$, the L2-loss SVM model seeks $(\omega, b) \in \mathbb{R}^n \times \mathbb{R}$ that minimises

$$F(\omega, b) = \frac{1}{2} \|\omega\|^2_{\ell^2} + \gamma \sum_{i=1}^{\ell} \max(1 - y_i(\omega^\top x_i + b), 0)^2.$$

We visualise the solution to one data set where $n = 2$ and $\ell = 10,000$ in Figure 5a. In Figure 5b, we test Algorithm 1 for $\ell = 10,000$ with $n \in \{2, 20, 200, 2000\}$ and $\gamma \in \{10^{-4}, 10^{-2}, 1, 10^2, 10^4\}$. The data x_i and y_i, $i \in \{1, \ldots, \ell\}$ are generated by a reproducible classification problem generator

from the Python package `scikit-learn` [81]. We observe that Algorithm 1 always converges for the parameters we consider. More iterations are required as both n and γ become larger.

(a) The data points and separating hyperplane.

	n			
γ	2	20	200	2000
10^{-4}	5	4	7	10
10^{-2}	6	10	18	21
10^{0}	7	21	24	30
10^{2}	8	27	31	36
10^{4}	9	34	37	43

(b) Solver iteration counts.

Figure 5: Results for the Support Vector Machine classification problem of Section 7.3. (Left) Plot of the data points $x_\ell \in \mathbb{R}^2$ and the separating line $\omega^\top x + b$ when $n = 2$ and $\gamma = 10^2$. A red and blue dot indicates a value of $y_i = 1$ and $y_i = -1$, respectively. (Right) The number of linear system solves (including those leading to rejected updates) for Algorithm 1 to reach a residual vector ℓ^2-norm ($\|F'\|_{\ell^2}$) of 10^{-6} for various values of γ and n. Algorithm 1 is initialised with $\alpha = \beta = 10^{-1}$, $\Lambda_0 = 3\gamma \cdot (\sum_{i=1}^{\ell} \|x_i\|_{\ell^2}^2)^{1/2}$, and $(\omega_0, b_0) = (\mathbf{1}/2, 1/2)$. We use the Euclidean norm $\|z\|^2 = \|z\|_*^2 = \|z\|_{\ell^2}^2$ for the primal and dual norms.

Conclusion

In this paper, we introduced our SSN-type algorithm LeAP-SSN (Algorithm 1) and showed that it possesses classical local superlinear convergence as well as global convergence rates and, in some settings, accelerates to global superlinear convergence. Notably, these results are achieved with a relaxation of the standard strong convexity assumption to a PL condition. A key to our global convergence rates is Assumption 2, which is satisfied for a large class of semismooth functions. Another important feature of our algorithm is its adaptivity not only to the parameters of the function class (such as μ or L), but also to the class itself. In other words, we have one algorithm for all the considered settings without requiring knowledge of whether the function is convex or not, whether it satisfies the PL condition or not, nor whether the function is semismooth or not. We also explored how partial smoothness and active manifold identification can be utilised in the nonsmooth context. We concluded with some numerical experiments in both finite and infinite-dimensional contexts. We believe that this paper is a good starting point for future research on algorithms that combine global convergence rates that are well studied in the world of second-order methods for functions with Lipschitz Hessians and improved asymptotic convergence rates well developed in the works on SSN-type methods.

References

[1]   R. A. Adams. *Sobolev spaces*. New York: Academic Press, 1975. ISBN: 978-0-120-44150-1.

[2] A. Alphonse et al. "A Globalized Inexact Semismooth Newton Method for Nonsmooth Fixed-point Equations involving Variational Inequalities". In: *arXiv e-prints*, arXiv:2409.19637 (Sept. 2024), arXiv:2409.19637. DOI: `10.48550/arXiv.2409.19637`. arXiv: `2409.19637 [math.NA]`.

[3] A. Alphonse et al. *LeAP-SSN*. Version 2025.0.1. 2025. URL: `https://github.com/amal-alphonse/leapssn`.

