REVISITING SUPERLINEAR CONVERGENCE OF PROXIMAL NEWTON-LIKE METHODS TO DEGENERATE SOLUTIONS*

CHING-PEI LEE† AND STEPHEN J. WRIGHT‡

Abstract. We describe inexact proximal Newton-like methods for solving degenerate regularized optimization problems and for the broader problem of finding a zero of a generalized equation that is the sum of a continuous map and a maximal monotone operator. Superlinear convergence for both the distance to the solution set and a certain measure of first-order optimality can be achieved under a Hölderian error bound condition, including for problems in which the continuous map is nonmonotone, with Jacobian singular at the solution and not Lipschitz. Superlinear convergence is attainable even when the Jacobian is merely uniformly continuous, relaxing the standard Lipschitz assumption to its theoretical limit. For convex regularized optimization problems, we introduce a novel globalization strategy that ensures strict objective decrease and avoids the Maratos effect, attaining local Q-superlinear convergence without prior knowledge of problem parameters. Unit step size acceptance in our line search strategy does not rely on continuity or even existence of the Hessian of the smooth term in the objective, making the framework compatible with other potential candidates for superlinearly convergent updates.

Key words. proximal-Newton methods, regularized optimization, degenerate problems, superlinear convergence, Maratos effect, error bound

MSC codes. 90C53, 90C30

1. Introduction.

Consider the following generalized equation:

$$\text{Find } x \in \mathcal{H} \quad \text{such that} \quad 0 \in (A+B)(x), \tag{1.1}$$

where $\mathcal{H}$ is a real Hilbert space with an inner product $\langle\cdot,\cdot\rangle$ and induced norm $\|\cdot\|$, $A:\mathcal{H}\to\mathcal{H}$ is continuously differentiable, $B:\mathcal{H}\rightrightarrows 2^{\mathcal{H}}$ is a maximal-monotone set-valued operator, the solution set $\mathcal{S}$ is nonempty (and closed), and A is locally L-Lipschitz continuous in a neighborhood of $\mathcal{S}$. A point x solves (1.1) if and only if the forward-backward step $R(x)$ defined by

$$R(x) \coloneqq x - (\mathrm{Id}+B)^{-1}(\mathrm{Id}-A)(x) \tag{1.2}$$

is zero, where Id is the identity operator for $\mathcal{H}$. We define the norm of the forward-backward step as

$$r(x) \coloneqq \|R(x)\|. \tag{1.3}$$

Denoting the solution set of (1.1) by $\mathcal{S} \coloneqq \{x \in \mathcal{H} \,|\, R(x) = 0\}$, we further assume that the following order-q Hölderian error bound condition holds locally to $\mathcal{S}$:

$$\mathrm{dist}(x,\mathcal{S}) = \mathrm{dist}\big(x,(A+B)^{-1}(0)\big) \le \kappa r(x)^q, \quad \forall x \in \{x \mid r(x) \le \epsilon\}, \tag{1.4}$$

for some κ, $\epsilon > 0$ and $q \in (0,1]$, where $\mathrm{dist}(x,\mathcal{S})$ is the distance between the point x and the set $\mathcal{S}$, measured by the endowed norm on $\mathcal{H}$. Without assuming knowledge of the values of κ, ϵ, or q, we propose algorithms that attain fast local rates for a certain range of values of the exponent q. A well-known sufficient condition for (1.4) is the Hölderian metric subregularity condition of the same order q.

Some of our results assume also that the Jacobian ∇A is p-Hölder continuous in a neighborhood U of $\mathcal{S}$ for some $p \in (0,1]$, while others only assume uniform continuity. We say that ∇A is p-Hölder continuous in U if, for some $\zeta \ge 0$, we have

$$\|\nabla A(x) - \nabla A(y)\| \le \zeta \|x-y\|^p, \quad \forall x,y \in U \supset \mathcal{S}. \tag{1.5}$$

For simplicity, we often assume that U is the same as the neighborhood in which (1.4) holds, that is, $U = \{x \mid r(x) \le \epsilon\}$. Through simple calculus, we then have

$$\begin{aligned} \|A(x) - A(y) - \nabla A(y)(x-y)\| &\le \frac{\zeta}{1+p}\|x-y\|^{1+p} \\ &= O(\|x-y\|^{1+p}), \quad \forall x,y \in U. \end{aligned} \tag{1.6}$$

*Version of February 11, 2026.

†Institute of Statistical Mathematics, and Graduate University for Advanced Studies, SOKENDAI. Email: leechingpei@gmail.com

‡University of Wisconsin-Madison. Email: swright@cs.wisc.edu

∇A is Lipschitz continuous when it is Hölder continuous with $p = 1$.

In this work, we analyze a damped variant of a forward-backward method with Newton-like scaling for (1.1) under conditions of possible degeneracy. We account for cases in which the Jacobian of A at any $x^* \in \mathcal{S}$ could be singular and the solution set $\mathcal{S}$ may not be compact. We do not assume that the iterates $\{x_t\}$ have a limit or an accumulation point. We will first analyze local convergence under such degeneracy conditions, assuming that the Hölderian error bound (1.4) and the Hölder continuity condition on the Jacobian (1.5) are satisfied in a neighborhood of the solution set $\mathcal{S}$. For a certain range of values of p and q, we prove that both $r(x_t)$ and $\operatorname{dist}(x_t, \mathcal{S})$ converge superlinearly to 0 and that the full sequence of iterates $\{x_t\}$ converges. We show further that for the most widely considered case of $q = 1$ in (1.4), superlinear convergence is still attainable even if the assumption of Hölder continuity of ∇A is relaxed to local uniform continuity, provided that some algorithm parameters are adapted.

We then discuss the special case of regularized optimization

$$\min_{x \in \mathcal{H}} F(x) \coloneqq f(x) + \Psi(x), \tag{1.7}$$

where $\Psi : \mathcal{H} \to [-\infty, \infty]$ is convex, proper, and closed; and $f : \mathcal{H} \to \mathbb{R}$ is twice continuously differentiable with Lipschitz continuous gradient in an open set containing the domain of Ψ. We can express (1.7) in the form (1.1) by setting $A(x) \coloneqq \nabla f(x)$ and $B(x) \coloneqq \partial\Psi(x)$. When specialized to (1.7), the algorithmic framework considered in this work (see (1.8) below) is often called *proximal-Newton* or *sequential quadratic approximation*. We describe a globalization strategy for (1.7) that ensures global convergence and strict decrease of the objective even when conditions (1.4) and (1.5) do not hold. When (1.4) *is* satisfied and f is convex, the strategy guarantees eventual acceptance of the unit step size without requiring knowledge of problem-dependent parameters. This leads to fast local superlinear convergence for not only $r(x_t)$ and $\operatorname{dist}(x_t, \mathcal{S})$ but also the objective value $F(x_t)$, provided $\nabla^2 f$ is Hölder (or uniformly) continuous. Notably, in the scenario $q = p = \rho = 1$, our strategy achieves quadratic convergence, unlike existing backtracking and trust-region-like approaches.

Our algorithm for (1.1) is an inexact forward-backward method with Newton-like scaling, for which iteration t has the following form:

$$x_{t+1} \approx (H_t + B)^{-1} (H_t - A)(x_t), \quad H_t \coloneqq (\mu_t \mathrm{Id} + J_t), \quad \mu_t \coloneqq c r(x_t)^\rho, \tag{1.8}$$

where $r(x_t)$ is defined in (1.2) and (1.3), $c > 0$ and $\rho \geq 0$ are parameters, and J_t is a positive semidefinite but possibly non-Hermitian linear operator satisfying

$$\|J_t - \nabla A(x_t)\| = O\big(r(x_t)^\theta\big), \quad \text{for some } \theta \geq \rho. \tag{1.9}$$

When A is maximal monotone and differentiable, $\nabla A(x)$ is positive semidefinite for any x [15, see Lemma 3.5], so we could simply set $J_t = \nabla A(x_t)$ in this case. But we allow other scenarios too, such as when A is maximal monotone only in some region. The scheme (1.8) can be viewed as replacing $L \cdot \mathrm{Id}$ in the classical forward-backward splitting scheme (where L is an upper bound on the Lipschitz constant for A) by $H_t = \mu_t \mathrm{Id} + J_t$ as a variable-metric variant. Because the resolvent $(\mathrm{Id} + B)^{-1}$ of B is well-defined and single-valued due to the maximal monotonicity of B, so is $(H_t + B)^{-1}$, by positive semidefiniteness of J_t.

We can also view the scheme (1.8) as a generalization of the Newton method for solving the nonlinear equation

$$\text{Find } x \in \mathcal{H} \quad \text{such that} \quad A(x) = 0, \tag{1.10}$$

which is a special case of (1.1) with $B \equiv 0$.

We denote the *exact* solution for the next-iterate formula (1.8) by $\hat{x}_{t+1}$, that is,

$$\hat{x}_{t+1} \coloneqq (H_t + B)^{-1} (H_t - A)(x_t),$$

equivalently,

$$\text{find } \hat{x}_{t+1} \in \mathcal{H} \quad \text{such that} \quad 0 \in (B + H_t)(\hat{x}_{t+1}) - (H_t - A)(x_t).$$

If we define

$$\hat{R}_t(x) \coloneqq x - (\mathrm{Id} + B)^{-1} \left((\mathrm{Id} - H_t)(x) + (H_t - A)(x_t) \right), \quad \hat{r}_t(x) \coloneqq \|\hat{R}_t(x)\|, \tag{1.11}$$

then we have $\hat{r}_t(\hat{x}_{t+1}) = 0$. Using this residual, we consider the following requirement on the inexactness of the next iterate x_{t+1}:

$$\hat{r}_t(x_{t+1}) \leq \nu r(x_t)^{1+\rho}, \tag{1.12}$$

for some parameter $\nu \geq 0$, where ρ has the same meaning as in (1.8), where it is used to define μ_t.

1.1. Related Works. Most analyses of (1.1) focus on fixed-point algorithms and global convergence properties. There are fewer works on the variable-metric framework of (1.8), despite its practical interest. Newton and quasi-Newton approaches for the special case (1.10) (where $B \equiv 0$) are well-studied for both local and global properties under various globalization strategies (see, for example, [12, Chapter 11]). To our knowledge, there is no systematic treatment for extensions (1.8) of these approaches to the general case (1.1).

For regularized optimization (1.7), there is a substantial literature on algorithms like (1.8), sometimes known in this case as proximal-(quasi-)Newton or sequential quadratic approximation. Following the classical analysis for Newton's method, Lee et al. [7] assumed f strongly convex and Lipschitz twice continuously differentiable, and showed that x_t approaches the unique minimizer x^* Q-superlinearly and Q-quadratically for inexact and exact proximal-Newton methods, respectively, where $\mu_t \equiv 0$ in (1.8). Later, Yue et al. [16] showed that for f convex and Lipschitz twice continuously differentiable with a Lipschitz gradient, convergence of a variant of the damped proximal-Newton approach (1.8) (which they termed IRPN) is locally Q-superlinear for $\rho \in (0, 1)$, and Q-quadratic if $\rho = 1$ and the unit step size is eventually always accepted in their Armijo line search. Convergence here is for both $r(x_t)$ and $\mathrm{dist}(x_t, \mathcal{S})$ to 0 when the Luo-Tseng error-bound condition holds. (The latter condition corresponds to (1.4) with $q = 1$.) A disadvantage of the quadratic convergence result in [16], pointed out by Mordukhovich et al. [11], is that the Maratos effect must be avoided by selecting the line search parameters carefully to satisfy certain conditions involving the Lipschitz constants and the coefficient κ in (1.4). (The Maratos effect occurs when the unit step size might not yield sufficient decrease in the objective, causing unit steps that make good progress toward the solution to be rejected.) Selection of these parameters can be difficult in practice.

Mordukhovich et al. [11] showed that, under the same assumptions on f as [16], the conditions for superlinear convergence can be relaxed to (1.4) with $q > 1/2$, although a slightly weaker, R-superlinear convergence for $\mathrm{dist}(x_t, \mathcal{S})$ is obtained. They also established a convergence rate faster than quadratic when $q > 1$, but as we will see in the proof of Lemma 2.2, $r(x) = O(\mathrm{dist}(x, S))$ always holds, so (1.4) with $q > 1$ cannot hold in regimes of interest.[1]

To avoid the Maratos effect in the regime of quadratic convergence, [11] proposed a hybrid strategy that accepts unit steps when $r(x)$ decreases at a specified linear rate; they do not always check the objective value in (1.7), so it is not guaranteed to be monotonically decreasing. Doikov and Nesterov [3] studied the global convergence of a damped Newton approach for (1.7) that is a special case of (1.8). When both f and Ψ are convex and f is H-Lipschitz twice continuously differentiable, setting $\rho = 0.5$ and $c = \sqrt{H/3}$ in (1.8) and solving the subproblem exactly (that is, $\hat{r}_t(x_{t+1}) = 0$ in (1.11)) leads to strictly decreasing objective values and a global convergence rate of $O(t^{-2})$ for the objective value to the optimum. No special globalization strategies are required for this approach.

More recently, Liu et al. [8] extended the algorithm of [11] to nonconvex f restricted to the specific form of $f(x) = h(Ax - b)$ for some matrix A, vector b and a twice-differentiable function h whose Hessian is implicitly assumed to be locally Lipschitz. T Their approach requires computing the smallest eigenvalue of $\nabla^2 h$. Following an earlier preprint version of [11], they proved Q-superlinear convergence for $q > (\sqrt{5} - 1)/2 \approx .618$ under the assumption that the iterates have an accumulation point x^* with $\nabla^2 f(x^*)$ positive semidefinite. This requirement of local convexity is similar to our assumption that A is maximal monotone locally. When $q = 1$, their algorithm does not provide quadratic convergence as those variants in [11, 16]. Vom Dahl and Kanzow [14] proposed a trust-region-like variant of the method in [8] and obtained convergence results similar to those of that paper.

1.2. Contributions. We advance the state of the art in several respects.

1. In the degenerate setting where the Jacobian ∇A is singular at the solutions,We establish that superlinear convergence of proximal-Newton methods is significantly more robust to lack of smoothness than previously understood. Our analysis relaxes the standard assumption of Lipschitz continuity of the Jacobian to Hölder continuity of any order while retaining superlinear rates. We further show that superlinear convergence remains attainable even when this assumption is relaxed to merely uniform continuity, provided that the damping term and the stopping tolerance decay sufficiently slowly.
2. We propose a novel general line search strategy for convex regularized optimization that ensures strict function decrease at every iteration, guarantees global convergence to $\mathcal{S}$, and avoids the Maratos effect for any "fast direction" yielding superlinear convergence. Our strategy is agnostic to problem-dependent parameters and maintains acceptance of the unit step size even in difficult scenarios where existing back-

[1] Technically speaking, they assumed a q-metric subregularity condition such that $\mathrm{dist}(x, \mathcal{S}) \leq \kappa \mathrm{dist}(0, (A + B)(x))^q$ and showed that this condition is equivalent to (1.4) with the same q *provided that* $q \in (0, 1]$ in their Proposition 2.4. They then utilized (1.4) to obtain their convergence rates under the assumption of q-metric subregularity. For the case of $q > 1$, they leveraged the same proposition and claimed that such an equivalence continues to hold to obtain their rate — but the proof of their Proposition 2.4 shows that q-metric subregularity with $q > 1$ implies (1.4) only with $q = 1$. Therefore, for their proof of the faster-than-quadratic rate to hold, it is necessary to assume that (1.4) holds with $q > 1$.

tracking and trust-region approaches fail, such as the quadratic convergence regime and the non-Lipschitz Hessian setting described above. Notably, our analysis of unit step size acceptance does not rely on continuity properties or even the existence of the Hessian, but instead leverages the interplay between the Hölderian error bound condition and properties of the update direction. This feature makes the framework compatible with any direction yielding superlinear convergence, including potential candidates like proximal semismooth Newton or proximal quasi-Newton methods for more difficult scenarios, such as when f is not twice-differentiable.

3. Our line-search framework achieves Q-superlinear convergence for the objective value F in (1.7). We believe this result to be the first of its kind, even for nondegenerate instances of (1.7) in which $\nabla^2 f$ is nonsingular.
4. We extend local convergence rates previously established for regularized optimization to the generalized equation setting (1.1) in Hilbert spaces. We thus allow the use of a non-Hermitian Jacobian in (1.8). Through a refined error analysis, we broaden the range of the error bound exponent for Q-superlinear convergence of $\operatorname{dist}(x_t, \mathcal{S})$ to $1 \geq q > (\sqrt{33}-1)/8 \approx 0.593$, improving upon $q = 1$ in [16] and $1 > q > (\sqrt{5}-1)/2 \approx 0.618$ in [8, 14]. We obtain the same range of $q > 1/2$ for R-superlinear convergence of $\operatorname{dist}(x_t, \mathcal{S})$ and Q-superlinear convergence of $r(x_t)$, matching the range in [11] but for the more general problem (1.1).

1.3. Notation. For bounded nonnegative scalar sequences $\{\sigma_k\}$ and $\{\tau_k\}$, we say that $\sigma_k = o(\tau_k)$ if for any $\beta > 0$ we have $\sigma_k \leq \beta\tau_k$ for all k sufficiently large. We say $\sigma_k = O(\tau_k)$ if this bound holds for *some* $\beta > 0$ and all k sufficiently large. We say $\sigma_k = \Omega(\tau_k)$ if there is some $\beta > 0$ such that $\sigma_k \geq \beta\tau_k$ for all k sufficiently large.

1.4. Organization. Section 2 discusses local convergence rates of (1.8), including the ranges of q, p, and ρ that yield R- and Q-superlinear convergence, respectively. In Section 3 we propose a new globalization strategy for (1.7), analyzing its global convergence as well as local Q-superlinear convergence for $r(x_t)$, $\operatorname{dist}(x_t, \mathcal{S})$, and $F(x_t)$. Simplification of our algorithm for the case of smooth optimization is discussed in Section 4. The result of superlinear convergence when ∇A is only uniform continuity is in Appendix A.

2. Local Convergence. We describe local convergence properties of our basic algorithm (1.8), (1.12) in this section. Section 2.1 shows convergence of $\{x_t\}$ to the solution set $\mathcal{S}$, and of the residual $r(x_t)$ to zero, at superlinear rates, under certain conditions.

2.1. Superlinear Convergence. For the problem (1.1), we start by showing that under suitable conditions, the sequence $\{x_t\}$ defined by (1.8), (1.12) exhibits local superlinear convergence to the solution set $\mathcal{S}$ (and quadratic convergence, under stronger assumptions) and that $r(x_t)$ converges to 0 at the same rate. We define the notation

$$d_t := \operatorname{dist}(x_t, \mathcal{S}), \quad r_t := r(x_t), \quad p_t := x_{t+1} - x_t, \tag{2.1}$$

and derive estimates of some key quantities in terms of these values. Lemma 2.1 is a generalization to the setting of (1.1) with Hölder continuous ∇A of [16, Lemma 4] and [11, Lemma 4.1], which apply to (1.7) with Lipchitz continuous $\nabla^2 f$.