[4] A. Alphonse et al. *LeAP-SSN (Zenodo)*. Version 2025.0.1. 2025. DOI: `10.5281/zenodo.16918760`.

[5] V. Apidopoulos, N. Ginatta, and S. Villa. "Convergence rates for the heavy-ball continuous dynamics for non-convex optimization, under Polyak–Łojasiewicz condition". In: *Journal of Global Optimization* 84.3 (2022), pp. 563–589. ISSN: 1573-2916. DOI: `10.1007/s10898-022-01164-w`. URL: `https://doi.org/10.1007/s10898-022-01164-w`.

[6] H. Attouch and J. Bolte. "On the convergence of the proximal algorithm for nonsmooth functions involving analytic features". In: *Mathematical Programming* 116.1 (2009), pp. 5–16.

[7] H. Attouch, J. Bolte, and B. F. Svaiter. "Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods". In: *Mathematical programming* 137.1 (2013), pp. 91–129.

[8] H. Attouch et al. "Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality". In: *Mathematics of operations research* 35.2 (2010), pp. 438–457.

[9] S. Balay et al. "PETSc users manual". In: (2019).

[10] R. J. Baraldi and D. P. Kouri. "A proximal trust-region method for nonsmooth optimization with inexact function and gradient evaluations". In: *Mathematical Programming* 201.1 (2023), pp. 559–598.

[11] G. Bareilles, F. Iutzeler, and J. Malick. "Newton acceleration on manifolds identified by proximal gradient methods". In: *Mathematical Programming* 200.1 (2023), pp. 37–70.

[12] A. Beck. *First-order methods in optimization*. SIAM, 2017.

[13] G Bento et al. *Convergence of Descent Optimization Algorithms under Polyak-Lojasiewicz-Kurdyka Conditions*. 2025.

[14] M. Berger and B. Gostiaux. *Differential Geometry: Manifolds, Curves, and Surfaces: Manifolds, Curves, and Surfaces*. Vol. 115. Springer Science & Business Media, 2012.

[15] J. Bolte et al. "From error bounds to the complexity of first-order descent methods for convex functions". In: *Mathematical Programming* 165 (2017), pp. 471–507.

[16] J. Borwein and A. Lewis. *Convex Analysis and Nonlinear Optimization: Theoryand Examples*. Springer, 2006.

[17] M. Brokate and M. Ulbrich. "Newton differentiability of convex functions in normed spaces and of a class of operators". In: *SIAM J. Optim.* 32.2 (2022), pp. 1265–1287. ISSN: 1052-6234,1095-7189. DOI: `10.1137/21M1449531`. URL: `https://doi.org/10.1137/21M1449531`.

[18] P. R. Brune et al. "Composing scalable nonlinear algebraic solvers". In: *SIAM Review* 57.4 (2015), pp. 535–565. DOI: `10.1137/130936725`.

[19]   Y. Carmon et al. "Optimal and adaptive monteiro-svaiter acceleration". In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. NIPS '22. New Orleans, LA, USA: Curran Associates Inc., 2022. ISBN: 9781713871088.

[20]   C. Cartis, N. I. Gould, and P. L. Toint. "Adaptive cubic regularisation methods for unconstrained optimization. Part I: motivation, convergence and numerical results". In: *Mathematical Programming* 127.2 (2011), pp. 245–295.

[21]   C. Cartis, N. I. Gould, and P. L. Toint. *Evaluation Complexity of Algorithms for Nonconvex Optimization: Theory, Computation and Perspectives*. SIAM, 2022.

[22]   X. Chen, Z. Nashed, and L. Qi. "Smoothing methods and semismooth methods for nondifferentiable operator equations". In: *SIAM J. Numer. Anal.* 38.4 (2000), pp. 1200–1216. ISSN: 0036-1429,1095-7170. DOI: 10.1137/S0036142999356719. URL: https://doi.org/10.1137/S0036142999356719.

[23]   G. Chierchia et al. *The proximity operator repository*. 2020. URL: http://proximity-operator.net/download/guide.pdf.

[24]   R. Cibulka, A Dontchev, and M. H. Geoffroy. "Inexact Newton Methods and Dennis–Moré Theorems for Nonsmooth Generalized Equations". In: *SIAM Journal on Control and Optimization* 53.2 (2015), pp. 1003–1019.

[25]   A. R. Conn, N. I. Gould, and P. L. Toint. *Trust region methods*. SIAM, 2000.

[26]   C. Cortes and V. Vapnik. "Support-vector networks". In: *Machine learning* 20.3 (1995), pp. 273–297. DOI: 10.1007/BF00994018.

[27]   Y. Cui and J.-S. Pang. *Modern nonconvex nondifferentiable optimization*. SIAM, 2021.

[28]   S. vom Dahl and C. Kanzow. "An inexact regularized proximal Newton method without line search". In: *Computational Optimization and Applications* 89.3 (2024), pp. 585–624.

[29]   J. E. Dennis Jr. and J. J. Moré. "A characterization of superlinear convergence and its application to quasi-Newton methods". In: *Math. Comp.* 28 (1974), pp. 549–560. ISSN: 0025-5718,1088-6842. DOI: 10.2307/2005926. URL: https://doi.org/10.2307/2005926.

[30]   N. Doikov. *super-newton*. 2022. URL: https://github.com/doikov/super-newton.

[31]   N. Doikov, K. Mishchenko, and Y. Nesterov. "Super-universal regularized newton method". In: *SIAM Journal on Optimization* 34.1 (2024), pp. 27–56.

[32]   N. Doikov and Y. Nesterov. "Gradient regularization of Newton method with Bregman distances". In: *Mathematical Programming* 204.1 (2024), pp. 1–25. ISSN: 1436-4646. DOI: 10.1007/s10107-023-01943-7. URL: https://doi.org/10.1007/s10107-023-01943-7.

[33]   A. L. Dontchev. "Generalizations of the Dennis-Moré theorem". In: *SIAM J. Optim.* 22.3 (2012), pp. 821–830. ISSN: 1052-6234,1095-7189. DOI: 10.1137/110833567. URL: https://doi.org/10.1137/110833567.

[34]   D. Drusvyatskiy, A. D. Ioffe, and A. S. Lewis. "Nonsmooth optimization using Taylor-like models: error bounds, convergence, and termination criteria". In: *Math. Program.* 185.1-2 (2021), pp. 357–383. ISSN: 0025-5610,1436-4646. DOI: 10.1007/s10107-019-01432-w. URL: https://doi.org/10.1007/s10107-019-01432-w.

[35]   D. Drusvyatskiy and A. S. Lewis. "Error bounds, quadratic growth, and linear convergence of proximal methods". In: *Mathematics of Operations Research* 43.3 (2018), pp. 919–948.

[36]   D. Drusvyatskiy and A. S. Lewis. "Optimality, identifiability, and sensitivity". In: *Mathematical Programming* 147.1 (2014), pp. 467–498.

[37]   A. Ern and J.-L. Guermond. *Finite Elements I*. Vol. 72. Springer, 2021. ISBN: 978-3-030-56340-0. DOI: `10.1007/978-3-030-56341-7`.

[38]   F. Facchinei. "Minimization of SC1 functions and the Maratos effect". In: *Operations Research Letters* 17.3 (1995), pp. 131–137.

[39]   F. Facchinei and J.-S. Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer.

[40]   J. Fadili, J. Malick, and G. Peyré. "Sensitivity analysis for mirror-stratifiable convex functions". In: *SIAM J. Optim.* 28.4 (2018), pp. 2975–3000. ISSN: 1052-6234,1095-7189. DOI: `10.1137/17M113825X`. URL: `https://doi.org/10.1137/17M113825X`.

[41]   G. Fichera. "Sul problema elastostatico di Signorini con ambigue condizioni al contorno". In: *Atti della Accademia Nazionale dei Lincei, Classe di Scienze Fisiche, Matematiche e Naturali,* 34 (1963), pp. 138–142.