LEMMA 2.1. *Fix an iterate x_t and consider the update scheme (1.8), (1.12) for (1.1) with $\nu \geq 0$, $\rho \geq 0$, A single-valued and continuously differentiable, B maximal monotone, and $\mathcal{S} \neq \emptyset$. Assume that in a neighborhood containing both x_t and x_{t+1}, A is L-Lipschitz continuous for some $L \geq 0$ and ∇A is p-Hölder continuous for some $p \in (0, 1]$. Then we obtain*

$$\|p_t\| = \|x_{t+1} - x_t\| \leq O(d_t) + O\big(\mu_t^{-1} d_t^{1+p}\big) + O\big(\mu_t^{-1} r_t^{1+\rho}\big) + O\big(r_t^{1+\rho}\big). \tag{2.2}$$

Proof. Let $\bar{x}_t \in P_{\mathcal{S}}(x_t)$, where $P_{\mathcal{S}}$ denotes projection onto the solution set $\mathcal{S}$ using the endowed norm, so that $d_t = \|x_t - \bar{x}_t\|$.[2] Denoting by $\bar{U}$ the neighborhood assumed in the lemma, we have from the local Lipschitz continuity of A that

$$\|\nabla A(x)\| \leq L, \quad \forall x \in \bar{U}. \tag{2.3}$$

Defining

$$\xi_t := \hat{R}_t\left(x_{t+1}\right), \tag{2.4}$$

[2] The solution set $\mathcal{S}$ might be nonconvex and thus the projection could be non-unique. However, according to our assumption that $\mathcal{S} \neq \emptyset$, there must exist at least one such $\bar{x}_t \in P_{\mathcal{S}}(x_t)$. Regardless, d_t is uniquely defined.

we have after rearranging (1.11) that

$$\begin{aligned}&(H_t - A)(x_t) - (H_t - \mathrm{Id})(x_{t+1}) \in (\mathrm{Id} + B)(x_{t+1} - \xi_t)\\ \Rightarrow\quad &\xi_t - H_t(\xi_t) \in (A - H_t)(x_t) + (H_t + B)(x_{t+1} - \xi_t),\end{aligned} \tag{2.5}$$

while the condition (1.12) implies that

$$\|\xi_t\| \le \nu r_t^{1+\rho}. \tag{2.6}$$

Because $\bar{x}_t \in \mathcal{S}$, we have from the optimality condition of (1.1) that

$$\begin{aligned}&-A(\bar{x}_t) \in B(\bar{x}_t)\\ \Rightarrow\ &H_t(\bar{x}_t - x_t) + A(x_t) - A(\bar{x}_t) \in (A - H_t)(x_t) + (H_t + B)(\bar{x}_t).\end{aligned} \tag{2.7}$$

Since J_t is positive semidefinite and B is monotone, we have from the definition of H_t in (1.8) that

$$\langle v - u,\, H_t(v-u)\rangle + \langle v - u,\, z - w\rangle \ge \mu_t \|u - v\|^2, \quad \text{for any } (u,w) \text{ and } (v,z) \text{ in } \mathrm{graph}(B).$$

We thus obtain from (2.5) and (2.7) that

$$\begin{aligned}&\langle H_t(\bar{x}_t - x_t) + A(x_t) - A(\bar{x}_t) - \xi_t + H_t(\xi_t),\, \bar{x}_t - x_{t+1} + \xi_t\rangle\\ &\ge \mu_t \|\bar{x}_t - x_{t+1} + \xi_t\|^2,\end{aligned} \tag{2.8}$$

which together with (1.8), (1.9), and (2.6) implies that

$$\begin{aligned}&\|\bar{x}_t - x_{t+1} + \xi_t\|\\ \overset{(1.8),(1.9)}{\le}\ &\mu_t^{-1}\big(\big\|(\mu_t + r_t^\theta)(\bar{x}_t - x_t) + \nabla A(x_t)(\bar{x}_t - x_t) + A(x_t) - A(\bar{x}_t)\big\|\\ &\qquad +(1 + \|H_t\|)\|\xi_t\|\big)\\ \overset{(1.6),(2.3),(2.6)}{\le}\ &\mu_t^{-1}\big(O(\mu_t d_t) + O\big(d_t^{1+p}\big) + \nu\,(1 + L + O(\mu_t))\, r_t^{1+\rho}\big),\end{aligned} \tag{2.9}$$

where in the final inequality, we used the fact that $\theta \ge \rho$ implies $r_t^\theta = O(r_t^\rho) = O(\mu_t)$ and set $y = x_t, x = \bar{x}_t$ in applying (1.6). We also used the fact $\|H_t\| = O(1)$ derived from the below with $\theta \ge \rho > 0$:

$$\|H_t\| \overset{(1.8),(1.9)}{=} \big\|\mu_t I + \nabla A(x_t) + O(r_t^\theta)\big\| \overset{(1.8)}{\le} O(r_t^\rho) + \|\nabla A(x_t)\| \overset{(2.3)}{\le} O(r_t^\rho) + L = O(1).$$

Finally, using the triangle inequality, we can bound $\|x_{t+1} - x_t\|$ by

$$\begin{aligned}&\|x_{t+1} - x_t\|\\ \le\ &\|\bar{x}_t - x_{t+1} + \xi_t\| + \|x_t - \bar{x}_t\| + \|\xi_t\|\\ \overset{(2.9),(2.6)}{\le}\ &\mu_t^{-1}\Big(O(\mu_t d_t) + O\big(d_t^{1+p}\big) + O\big(1 + \mu_t\big)O\big(r_t^{1+\rho}\big)\Big) + d_t + O\big(r_t^{1+\rho}\big),\end{aligned}$$

verifying (2.2). ∎

LEMMA 2.2. *Suppose the assumptions of Lemma 2.1 hold. Then for x_{t+1} satisfying (1.12), we have*

$$r_{t+1} = O\big(d_t^{1+\rho}\big) + O\big(d_t^{1+p}\big) + O\big(\big(r_t^{-\rho} d_t^{1+p}\big)^{1+p}\big). \tag{2.10}$$

Proof. We have

$$\begin{aligned}r_{t+1} &\le \|R(x_{t+1}) - \hat{R}_t(x_{t+1})\| + \|\hat{R}_t(x_{t+1})\|\\ &= \big\|(\mathrm{Id} + B)^{-1}(\mathrm{Id} - A)(x_{t+1}) - (\mathrm{Id} + B)^{-1}\left((H_t - A)(x_t) - (H_t - \mathrm{Id})(x_{t+1})\right)\big\|\\ &\quad + \hat{r}_t(x_{t+1}).\end{aligned} \tag{2.11}$$

From the nonexpansiveness of the resolvent of B, we can bound the first term on the right-hand side of (2.11) by

$$\begin{aligned}&\big\|(\mathrm{Id} + B)^{-1}(\mathrm{Id} - A)(x_{t+1}) - (\mathrm{Id} + B)^{-1}\left((H_t - A)(x_t) - (H_t - \mathrm{Id})(x_{t+1})\right)\big\|\\ \le\ &\|A(x_t) - A(x_{t+1}) - H_t(x_t - x_{t+1})\|.\end{aligned} \tag{2.12}$$

Next, the fact that $\bar{x}_t \in P_{\mathcal{S}}(x_t)$ is a solution of (1.1) indicates that

$$(\mathrm{Id}+B)^{-1}(\mathrm{Id}-A)(\bar{x}_t) = \bar{x}_t.$$

From the nonexpansiveness of the resolvent of B and the expressions above, we obtain by applying the triangle inequality again that

$$\begin{aligned} r(x_t) &= \left\|x_t - (\mathrm{Id}+B)^{-1}(\mathrm{Id}-A)(x_t)\right\| \\ &= \left\|\left(x_t - (\mathrm{Id}+B)^{-1}(\mathrm{Id}-A)(x_t)\right) - \left(\bar{x}_t - (\mathrm{Id}+B)^{-1}(\mathrm{Id}-A)(\bar{x}_t)\right)\right\| \\ &\le \|\bar{x}_t - x_t\| + \|\bar{x}_t - x_t + A(x_t) - A(\bar{x}_t)\| \\ &\le (L+2)\|\bar{x}_t - x_t\| = (L+2)d_t, \end{aligned} \tag{2.13}$$

where we used local Lipschitz continuity of A in the last inequality. We also note from the definition of μ_t in (1.8) that

$$\mu_t^{-1} r_t^{1+\rho} = c^{-1} r_t \overset{(2.13)}{=} O(d_t). \tag{2.14}$$

By substituting (2.12) and (1.12) into (2.11) and using the definition $p_t \coloneqq x_{t+1} - x_t$, we can prove (2.10) as follows:

$$\begin{aligned} & r(x_{t+1}) \\ \le\ & \|A(x_t) - A(x_{t+1}) - H_t(x_t - x_{t+1})\| + \nu r_t^{1+\rho} \\ \overset{(1.8),(1.9)}{=}\ & \left\|\nabla A(x_t)(x_{t+1} - x_t) + A(x_t) - A(x_{t+1}) + (\mu_t + O(r_t^{\theta}))(x_{t+1} - x_t)\right\| \\ & + \nu r_t^{1+\rho} \\ \overset{(1.6),(1.8),(1.9),(2.1)}{\le}\ & O\left(\|p_t\|^{1+p}\right) + O(\mu_t)\|p_t\| + \nu r_t^{1+\rho}. \end{aligned} \tag{2.15}$$

Thus from Lemma 2.1, we obtain

$$\begin{aligned} r_{t+1} \quad \le\quad & O\left(d_t^{1+p} + \left(\mu_t^{-1} d_t^{1+p}\right)^{1+p} + \left(\mu_t^{-1} r_t^{1+\rho}\right)^{1+p} + \left(r_t^{1+\rho}\right)^{1+p}\right) \\ & + O(\mu_t d_t) + O\left(d_t^{1+p}\right) + O\left(r_t^{1+\rho}\right) + O\left(\mu_t r_t^{1+\rho}\right) \\ \overset{(2.13),(2.14)}{=}\ & O\left(d_t^{1+p}\right) + O\left(\left(r_t^{-\rho} d_t^{1+p}\right)^{1+p}\right) + O\left(d_t^{1+\rho}\right), \end{aligned}$$

where in the last equality we used $\rho > 0$ to deduce that $(1+\rho)(1+p) \ge (1+p)$ and $1 + 2\rho \ge 1 + \rho$. ∎

It follows from (2.13) that $q > 1$ in (1.4) is possible only in trivial circumstances. Our main superlinear convergence is as follows.