[42]   P. Frankel, G. Garrigos, and J. Peypouquet. "Splitting methods with variable metric for Kurdyka–Łojasiewicz functions and general convergence rates". In: *Journal of Optimization Theory and Applications* 165 (2015), pp. 874–900.

[43]   A. Gasnikov et al. "Near Optimal Methods for Minimizing Convex Functions with Lipschitz p-th Derivatives". In: *Proceedings of the Thirty-Second Conference on Learning Theory*. Ed. by A. Beygelzimer and D. Hsu. Vol. 99. Proceedings of Machine Learning Research. arXiv:1809.00382. Phoenix, USA: PMLR, 2019, pp. 1392–1393. URL: `http://proceedings.mlr.press/v99/gasnikov19b.html`.

[44]   C. Geiersbach and T. Scarinci. "Stochastic proximal gradient methods for nonconvex problems in Hilbert spaces". In: *Computational Optimization and Applications* 78.3 (2021), pp. 705–740. ISSN: 1573-2894. DOI: `10.1007/s10589-020-00259-y`. URL: `https://doi.org/10.1007/s10589-020-00259-y`.

[45]   J.-J. Godeme. "Inertial Bregman Proximal Gradient under Partial Smoothness". In: *arXiv e-prints*, arXiv:2501.17542 (Jan. 2025), arXiv:2501.17542. DOI: `10.48550/arXiv.2501.17542`. arXiv: `2501.17542 [math.OC]`.

[46]   G. N. Grapiglia and Y. Nesterov. "Regularized Newton Methods for Minimizing Functions with Hölder Continuous Hessians". In: *SIAM Journal on Optimization* 27.1 (2017), pp. 478–506. DOI: `10.1137/16M1087801`. eprint: `https://doi.org/10.1137/16M1087801`. URL: `https://doi.org/10.1137/16M1087801`.

[47]   G. N. Grapiglia and Y. Nesterov. "Accelerated Regularized Newton Methods for Minimizing Composite Convex Functions". In: *SIAM Journal on Optimization* 29.1 (2019), pp. 77–99. DOI: `10.1137/17M1142077`. eprint: `https://doi.org/10.1137/17M1142077`. URL: `https://doi.org/10.1137/17M1142077`.

[48]   D. A. Ham et al. *Firedrake User Manual*. First. Imperial College London et al. May 2023. DOI: `10.25561/104839`.

[49] M. Hintermüller, K. Ito, and K. Kunisch. "The primal-dual active set strategy as a semismooth Newton method". In: *SIAM J. Optim.* 13.3 (2002), pp. 865–888. ISSN: 1052-6234,1095-7189. DOI: 10.1137/S1052623401383558. URL: https://doi.org/10.1137/S1052623401383558.

[50] M. Hintermüller. *Semismooth Newton methods and applications*. URL: https://www.math.uni-hamburg.de/home/hinze/Psfiles/Hintermueller_OWNotes.pdf.

[51] J. Hu et al. "On the analysis of semismooth Newton-type methods for composite optimization". In: *J. Sci. Comput.* 103.2 (2025), Paper No. 59, 31. ISSN: 0885-7474,1573-7691. DOI: 10.1007/s10915-025-02867-4. URL: https://doi.org/10.1007/s10915-025-02867-4.

[52] K. Ito and K. Kunisch. "Semi-smooth Newton methods for variational inequalities of the first kind". In: *M2AN Math. Model. Numer. Anal.* 37.1 (2003), pp. 41–62. ISSN: 0764-583X,1290-3841. DOI: 10.1051/m2an:2003021. URL: https://doi.org/10.1051/m2an:2003021.

[53] Q. Jin. "On growth error bound conditions with an application to heavy ball method". In: *arXiv preprint arXiv:2310.03947* (2023).

[54] H. Karimi, J. Nutini, and M. Schmidt. "Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition". In: *Joint European conference on machine learning and knowledge discovery in databases*. Springer. 2016, pp. 795–811.

[55] P. D. Khanh et al. "Globally convergent coderivative-based generalized Newton methods in nonsmooth optimization". In: *Mathematical Programming* 205.1 (2024), pp. 373–429.

[56] D. Kovalev and A. Gasnikov. "The first optimal acceleration of high-order methods in smooth convex optimization". In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. NIPS '22. New Orleans, LA, USA: Curran Associates Inc., 2022. ISBN: 9781713871088.

[57] A. Y. Kruger. "On Fréchet subdifferentials". In: vol. 116. 3. Optimization and related topics, 3. 2003, pp. 3325–3358. DOI: 10.1023/A:1023673105317. URL: https://doi.org/10.1023/A:1023673105317.

[58] K. Kunisch and M. Hintermüller. "Total bounded variation regularization as a bilaterally constrained optimization problem". In: *SIAM Journal on Applied Mathematics* 64.4 (2004), pp. 1311–1333. DOI: 10.1137/S0036139903422784.

[59] C.-p. Lee. "Accelerating inexact successive quadratic approximation for regularized optimization through manifold identification". In: *Math. Program.* 201.1-2 (2023), pp. 599–633. ISSN: 0025-5610,1436-4646. DOI: 10.1007/s10107-022-01916-2. URL: https://doi.org/10.1007/s10107-022-01916-2.

[60] A. S. Lewis and S. Zhang. "Partial smoothness, tilt stability, and generalized Hessians". In: *SIAM J. Optim.* 23.1 (2013), pp. 74–94. ISSN: 1052-6234,1095-7189. DOI: 10.1137/110852103. URL: https://doi.org/10.1137/110852103.

[61] A. S. Lewis. "Active sets, nonsmoothness, and sensitivity". In: *SIAM Journal on Optimization* 13.3 (2002), pp. 702–725.

[62] A. S. Lewis, J. Liang, and T. Tian. "Partial smoothness and constant rank". In: *SIAM Journal on Optimization* 32.1 (2022), pp. 276–291.

[63] A. S. Lewis and M. L. Overton. "Nonsmooth optimization via quasi-Newton methods". In: *Mathematical Programming* 141.1 (2013), pp. 135–163.

[64] A. S. Lewis and C. Wylie. "Active-set Newton methods and partial smoothness". In: *Math. Oper. Res.* 46.2 (2021), pp. 712–725. ISSN: 0364-765X,1526-5471. DOI: 10.1287/moor.2020.1075. URL: https://doi.org/10.1287/moor.2020.1075.

[65] J. Liang, J. Fadili, and G. Peyré. "Activity identification and local linear convergence of forward–backward-type methods". In: *SIAM Journal on Optimization* 27.1 (2017), pp. 408–437.

[66] J. Liang, J. M. Fadili, and G. Peyré. "Local linear convergence of forward–backward under partial smoothness". In: *Advances in neural information processing systems* 27 (2014).

[67] F.-Y. Liao, L. Ding, and Y. Zheng. "Error bounds, PL condition, and quadratic growth for weakly convex functions, and linear convergences of proximal point methods". In: *6th Annual Learning for Dynamics & Control Conference*. arXiv:2312.16775. PMLR. 2024, pp. 993–1005.

[68] R. Liu, S. Pan, and Y. Qian. "An Inexact-Order Regularized Proximal Newton Method for Nonconvex Composite Optimization". In: *SIAM Journal on Optimization* 35.2 (2025), pp. 959–988.

[69] R. Liu et al. "An inexact regularized proximal Newton method for nonconvex and nonsmooth optimization". In: *Computational Optimization and Applications* 88.2 (2024), pp. 603–641.

[70] S. A. Miller and J. Malick. "Newton methods for nonsmooth convex minimization: connections among-Lagrangian, Riemannian Newton and SQP methods". In: *Mathematical programming* 104.2 (2005), pp. 609–633.