THEOREM 2.3. *Consider the update scheme (1.8), (1.12) for (1.1) for some $\nu, \rho \ge 0$, with A single-valued and continuously differentiable, B maximal monotone, and $\mathcal{S} \ne \emptyset$. Assume that (1.4) holds for some $q > 0$ in a neighborhood $V \coloneqq \{x \mid r(x) \le \epsilon\}$ of $\mathcal{S}$ for some $\epsilon > 0$, and that A is L-Lipschitz continuous for some $L \ge 0$ and ∇A is p-Hölder continuous (1.5) for some $p \in (0, 1]$ within the same neighborhood. Suppose that the following inequalities are satisfied:*

$$\begin{cases} (1+\rho)q > 1, \\ (1+p)q > 1, \\ (q + pq - \rho)(1+p) > 1. \end{cases} \tag{2.16}$$

Then if r_0 is sufficiently small, we have Q-superlinear convergence of $\{r_t\}$ and $\{d_t\}$ according to

$$r_{t+1} = O\left(r_t^{1+s}\right), \quad d_{t+1} = O\left(d_t^{1+s}\right), \tag{2.17}$$

where s is the smallest gap between the left- and right-hand sides in (2.16), that is,

$$s \coloneqq \min\left\{(1+\rho)q,\ (1+p)q,\ (q+pq-\rho)(1+p)\right\} - 1 > 0.$$

Proof. Suppose for some $t \geq 0$ that $r_t \leq \epsilon_1$ for some $\epsilon_1 \in (0, \epsilon]$ whose value is defined below. From (1.4) we have

$$r_t^{-\rho} = (r_t^q)^{-\frac{\rho}{q}} = O\big(d_t^{-\frac{\rho}{q}}\big). \tag{2.18}$$

By substituting (2.18) into (2.10) we then have

$$\begin{aligned} r_{t+1} &= O\big(d_t^{1+\rho}\big) + O\big(d_t^{1+p}\big) + O\big(\big(d_t^{-\frac{\rho}{q}} d_t^{1+p}\big)^{1+p}\big) \\ &= O\Big(d_t^{1+\rho} + d_t^{1+p} + d_t^{\left(1+p-\frac{\rho}{q}\right)(1+p)}\Big). \end{aligned} \tag{2.19}$$

By applying (1.4) to the right-hand side of (2.19), we obtain the first equation in (2.17), as well as $r_{t+1} = O(\epsilon_1^s r_t)$. Because $s > 0$, we can decrease ϵ_1 if necessary to ensure that $r_{t+1} < \frac{99}{100} r_t \leq \frac{99}{100}\epsilon_1$ for *all* x_t with $r_t \leq \epsilon_1$, showing that $x_{t+1} \in V$. We can thus also apply (2.19) to (1.4) to obtain

$$d_{t+1} = O\left(r_{t+1}^q\right) = O\Big(d_t^{(1+p)q} + d_t^{(1+\rho)q} + d_t^{(q+pq-\rho)(1+p)}\Big) = O\left(d_t^{1+s}\right),$$

proving the case of Q-superlinear convergence for $\{d_t\}$. The proof is completed by noting that if x_0 is such that $r_0 \leq \epsilon_1$, then $\{r_t\}$ decreases monotonically to zero. ∎

If we seek only R-superlinear convergence for $\{d_t\}$, while retaining Q-superlinear convergence for $\{r_t\}$, a different range of parameters is allowed.

THEOREM 2.4. *Suppose that the assumptions of Theorem 2.3 hold, except that in place of (2.16), the following inequalities are satisfied:*

$$\begin{cases} (1+p)q > 1, \\ (q+pq-\rho)(1+p) > 1, \\ \rho + q > 1, \\ \rho > 0. \end{cases} \tag{2.20}$$

Then we obtain Q-superlinear convergence within V for $\{r_t\}$ according to

$$r_{t+1} = O\left(r_t^{1+\bar{s}}\right), \tag{2.21}$$

where $\bar{s}$ is the smallest gap between the left- and right-hand sides in (2.20), that is,

$$\bar{s} := \min\big\{(1+p)q,\, (q+pq-\rho)(1+p),\, \rho+q,\, 1+\rho\big\} - 1 > 0. \tag{2.22}$$

Moreover, we have R-superlinear convergence within V for $\{d_t\}$.

Proof. As in the proof of Theorem 2.3, suppose that $r_t \leq \epsilon_1 \leq \epsilon$ for some $\epsilon_1 > 0$. From (1.4) we have $d_t = O(r_t^q)$, so that $r_t^{-1} = O(d_t^{-1/q})$. Using this bound together with Lemma 2.1 and (2.13), we obtain

$$p_t := x_{t+1} - x_t = O(d_t) + O\Big(d_t^{1+p-\frac{\rho}{q}}\Big) = O\Big(d_t^{\min\left\{1,\,1+p-\frac{\rho}{q}\right\}}\Big). \tag{2.23}$$

Substitution of (2.23) into (2.15) then leads to

$$\begin{aligned} r_{t+1} &\leq O(d_t^{1+p}) + O\Big(d_t^{(1+p-\frac{\rho}{q})(1+p)}\Big) + \mu_t O(d_t) + \mu_t O\Big(d_t^{1+p-\frac{\rho}{q}}\Big) + \nu r_t^{1+\rho} \\ &\overset{(1.4)}{=} O(r_t^{q(1+p)}) + O\Big(r_t^{(q+pq-\rho)(1+p)}\Big) + O(r_t^{\rho+q}) + O\Big(r_t^{\rho+(q+pq-\rho)}\Big) + O(r_t^{1+\rho}) \\ &\overset{(2.22)}{=} O(r_t^{1+\bar{s}}), \end{aligned}$$

which is exactly (2.21), and $\bar{s} > 0$ if and only if (2.20) holds. Note that by choosing ϵ_1 sufficiently small, we can ensure that $r_{t+1} \leq \frac{99}{100} r_t \leq \frac{99}{100}\epsilon_1$, so by requiring that $r_0 \leq \epsilon_1$, we have that $\{r_t\}$ decreases to zero, and in fact converges superlinearly to zero since $\bar{s} > 0$.

To prove R-superlinear convergence of $\{d_t\}$, we note that by defining $\hat{d}_t = \kappa r_t^q$ and combining (2.21) and (1.4), we have

$$d_t \leq \hat{d}_t, \quad \hat{d}_{t+1} = O(\hat{d}_t^{1+\bar{s}}), \quad \forall t \geq 0,$$

thus showing that $\{d_t\}$ is dominated by a Q-superlinearly convergent sequence. ∎

When the conditions of either Theorem 2.3 or Theorem 2.4 hold, we also obtain strong convergence for the iterates to a solution point.

THEOREM 2.5. *Assume that the conditions of either Theorem 2.3 or Theorem 2.4 hold. Then $\{x_t\}$ converges strongly to some point $x^* \in \mathcal{S}$.*

Proof. Theorem 2.3 and Theorem 2.4 both show that $x_t \in V$ for all t and that $\{r_t\}$ converges superlinearly to zero. By using (1.4) in (2.2), we see that p_t is bounded by r_t as follows:

$$\|p_t\| \leq O(r_t^q) + O\big(r_t^{(1+p)q-\rho}\big) + O(r_t) + O\big(r_t^{1+\rho}\big) = O\big(r_t^{\min\{q,1,1+\rho,(1+p)q-\rho\}}\big).$$

From the constraints in (2.16) and (2.20), together with $\rho \geq 0$ and $q > 0$, we obtain

$$\min\big\{1+\rho, (1+p)q-\rho, q, 1\big\} > 0,$$

so (1.4) and (2.23) indicate that there is $\tau > 0$ such that $\|p_t\| = O(r_t^\tau)$. Theorems 2.3 and 2.4 show that $\{r_t\}$ converges superlinearly to 0. Therefore, we can find $t_1 \geq 0$ and $\eta \in [0, 1)$ such that

$$r_{t+1} \leq \eta r_t, \quad \forall t \geq t_1.$$

Thus we get

$$\sum_{t=t_1}^{\infty} \|p_t\| = O\bigg(\sum_{t=t_1}^{\infty} \eta^{\tau(t-t_1)} r_{t_1}^\tau\bigg) = O\bigg(\frac{r_{t_1}^\tau}{1-\eta^\tau}\bigg) < \infty,$$

so $\{\|p_t\|\}$ is summable. Thus $\{x_t\}$ is a Cauchy sequence, so it converges strongly to a point x^* because of completeness of the Hilbert space $\mathcal{H}$. Moreover, $\operatorname{dist}(x_t, \mathcal{S}) \to 0$ implies that $\operatorname{dist}(x^*, \mathcal{S}) = 0$, so by closedness of $\mathcal{S}$ we have $x^* \in \mathcal{S}$. ∎

Comparing (2.16) and (2.20), we see that the difference is that the inequality

$$(1+\rho)q > 1 \tag{2.24}$$

in the former is replaced by the two inequalities

$$\rho + q > 1, \quad \rho > 0 \tag{2.25}$$

in the latter. Since $q \leq 1$, we have $\rho \geq \rho q$, so (2.24) implies (2.25). We discuss several interesting realizations of (2.16) and (2.20) below. (To our knowledge, the analysis for generalized equation (1.1) is new, and for the special case (1.7), our bounds are new.)

Remark 2.6 (Convergence Regimes).

1. Quadratic convergence occurs when $p = \rho = q = 1$, that is, when the Luo-Tseng error bound holds and $\nabla^2 f$ is Lipschitz continuous (for (1.7)) or when the error bound condition with $q = 1$ holds with ∇A Lipschitz continuous (for (1.1)). (Our results show that $p = 1$ and $q \geq \rho \geq 1$ imply a convergence rate that is at least quadratic.)
2. If $p = 1$ and $q \leq 1$, (2.16) implies

$$\begin{cases} 2q > 1 \\ (1+\rho)q > 1 \\ 2(2q-\rho) > 1 \end{cases} \Leftrightarrow \begin{cases} q > \frac{1}{2} \\ \rho > \frac{1-q}{q} \\ 2q - \frac{1}{2} > \rho \end{cases} \Leftrightarrow \begin{cases} q > \frac{1}{2} \\ q > \frac{-1+\sqrt{33}}{8} \\ \rho \in (\frac{1-q}{q}, 2q - \frac{1}{2}). \end{cases}.$$

(We obtain the second formula in the last column by ensuring the ρ-interval is nonempty.) We thus attain Q-superlinear convergence of $\{d_t\}$ provided that $q > \frac{1}{8}(\sqrt{33}-1)) \approx 0.593$, as long as $\rho < 2q - \frac{1}{2}$. (We can set $\rho = \frac{1}{4}(\sqrt{33}-1) - \frac{1}{2} = \frac{1}{4}(\sqrt{33}-3)$ to maximize the allowable range of q.) By comparison with the result of [8, 11, 14], which required $\rho \leq q$, our analysis shows that superlinear convergence can be obtained even in some situations where $\rho > q$. In addition, [11] does not guarantee Q-superlinear convergence of $\{d_t\}$, while the result of Yue et al. [16] has such a guarantee only for $q = 1$ and [8, 14] for $q > \frac{1}{2}(\sqrt{5}-1) \approx 0.618$. By contrast, our bound allows a wider range for q. When our analysis is applied to smooth optimization

problems, the allowed range is also wider than that in [5, Theorem 5]. For (2.20), $p = 1$ and $q \leq 1$ then imply that

$$\begin{cases} \rho > 0, \\ q > \frac{1}{2}, \\ \rho + q > 1, \\ 4q - 2\rho > 1, \end{cases}$$

so $q > 1/2$ with $\rho = 1/2$ implies Q-superlinear convergence for $\{r_t\}$ and R-superlinear convergence for $\{d_t\}$.

3. If $q = 1$ and $p > 0$, then any $\rho \in (0, p)$ implies superlinear convergence. In contrast, if $q = 1$, $p = \rho = 0$, the damping becomes a positive constant, making J_t uniformly bounded. Linear convergence could then by obtained from standard analysis of first-order-like variable metric approaches.

Inspired by the last item in Remark 2.6, we further demonstrate that the p-Hölder assumption on ∇A can be relaxed to merely *uniform continuity*, provided the damping and stopping tolerance decay sufficiently slowly. We present this result in Appendix A as its proof is technical but largely mirrors the analysis in this section.

3. A Line Search Strategy for Convex Regularized Optimization. In this section, we focus on (1.7), the special case of (1.1) obtained by setting $A = \nabla f$ and $B = \partial\Psi$. Although solutions of (1.1) in general only correspond to stationary points of (1.7), we make the further assumption that f is convex (similar to [11, 16]), which causes the solution sets of (1.1) and (1.7) to coincide in this case. In this scenario, A is also maximal monotone, so [4, Theorem 3.5] indicates that

$$r(x) \leq \operatorname{dist}(0, (A + B)(x)), \tag{3.1}$$

or equivalently $r(x) \leq \operatorname{dist}(0, \partial F(x))$ for (1.7), for all x. Therefore, the error bound (1.4) further implies a more intuitive local upper bound of $\operatorname{dist}(x, S)$ by ∂F. Moreover, [10, Theorem 3.4] shows that this bound, together with convexity, implies a growth condition that effectively upper bounds $\operatorname{dist}(x, S)$ by the objective gap. This characterization serves as a critical tool in our subsequent analysis.

We describe algorithms that attain global convergence to the solution set, and examine their rates of convergence. To simplify the description of our globalization strategy, we further assume that ∇f is globally L-Lipschitz continuous in its domain, not just locally Lipschitz, as in our discussion above.

We first discuss an existing algorithm due to [11], and argue that it guarantees global convergence and ensures local superlinear convergence of both r_t and d_t to 0 when the conditions in Theorem 2.3 or Theorem 2.4 hold. We then propose a novel strategy that retains these properties and adds another property, namely, strict decrease of the objective function at each iteration. Importantly, our new approach exhibits Q-superlinear convergence of the objective value to its optimal value, which we believe to be a new result even for nondegenerate problems. Existing analyses for ensuring local superlinear convergence in proximal-Newton-type methods require Lipschitz continuity of the Hessian of f and depend on a Taylor expansion to guarantee acceptance of the unit step size. We use instead a novel mechanism and analysis that relies on the convexity of F and the Hölderian error bound condition to guarantee sufficient function decrease for a unit step. Note that we do not need to assume that the Hessian $A(x) = \nabla^2 f(x)$ is Lipschitzian.

3.1. Two Algorithms. In our following discussion of both the existing approach and our new globalization approaches for (1.7), a tentative iterate $\tilde{x}_{t+1}$ is first obtained from the approximate minimization

$$\tilde{x}_{t+1} \approx \arg\min_x q_t(x), \tag{3.2}$$

with

$$\begin{aligned} q_t(x) &\coloneqq \langle g_t,\, x - x_t\rangle + \frac{1}{2}\langle H_t\,(x - x_t)\,,\, x - x_t\rangle + \Psi\,(x)\,, \\ g_t &\coloneqq \nabla f(x_t), \quad H_t \coloneqq J_t + \mu_t \mathrm{Id}, \quad \mu_t \coloneqq c r(x_t)^\rho, \end{aligned} \tag{3.3}$$

where $r(\cdot)$ follows the definition in (1.2) and (1.3), which can be written equivalently in this setting as follows:

$$R(x) = x - \operatorname{prox}_\Psi\left(x - \nabla f(x)\right), \quad r(x) = \|R(x)\|, \tag{3.4}$$

and J_t is a linear operator satisfying[3]

$$J_t \text{ symmetric}, \quad J_t \succeq 0, \quad \left\|J_t - \nabla^2 f(x_t)\right\| = O\big(r(x_t)^\theta\big) \text{ for some } \theta \geq \rho. \tag{3.5}$$

A line search procedure is applied to the update direction

$$\tilde{p}_t \coloneqq \tilde{x}_{t+1} - x_t \tag{3.6}$$

to find a suitable step size $\alpha_t > 0$; we then set $x_{t+1} \leftarrow x_t + \alpha_t \tilde{p}_t$. To ensure a descent direction and global convergence, the point $\tilde{x}_{t+1}$ from (3.2) is required to satisfy the conditions

$$q_t(\tilde{x}_{t+1}) \leq q_t(x_t), \quad \hat{r}_t(\tilde{x}_{t+1}) \leq \nu \min\left\{r(x_t)^{1+\rho}, r(x_t)\right\}, \quad \nu \in [0, 1). \tag{3.7}$$

In the setting of regularized optimization, we have that $\hat{r}_t(\cdot)$ in (1.11) is equivalent to

$$\hat{r}_t(\tilde{x}_{t+1}) = \|\hat{R}_t(\tilde{x}_{t+1})\|, \; \hat{R}_t\left(\tilde{x}_{t+1}\right) = \tilde{x}_{t+1} - \operatorname{prox}_\Psi\left(\tilde{x}_{t+1} - g_t - H_t\left(\tilde{x}_{t+1} - x_t\right)\right). \tag{3.8}$$

The conditions above are just specific realizations of (1.8), (1.9), (1.11), and (1.12) for the case of (1.7), except that (3.7) contains additional requirements on q_t, ν, and $r_t = r(x_t)$. Thus, the results that we derived for p_t in Section 2, such as (2.23), apply to $\tilde{p}_t$ as well. Moreover, we require J_t to be symmetric, which is natural in this setting, since it is an approximation of the Hessian $\nabla^2 f(x_t)$.

The first approach we consider, shown in Algorithm 3.1, is due to [11]. (Both this approach and its successor, Algorithm 3.2, require the conditions (3.7) to hold.)

Algorithm 3.1 Proximal Newton Method in [11]

input : $x_0 \in \mathcal{H}$, $\beta, \gamma, \sigma \in (0, 1)$, $\rho > 0$, $\nu \geq 0$, $c > 0$, $\bar{C} > F(x_0)$

1 $\eta \leftarrow r(x_0)$
2 **for** $t = 0, 1, \ldots$ **do**
3 Select H_t satisfying (3.3) and (3.5) with $\theta = 1$ and find an approximate solution $\tilde{x}_{t+1}$ of (3.2) satisfying (3.7)
4 $\tilde{p}_t \leftarrow \tilde{x}_{t+1} - x_t$
5 **if** $t > 0$, $r(\tilde{x}_{t+1}) \leq \sigma\eta$, *and* $F(\tilde{x}_{t+1}) \leq \bar{C}$ **then**
6 $\alpha_t \leftarrow 1$, $\eta \leftarrow r(\tilde{x}_{t+1})$
7 **else**
8 $\alpha_t \leftarrow \beta^{m_t}$, where m_t is the smallest nonnegative integer m such that

$$F\left(x_t + \beta^m \tilde{p}_t\right) \leq F\left(x_t\right) - \gamma\mu_t\beta^m\|\tilde{p}_t\|^2. \tag{3.9}$$

9 Set $x_{t+1} \leftarrow x_t + \alpha_t \tilde{p}_t$.

If the conditions of Theorems 2.3 and 2.4 hold, then when $r(x_t)$ is small enough, all subsequent iterations of Algorithm 3.1 will accept the unit step through the condition in Line 5 (no backtracking required), as Q-superlinearly convergence of $\{r(x_t)\}$ implies that this sequence also converges Q-linearly with an arbitrarily fast rate, and global convergence of $r(x_t)$ to 0 is guaranteed by [11, Theorem 3.1]. In the latter reference, local superlinear convergence requires local Lipschitz continuity of $\nabla^2 f$, so our analysis in Section 2 (which required Hölder continuity of $\nabla^2 f$) slightly broadens the problem class for which Algorithm 3.1 is superlinearly convergent. Theorem 2.3 provides additional guarantees for Q-superlinear convergence of $\{d_t\}$ not covered by [11]. Moreover, our requirement for $\theta \geq \rho$ in (3.5) is less restrictive than the choice $\theta = 1$ in [11].

Algorithm 3.1 does not require monotonic decrease of function values, as the conditions of Line 5 can be satisfied without having $F(\tilde{x}_{t+1}) < F(x_t)$. Our new approach, Algorithm 3.2, will ensure strict function decrease at each step while retaining Q-superlinear convergence of the function values.