[71] K. Mishchenko. "Regularized Newton Method with Global $\mathcal{O}(1/k^2)$ Convergence". In: *SIAM Journal on Optimization* 33.3 (2023), pp. 1440–1462. DOI: 10.1137/22M1488752. eprint: https://doi.org/10.1137/22M1488752. URL: https://doi.org/10.1137/22M1488752.

[72] R. Monteiro and B. Svaiter. "An Accelerated Hybrid Proximal Extragradient Method for Convex Optimization and Its Implications to Second-Order Methods". In: *SIAM Journal on Optimization* 23.2 (2013), pp. 1092–1125. DOI: 10.1137/110833786. eprint: https://doi.org/10.1137/110833786. URL: https://doi.org/10.1137/110833786.

[73] B. S. Mordukhovich. *Second-order variational analysis in optimization, variational stability, and control: theory, algorithms, applications*. Springer Nature, 2024.

[74] B. S. Mordukhovich. *Variational analysis and generalized differentiation. I*. Vol. 330. Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Basic theory. Springer-Verlag, Berlin, 2006, pp. xxii+579.

[75] B. S. Mordukhovich et al. "A globally convergent proximal Newton-type method in nonsmooth convex optimization". In: *Mathematical Programming* 198.1 (2023), pp. 899–936.

[76] Y. Nesterov. "Accelerating the cubic regularization of Newton's method on convex problems". In: *Mathematical Programming* 112.1 (2008), pp. 159–181. ISSN: 1436-4646. DOI: 10.1007/s10107-006-0089-x. URL: https://doi.org/10.1007/s10107-006-0089-x.

[77] Y. Nesterov. *How to make the gradients small*. Optima, 88:10–11. 2012. URL: https://www.mathopt.org/Optima-Issues/optima88.pdf.

[78] Y. Nesterov and B. T. Polyak. "Cubic regularization of Newton method and its global performance". In: *Mathematical programming* 108.1 (2006), pp. 177–205.

[79]   W. Ouyang and A. Milzarek. "A trust region-type normal map-based semismooth Newton method for nonsmooth nonconvex composite optimization: W. Ouyang, A. Milzarek". In: *Mathematical Programming* 212.1 (2025), pp. 389–435.

[80]   J.-S. Pang and L. Qi. "A globally convergent Newton method for convex SC1 minimization problems". In: *Journal of Optimization Theory and Applications* 85.3 (1995), pp. 633–648.

[81]   F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[82]   B. T. Polyak. "Gradient methods for minimizing functionals". In: *Zhurnal vychislitel'noi matematiki i matematicheskoi fiziki* 3.4 (1963), pp. 643–653.

[83]   B. Pötzl, A. Schiela, and P. Jaap. "Inexact proximal Newton methods in Hilbert spaces". In: *Computational Optimization and Applications* 87.1 (2024), pp. 1–37.

[84]   B. Pötzl, A. Schiela, and P. Jaap. "Second order semi-smooth Proximal Newton methods in Hilbert spaces". In: *Computational Optimization and Applications* 82.2 (2022), pp. 465–498.

[85]   L. Q. Qi. "Convergence analysis of some algorithms for solving nonsmooth equations". In: *Math. Oper. Res.* 18.1 (1993), pp. 227–244. ISSN: 0364-765X,1526-5471. DOI: 10.1287/moor.18.1.227. URL: https://doi.org/10.1287/moor.18.1.227.

[86]   L. Q. Qi and J. Sun. "A nonsmooth version of Newton's method". In: *Math. Programming* 58.3 (1993), pp. 353–367. ISSN: 0025-5610,1436-4646. DOI: 10.1007/BF01581275. URL: https://doi.org/10.1007/BF01581275.

[87]   L. Qi. "Trust region algorithms for solving nonsmooth equations". In: *SIAM Journal on Optimization* 5.1 (1995), pp. 219–230.

[88]   Q. Rebjock and N. Boumal. "Fast convergence to non-isolated minima: four equivalent conditions for c 2 functions". In: *Mathematical Programming* (2024), pp. 1–49.