The first distinctive element in Algorithm 3.2 is to replace the sufficient decrease condition (3.9) with

$$F(x_{t+1}) \leq F\left(x_t\right) - \gamma\alpha_t^2\|\tilde{p}_t\|^{2+\delta} \tag{3.10}$$

[3] For J_t to satisfy this definition, $\nabla^2 f$ needs to be positive semidefinite at any accumulation point $\bar{x}$ of the sequence $\{x_t\}$ that satisfies the stationarity condition $r(\bar{x}) = 0$. We refer to this property as "local convexity." Our algorithm can be extended to nonconvex problems by considering other choices of upper-bounded and positive-definite J_t, with the global convergence results of the next subsection continuing to hold. However, this local convexity assumption is necessary for superlinear convergence.

(where γ, $\tilde{p}_t$, and α_t are defined as in Algorithm 3.1, but x_{t+1} is defined differently; see below) for some given $\delta \geq 0$. Our analysis in the next subsection will show that when the conditions (2.16) are satisfied by p, q, and ρ, we can set $\delta = 2$. Since $\tilde{p}_t \neq 0$ for all t, the objective value is always decreasing. The second distinctive element of Algorithm 3.2, inspired by a technical result from [1, Lemma 2.3] on how $r(x)$ and $\mathrm{dist}(x, S)$ bound the objective gap, is to test the condition (3.10) on a point obtained from a proximal gradient step. In this vein, we first compute

$$y_{t+1}(\alpha_t) \leftarrow x_t + \alpha_t (\tilde{x}_{t+1} - x_t) = x_t + \alpha_t \tilde{p}_t, \tag{3.11}$$

and then compute a proximal gradient step from $y_{t+1}(\alpha_t)$ to obtain the candidate $\bar{x}_{t+1}(\alpha_t)$ for the next iterate, as follows

$$\begin{aligned} \bar{x}_{t+1}(\alpha_t) &\leftarrow \mathrm{prox}_{(\Psi/L)}\big(y_{t+1}(\alpha_t) - \tfrac{1}{L}\nabla f(y_{t+1}(\alpha_t))\big) \\ &= y_{t+1}(\alpha_t) - \tfrac{1}{L} G_L(y_{t+1}(\alpha_t)), \end{aligned} \tag{3.12}$$

where L is the Lipschitz constant for ∇f,

$$G_L(x) \coloneqq L\Big(x - \mathrm{prox}_{\frac{\Psi}{L}}\Big(x - \frac{1}{L}\nabla f(x)\Big)\Big) \tag{3.13}$$

is the proximal gradient, and prox_g denotes the proximal operator

$$\mathrm{prox}_g(x) \coloneqq \min_{y \in \mathcal{H}} \tfrac{1}{2}\|y - x\|^2 + g(y)$$

for any function g.[4]

The full algorithm can be specified as follows.

Algorithm 3.2 A Proximal Newton Method Guaranteeing Strict Decrease and Superlinear Convergence for the Objective Value

input : $x_0 \in \mathcal{H}$, $\beta \in (0,1)$, $\gamma \in (0,1)$, $\nu \in [0,1)$, $\rho \in (0,1]$, $c > 0$, $\delta \geq 0$, Lipschitz constant L for ∇f

1 **for** $t = 0, 1, \ldots$ **do**
2 Select H_t satisfying (3.3) and (3.5) with $\theta \geq \rho$ and find an approximate solution $\tilde{x}_{t+1}$ of (3.2) satisfying (3.7)
3 $\alpha_t \leftarrow 1, \quad \tilde{p}_t \leftarrow \tilde{x}_{t+1} - x_t$
4 Terminate if $\tilde{p}_t = 0$
5 **while** *True* **do**
6 Compute $y_{t+1}(\alpha_t)$ from (3.11) and $\bar{x}_{t+1}(\alpha_t)$ from (3.12)
7 **if** $F(\bar{x}_{t+1}(\alpha_t)) \leq F(x_t) - \gamma\alpha_t^2\|\tilde{p}_t\|^{2+\delta}$ **then**
8 $x_{t+1} \leftarrow \bar{x}_{t+1}(\alpha_t)$
9 Break
10 **else** $\alpha_t \leftarrow \beta\alpha_t$

The step taken at each iteration of Algorithm 3.2 is a composition of a prox-Newton step (with backtracking) and a short prox-gradient step (with the conservative choice $1/L$ for the steplength parameter). (One could instead use $\min\{\|\tilde{p}_t\|^2, \|\tilde{p}_t\|^{2+\delta}\}$ in Line 7 to make the acceptance condition easier to achieve in the early stage when $\|\tilde{p}_t\|$ is still large, and our analysis below still remains valid.)

3.2. Analysis. We show first that the line search criterion in Algorithm 3.2 is satisfied for all α_t sufficiently small (Lemma 3.1), and then deduce global convergence (Lemma 3.2). The first lemma does not require convexity of f; the second requires the property (3.5), which holds only if $\nabla^2 f$ is positive semidefinite at stationary accumulation points of the sequence $\{x_t\}$. From standard analysis of proximal gradient, we know that

$$F(\bar{x}_{t+1}(\alpha_t)) \leq F(y_{t+1}(\alpha_t)), \quad \forall \alpha_t \geq 0, \tag{3.14}$$

where $\bar{x}_{t+1}$ is defined in (3.12). This fact will play an important role in the following two results.

[4] For simplicity, we use a given (possibly rough) upper bound L on the actual Lipschitz constant here, but we can also conduct another backtracking on L to remove dependency on knowledge of this problem parameter. Our analysis still holds (with some additional calculations and notations) as long as L satisfies the following condition:

$$F(\bar{x}_{t+1}(\alpha_t)) \leq F(y_{t+1}(\alpha_t)) + \langle \nabla f(y_{t+1}(\alpha_t)), z_{t+1}(\alpha_t)\rangle + \frac{L}{2}\|z_{t+1}(\alpha_t)\|^2 + \Psi(\bar{x}_{t+1}(\alpha_t)),$$

where $z_{t+1}(\alpha_t) \coloneqq \bar{x}_{t+1}(\alpha_t) - y_{t+1}(\alpha_t)$.

LEMMA 3.1. *Given $\beta \in (0,1)$ and $\gamma \in (0,1)$, assume that f is L-Lipschitz-continuously differentiable for some $L > 0$ and that Ψ is convex, proper, and closed. At iteration t of Algorithm 3.2, the final value of α_t satisfying the sufficient decrease condition (3.10) satisfies the following condition:*

$$\text{either } \alpha_t = 1 \quad \text{or} \quad \alpha_t \geq \beta\mu_t\big(L + 2\gamma\|\tilde{p}_t\|^{\delta}\big)^{-1}. \tag{3.15}$$

Proof. Since f is L-Lipschitz-continuously differentiable, we have

$$f\left(x_t + \alpha_t\tilde{p}_t\right) \leq f\left(x_t\right) + \alpha_t\langle\nabla f(x_t),\, \tilde{p}_t\rangle + \frac{\alpha_t^2 L}{2}\|\tilde{p}_t\|^2.$$

Convexity of Ψ implies that

$$\Psi(x_t + \alpha_t\tilde{p}_t) - \Psi(x_t) \leq \alpha_t\left(\Psi(x_t + \tilde{p}_t) - \Psi(x_t)\right) = \alpha_t\left(\Psi\left(\tilde{x}_{t+1}\right) - \Psi\left(x_t\right)\right).$$

By summing these two inequalities, we obtain

$$\begin{aligned} F\left(x_t + \alpha_t\tilde{p}_t\right) &\leq F\left(x_t\right) + \alpha_t\left(\langle\nabla f(x_t),\, \tilde{p}_t\rangle + \Psi\left(\tilde{x}_{t+1}\right) - \Psi\left(x_t\right)\right) + \frac{\alpha_t^2 L}{2}\|\tilde{p}_t\|^2 \\ &\overset{(3.7)}{\leq} F\left(x_t\right) - \frac{\alpha_t}{2}\langle\tilde{p}_t,\, H_t\tilde{p}_t\rangle + \frac{\alpha_t^2 L}{2}\|\tilde{p}_t\|^2. \end{aligned} \tag{3.16}$$

From $J_t \succeq 0$, we have that $H_t \succeq \mu_t I$ and thus $-\langle\tilde{p}_t,\, H_t\tilde{p}_t\rangle \leq -\mu_t\|\tilde{p}_t\|^2$. Therefore, (3.16) and (3.14) lead to

$$F\left(\bar{x}_{t+1}\left(\alpha_t\right)\right) - F\left(x_t\right) \leq F\left(x_t + \alpha_t\tilde{p}_t\right) - F\left(x_t\right) \leq \frac{\|\tilde{p}_t\|^2}{2}\left(\alpha_t^2 L - \alpha_t\mu_t\right),$$

implying that (3.10) is satisfied provided that

$$\frac{\|\tilde{p}_t\|^2}{2}\left(\alpha_t^2 L - \alpha_t\mu_t\right) \leq -\alpha_t^2\gamma\|\tilde{p}_t\|^{2+\delta} \quad \overset{\alpha_t,\|\tilde{p}_t\|\geq 0}{\Longleftrightarrow} \quad \alpha_t \leq \mu_t\big(L + 2\gamma\|\tilde{p}_t\|^{\delta}\big)^{-1}.$$

The bound (3.15) is then obtained after considering the overshoot of backtracking. ∎

In the following result, we denote by $\mathcal{S}$ the set of stationary points for (1.7), that is, points x for which $r(x) = 0$. Such points are solutions of (1.1). This result additionally requires $\theta \leq 1$, but we note that this condition is more flexible than existing results such as those in [11], which require $\theta = 1$. The condition $\theta \leq 1$ together with (3.5) imply that we also need $\rho \leq 1$. From Item 3 of Remark 2.6, we see that when $q = 1$ in (1.4) (that is, when the Luo-Tseng error bound holds), since $p \leq 1$, then $\rho \leq 1$ is already a necessary condition for superlinear convergence. Thus, our requirement is not more restrictive than existing conditions in the literature. In the proof of the following lemma, we use a technical result originally from [16, (12)] to upper bound r_t by $\|\tilde{p}_t\|$. For completeness, we state and prove it in Appendix B.