[89]   H. Rickmann, R. Herzog, and E. Herberg. "Global Convergence of Semismooth Newton Methods for Quadratic Problems". In: *GAMM Archive for Students* 7.1 (2025), p. 16. DOI: 10.14464/gammas.v7i1.810. URL: https://www.bibliothek.tu-chemnitz.de/ojs/index.php/GAMMAS/article/view/810.

[90]   R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*. Vol. 317. Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer-Verlag, Berlin, 1998, pp. xiv+733. ISBN: 3-540-62772-3. DOI: 10.1007/978-3-642-02431-3. URL: https://doi.org/10.1007/978-3-642-02431-3.

[91]   L. I. Rudin, S. Osher, and E. Fatemi. "Nonlinear total variation based noise removal algorithms". In: *Physica D: nonlinear phenomena* 60.1-4 (1992), pp. 259–268. DOI: 10.1016/0167-2789(92)90242-F.

[92]   R. B. Schnabel and J. E. Dennis. *Numerical methods for unconstrained optimization and nonlinear equations*. SIAM, 1996. ISBN: 978-0-89871-364-0. DOI: 10.1137/1.9781611971200.

[93]   T. Schwedes et al. *Mesh dependence in PDE-constrained optimisation*. Springer, 2017. DOI: 10.1007/978-3-319-59483-5.

[94]   L. A. Shepp and B. F. Logan. "The Fourier reconstruction of a head section". In: *IEEE Transactions on nuclear science* 21.3 (1974), pp. 21–43. DOI: 10.1109/TNS.1974.6499235.

[95]   M. Ulbrich. "Semismooth Newton methods for operator equations in function spaces". In: *SIAM J. Optim.* 13.3 (2002), pp. 805–842. ISSN: 1052-6234,1095-7189. DOI: 10.1137/S1052623400371569. URL: https://doi.org/10.1137/S1052623400371569.

[96]   M. Ulbrich. *Semismooth Newton methods for variational inequalities and constrained optimization problems in function spaces.* Vol. 11. MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics (SIAM), 2011, pp. xiv+308. ISBN: 978-1-611970-68-5. DOI: 10.1137/1.9781611970692. URL: https://doi.org/10.1137/1.9781611970692.

[97]   S. Vaiter, G. Peyré, and J. Fadili. "Model consistency of partly smooth regularizers". In: *IEEE Transactions on Information Theory* 64.3 (2017), pp. 1725–1737.

[98]   D. Wachsmuth. "A globalized inexact semismooth Newton for strongly convex optimal control problems". In: *arXiv preprint arXiv:2503.21612* (2025).

[99]   S. Wang, J. Fadili, and P. Ochs. "Convergence rates of regularized quasi-Newton methods without strong convexity". In: *arXiv preprint arXiv:2506.00521* (2025).

[100]   X. Xiao et al. "A regularized semi-smooth Newton method with projection steps for composite convex programs". In: *J. Sci. Comput.* 76.1 (2018), pp. 364–389. ISSN: 0885-7474,1573-7691. DOI: 10.1007/s10915-017-0624-3. URL: https://doi.org/10.1007/s10915-017-0624-3.

[101]   Y. Yamakawa and N. Yamashita. "Convergence analysis of a regularized Newton method with generalized regularization terms for unconstrained convex optimization problems". In: *Appl. Math. Comput.* 491 (2025), Paper No. 129219, 13. ISSN: 0096-3003,1873-5649. DOI: 10.1016/j.amc.2024.129219. URL: https://doi.org/10.1016/j.amc.2024.129219.

[102]   J. Yin and Q. Li. "A semismooth Newton method for support vector classification and regression". In: *Computational Optimization and Applications* 73.2 (2019), pp. 477–508. DOI: 10.1007/s10589-019-00075-z.

[103]   C. Zhang et al. "Riemannian trust region methods for SC^1 minimization". In: *J. Sci. Comput.* 101.2 (2024), Paper No. 32, 37. ISSN: 0885-7474,1573-7691. DOI: 10.1007/s10915-024-02664-5. URL: https://doi.org/10.1007/s10915-024-02664-5.