LEMMA 3.2. *Consider the setting of Lemma 3.1 and assume that $\theta \leq 1$, $\delta \leq 2$, F is lower bounded by some $\bar{F} > -\infty$, and J_t satisfies (3.5) for all t. Then either Algorithm 3.2 terminates at some x_t for which $r(x_t) = 0$ (thus $x_t \in \mathcal{S}$), or else*

$$\lim_{t\to\infty} r(x_t) = 0. \tag{3.17}$$

If in addition $\mathcal{S}$ is nonempty, then any accumulation point of $\{x_t\}$ is in $\mathcal{S}$.

Proof. In the case of finite termination, it follows from the condition (3.7) that $\tilde{x}_{t+1} = x_t$ only if $r(x_t) = 0$. Otherwise, by summing (3.10) from $t = 0$ to $t = \infty$ with the understanding that $x_{t+1} = \bar{x}_{t+1}(\alpha_t)$, and by telescoping, we have that

$$\gamma\sum_{t=0}^{\infty}\alpha_t^2\|\tilde{p}_t\|^{2+\delta} \leq F\left(x_0\right) - \bar{F} \quad \Rightarrow \quad \lim_{t\to\infty}\alpha_t^2\|\tilde{p}_t\|^{2+\delta} = 0. \tag{3.18}$$

Let $\mathcal{K}_1$ be the subsequence of iterates t for which $\alpha_t = 1$ and $\mathcal{K}_2$ be the complementary set, for which $\alpha_t \geq \beta\mu_t\big(L + 2\gamma\|\tilde{p}_t\|^{\delta}\big)^{-1}$, by Lemma 3.1. We further partition $\mathcal{K}_2$ as $\mathcal{K}_2 = \mathcal{K}_{2a} \cup \mathcal{K}_{2b}$, where for $t \in \mathcal{K}_{2a}$ we have $L/(2\gamma) \geq \|p_t\|^{\delta}$, while for $t \in \mathcal{K}_{2b}$ we have $L/(2\gamma) < \|p_t\|^{\delta}$.

If $\mathcal{K}_1$ is infinite, we have from (3.18) that

$$\lim_{t\to\infty,\, t\in\mathcal{K}_1} \|\tilde{p}_t\| = 0. \tag{3.19}$$

From our assumption that f is L-Lipschitz continuous, we have that $\left\|\nabla^2 f\right\| \leq L$, which together with (3.3) and (3.5) indicates that

$$\|H_t\| \leq L + \mu_t + O(r_t^\theta).$$

Combination of this inequality with Lemma B.1 then implies that

$$(1-\nu)r_t \leq (\|H_t\|+2)\|\tilde{p}_t\| \leq (L + \mu_t + O(r_t^\theta) + 2)\|\tilde{p}_t\|. \tag{3.20}$$

Assume for contradiction that $\{r_t\}_{t\in\mathcal{K}_1}$ is not upper bounded, then there is a subsequence $\mathcal{K}_1^r \subset \mathcal{K}_1$ such that

$$\lim_{t\to\infty, t\in\mathcal{K}_1^r} r_t = \infty. \tag{3.21}$$

The equation above together with (3.20), the definition of μ_t in (3.3), and the condition $\theta \geq \rho$ in (3.5) implies that there is a constant $C \geq 0$ such that

$$r_t \leq C r_t^\theta \|\tilde{p}_t\|, \quad \forall t \in \mathcal{K}_1^r. \tag{3.22}$$

The conditions (3.21) and (3.22) together imply that

$$\begin{aligned} \infty = \lim_{t\to\infty, t\in\mathcal{K}_1^r} r_t^{1-\theta} &\leq \lim_{t\to\infty, t\in\mathcal{K}_1^r} C\|\tilde{p}_t\|, && \text{if } \theta < 1, \\ 1 \leq \lim_{t\to\infty, t\in\mathcal{K}_1^r} C\|\tilde{p}_t\|, & && \text{if } \theta = 1, \end{aligned}$$

a contradiction with (3.19). Therefore, we know that $\{r_t\}_{t\in\mathcal{K}_1}$ is upper bounded, and (3.19) and (3.20) thus imply that

$$\lim_{t\to\infty, t\in\mathcal{K}_1} r_t = 0. \tag{3.23}$$

Now we turn to the situation of $\mathcal{K}_2$ infinite. From (3.18), we have

$$\lim_{t\to\infty, t\in\mathcal{K}_2} \|\tilde{p}_t\|^{2+\delta} \frac{\mu_t^2}{4} \left(\frac{L}{2} + \gamma\|\tilde{p}_t\|^\delta\right)^{-2} = 0. \tag{3.24}$$

When the subsequence $\mathcal{K}_{2a}$ is infinite, we have from (3.24) that

$$\lim_{t\to\infty, t\in\mathcal{K}_{2a}} \|p_t\|^{2+\delta} \frac{\mu_t^2}{4L^2} = 0 \quad \Rightarrow \quad \lim_{t\to\infty, t\in\mathcal{K}_{2a}} \|p_t\|^{2+\delta} \mu_t^2 = 0. \tag{3.25}$$

By applying (3.20) to (3.25), we then obtain

$$\lim_{t\to\infty, t\in\mathcal{K}_{2a}} \frac{\mu_t^2 r_t^{2+\delta}}{\left(2 + L + \mu_t + O\left(r_t^\theta\right)\right)^{2+\delta}} = 0. \tag{3.26}$$

By inserting the definition (3.3) of μ_t into (3.26), we get

$$\lim_{t\to\infty, t\in\mathcal{K}_{2a}} \frac{c^2 r_t^{2\rho+2+\delta}}{\left(2 + L + c r_t^\rho + O\left(r_t^\theta\right)\right)^{2+\delta}} = 0. \tag{3.27}$$

If the sequence $\{r_t\}_{t\to\infty, t\in\mathcal{K}_{2a}}$ has an accumulation point at some positive finite value, the limit (3.27) cannot hold. If there is a subsequence approaching ∞, then because $1 \geq \theta \geq \rho$, the denominator in (3.27) is $O(r_t^{2+\delta})$ and since the numerator is $c^2 r_t^{2\rho+2+\delta}$, so the limit in (3.27) over this subsequence is infinite, contradicting(3.27). We therefore conclude that

$$\lim_{t\to\infty, t\in\mathcal{K}_{2a}} r_t = 0. \tag{3.28}$$

(To double check, we see that when (3.28) holds, $(2 + L + cr_t^\rho + O(r_t^\theta)) = O(1)$, and it indeed leads to (3.27) and thus (3.26).)

Suppose next that the subsequence $\mathcal{K}_{2b}$ is infinite. If $\delta = 2$, we have from (3.24) that

$$\lim_{t\to\infty, t\in\mathcal{K}_{2b}} \mu_t^2 = 0 \quad \overset{(3.3)}{\Rightarrow} \quad \lim_{t\to\infty, t\in\mathcal{K}_{2b}} r_t = 0. \tag{3.29}$$

For $\delta < 2$, from the lower-boundedness of $\|\tilde{p}_t\|$ in $\mathcal{K}_{2b}$, we see that

$$\lim_{t\to\infty, t\in\mathcal{K}_{2b}} \|\tilde{p}_t\|^{2-\delta}\mu_t^2 = 0 \quad \Rightarrow \lim_{t\to\infty, t\in\mathcal{K}_{2b}} \mu_t^2 = 0 \quad \Rightarrow \lim_{t\to\infty, t\in\mathcal{K}_{2b}} r_t = 0. \tag{3.30}$$

The desired result for the full sequence $\{r_t\}$ is thus obtained by combining (3.23) and (3.28)–(3.30). The claim that accumulation points are in $\mathcal{S}$ follows from the continuity of $r(\cdot)$. ∎

The previous two lemmas require no Hölder continuity of $\nabla^2 f$ nor the Hölderian error bound condition, and are therefore valid even if the problem does not fall in the class that allows for superlinear convergence in our previous analysis.

Next, we show in Theorem 3.3 that the unit step $\alpha_t = 1$ is eventually always accepted when the conditions of Theorem 2.3 hold and f is convex, so that Q-superlinear convergence of $\{d_t\}$ and $\{r_t\}$, defined in (2.1), is guaranteed, and the objective function converges superlinearly to its optimal value. Theorem 3.3 requires δ to be large enough such that $\|\tilde{p}_t\|^{2+\delta}$ is dominated by $d_t^{(q+1)/q}$ for all large t, or equivalently when $\tilde{p}_t$ is sufficiently small, where $\tilde{p}_t$ is the prox-Newton update step defined in (3.6). Noting the bound (2.23) (which, as noted above, applies to $\tilde{p}_t$ as well as to p_t), we need to select δ to satisfy

$$(2+\delta)\min\left\{1, 1+p-\frac{\rho}{q}\right\} > 1 + q^{-1}. \tag{3.31}$$

It suffices for this inequality that the following two conditions hold:

$$\delta > q^{-1} - 1, \tag{3.32a}$$

$$2 + \delta \geq \frac{q+1}{q}(1+p)q = (1+p)(1+q). \tag{3.32b}$$

To confirm that (3.32b) suffices, we have from (2.16) that

$$(2+\delta)\left(1+p-\frac{\rho}{q}\right) \geq (1+p)(1+q)\frac{1+pq-\rho}{q} > \frac{1+q}{q}.$$

When $q \leq 1$, since $p \in (0, 1]$, the condition (3.32b) is satisfied as long as $\delta \geq 2$. For condition (3.32a), we see from Remark 2.6 that when $p = 1$ we have $q^{-1} < 8/(-1+\sqrt{33}) < 1.7$. (If $p < 1$, the lower bound for q is larger, making q^{-1} even smaller.) Thus, it suffices for (3.32a) that $\delta \geq 0.7$. Setting $\delta = 2$ thus works for both conditions in (3.32).

THEOREM 3.3. *Consider (1.7) and assume that the settings of Theorem 2.3 hold, with $A(x) = \nabla f(x)$ and $B(x) = \partial\Psi(x)$, and in particular that the quantities $\rho \geq 0$, $p \in (0, 1]$ and $q > 0$ satisfy (2.16). Assume too that the settings of Lemma 3.1 hold and f is convex. Consider Algorithm 3.2 and let $F^* := \min F(x)$. If $\delta > 0$ satisfies $\|\tilde{p}_t\|^{2+\delta} = o(d_t^{(q+1)/q})$, for all t sufficiently large, then there is $t_0 \geq 0$ such that $\alpha_t = 1$ is accepted for all $t \geq t_0$ and, for $s > 0$ defined as in Theorem 2.3, we have*

$$\begin{cases} d_{t+1} = O\left(d_t^{1+s}\right), \\ r_{t+1} = O\left(r_t^{1+s}\right), \\ F\left(x_{t+1}\right) - F^* = O\big(\left(F\left(x_{t+1}\right) - F^*\right)^{1+s}\big), \end{cases} \qquad \forall t \geq t_0. \tag{3.33}$$

Proof. For purposes of this proof, we use the abbreviations

$$\bar{x}_{t+1} := \bar{x}_{t+1}(1), \quad y_{t+1} := y_{t+1}(1).$$

We first list two auxiliary results. From [1, Lemma 2.3], we have that

$$F\Big(x - \tfrac{1}{L}G_L(x)\Big) - F^* \leq \|G_L(x)\|^2\left(\frac{\operatorname{dist}(x,\mathcal{S})}{\|G_L(x)\|} - \frac{1}{2L}\right), \quad \forall x \in \mathcal{H}. \tag{3.34}$$

Moreover, [13, Lemma 3] indicates that there are constants $C_1 \geq C_2 > 0$ such that $r(x)$ and the norm of $G_L(x)$ are related by

$$C_1 r(x) \geq \|G_L(x)\| \geq C_2 r(x), \quad \forall x. \tag{3.35}$$

Now, using (3.34) with $x = y_{t+1}$ and $\bar{x}_{t+1} = y_{t+1} - \frac{1}{L} G_L(y_{t+1})$ as in (3.12), we have that

$$\begin{aligned} F(\bar{x}_{t+1}) - F^* &\leq \|G_L(y_{t+1})\| \mathrm{dist}(y_{t+1}, \mathcal{S}) \\ &\overset{(3.35)}{\leq} C_1 r(y_{t+1}) \mathrm{dist}(y_{t+1}, \mathcal{S}) \\ &\overset{(1.4)}{\leq} C_1 \kappa r(y_{t+1})^{1+q}. \end{aligned} \tag{3.36}$$

When $\alpha_t = 1$, we have $y_{t+1} = \tilde{x}_{t+1}$. Therefore, we can apply (2.19) to (3.36) and obtain

$$F(\bar{x}_{t+1}) - F^* \leq O\Big(d_t^{(1+q)\min\left\{1+\rho, 1+p, \left(1+p-\frac{\rho}{q}\right)(1+p)\right\}}\Big) = O\big(d_t^{(1+s)(1+q)/q}\big), \tag{3.37}$$

where we used the definition of s from Theorem 2.3. On the other hand, (3.1) (due to convexity of f) and (1.4) imply that there is $\epsilon_2 > 0$ such that

$$\mathrm{dist}(x, \mathcal{S}) \leq \kappa \|z\|^q, \quad \forall z \in \partial F(x), \quad \forall x \in \{x \mid \mathrm{dist}(x, \mathcal{S}) \leq \epsilon_2\},$$

which together with convexity of F implies that

$$F(x) - F^* \leq \kappa \min_{z \in \partial F(x)} \|z\|^{1+q}, \quad \forall x \in \{x \mid \mathrm{dist}(x, \mathcal{S}) \leq \epsilon_2\}. \tag{3.38}$$

We then apply [10, Theorem 3.4] to (3.38) to conclude that there is $\kappa_2 > 0$ such that

$$F(x) - F^* \geq \kappa_2 \mathrm{dist}(x, \mathcal{S})^{\frac{1+q}{q}} \tag{3.39}$$

holds for all x close enough to $\mathcal{S}$. Since $\|\tilde{p}_t\|^{2+\delta} = o(d_t^{(1+q)/q})$, we have from (3.39) that

$$F(x_t) - F^* - \sigma \|\tilde{p}_t\|^{2+\delta} \geq \Omega\Big(d_t^{\frac{1+q}{q}}\Big) + o\Big(d_t^{\frac{1+q}{q}}\Big) = \Omega\Big(d_t^{\frac{1+q}{q}}\Big). \tag{3.40}$$

By comparing (3.37) and (3.40), and using the fact that (2.16) implies $s > 0$, we have for d_t sufficiently small that

$$F(\bar{x}_{t+1}) - F^* = o\big(F(x_t) - F^* - \sigma \|\tilde{p}_t\|^{2+\delta}\big).$$

Therefore, from (3.17) and (1.4), we have for all sufficiently large t that

$$F(\bar{x}_{t+1}) - F(x_t) + \sigma \|\tilde{p}_t\|^{2+\delta} = F(\bar{x}_{t+1}) - F^* - \big(F(x_t) - F^* - \sigma \|\tilde{p}_t\|^{2+\delta}\big) \leq 0,$$

meaning that the unit step size is accepted and we have $x_{t+1} = \bar{x}_{t+1}$. From convexity of f and [6, (34)], a short proximal-gradient step does not increase distance to the solution, so we have $d_{t+1} = \mathrm{dist}(x_{t+1}, \mathcal{S}) = \mathrm{dist}(\bar{x}_{t+1}, \mathcal{S}) \leq \mathrm{dist}(y_{t+1}, \mathcal{S})$. We can now apply (2.17), noting that d_{t+1} in that bound corresponds to $\mathrm{dist}(y_{t+1}, \mathcal{S})$ here because $\alpha_t = 1$ (while our d_{t+1} in this section corresponds to $\mathrm{dist}(x_{t+1}, \mathcal{S})$), to obtain

$$\underbrace{\mathrm{dist}(x_{t+1}, \mathcal{S})}_{d_{t+1} \text{ in this section}} \leq \underbrace{\mathrm{dist}(y_{t+1}, \mathcal{S})}_{d_{t+1} \text{in (2.17)}} = O(d_t^{1+s}) = o(d_t). \tag{3.41}$$

This proves the first claim in (3.33). When d_t is small enough, we further have from (3.41) that $d_{t+1} \leq d_t$ so that superlinear convergence and acceptance of the unit step propagates to the next iteration. We note that there must be t such that d_t is small enough to satisfy our requirement according to Lemma 3.2 and (1.4).

Superlinear convergence of the objective follows from (3.39) and (3.37). In particular, using $x_{t+1} = \bar{x}_{t+1}$ in (3.37), we see that

$$F(x_{t+1}) - F^* \leq O\left(d_t^{(1+s)(1+q)/q}\right) \overset{(3.39)}{=} O\big(\big(F(x_t) - F^*\big)^{1+s}\big).$$

Finally, to prove the convergence rate for r_t, we use the definition (3.4) and nonexpansiveness of prox_Ψ to obtain

$$\begin{aligned} &\|R\left(x_{t+1}\right) - R\left(y_{t+1}\right)\| \\ &= \|(x_{t+1} - y_{t+1}) - (\text{prox}_\Psi(x_{t+1} - \nabla f(x_{t+1})) - \text{prox}_\Psi(y_{t+1} - \nabla f(y_{t+1})))\| \\ &\le \|(x_{t+1} - y_{t+1}\| + \|(x_{t+1} - \nabla f(x_{t+1})) - (y_{t+1} - \nabla f(y_{t+1}))\| \\ &\le (2+L)\|x_{t+1} - y_{t+1}\|. \end{aligned} \tag{3.42}$$

Using from (3.12) that $x_{t+1} = \bar{x}_{t+1} = y_{t+1} - (1/L)G_L(y_{t+1})$, and using (3.35), we can bound $\|x_{t+1} - y_{t+1}\|$ by

$$\|x_{t+1} - y_{t+1}\| = \frac{1}{L}\|G_L(y_{t+1})\| \in \left[\frac{C_2}{L} r\left(y_{t+1}\right), \frac{C_1}{L} r\left(y_{t+1}\right)\right],$$

where $C_1 \ge C_2 > 0$. By substituting into (3.42), we obtain

$$\begin{aligned} r_{t+1} = \|R(x_{t+1})\| &\le r(y_{t+1}) + (2+L)\|x_{t+1} - y_{t+1}\| \\ &\le \Big(1 + \frac{(2+L)C_1}{L}\Big) r(y_{t+1}) \\ &= O\left(r(y_{t+1})\right) \\ &= O\left(r_t^{1+s}\right), \end{aligned}$$

(where the last step is from (2.17) in Theorem 2.3), proving (3.33). With the same argument for d_t, we see that (1.4) continues to hold for the next iterate when r_t is small enough to ensure $r_{t+1} \le r_t$. ∎

Finally, we show that when $q = 1$ in (1.4), our linesearch is versatile enough to accept the unit step size for *any* update directions that yield Q-superlinear convergence of both d_t and r_t, regardless of how these directions are generated.

COROLLARY 3.4. *Consider (1.7) and assume that the settings of Lemma 3.1 hold, f is convex, and (1.4) holds with $q = 1$. Consider Algorithm 3.2 but with $\{\tilde{x}_{t+1}\}$ generated in an arbitrary manner that satisfies*

$$\text{dist}(\tilde{x}_{t+1}, \mathcal{S}) = o(d_t) \quad \text{and} \quad r(\tilde{x}_{t+1}) = o(r_t). \tag{3.43}$$

Assume too that the parameter $\delta > 0$ in Algorithm 3.2 satisfies $\|\tilde{p}_t\|^{2+\delta} = o(d_t^2)$. Then for all t sufficiently large, we have $\alpha_t = 1$, and we have

$$d_{t+1} = o(d_t), \quad r_{t+1} = o(r_t), \quad F(x_{t+1}) - F^* = o(F(x_t) - F^*).$$

Proof. The proof mainly follows the argument for Theorem 3.3 with a few differences noted here. First, due to (3.43) and $q = 1$, and noting that $r_t = O(d_t)$ from (2.13), (3.37) becomes

$$F(\bar{x}_{t+1}) - F^* \le o(d_t^2).$$

On the other hand, our assumption that $\|\tilde{p}_t\|^{2+\delta} = o(d_t^2)$ and (3.39) with $q = 1$ imply (3.40) with the right-hand side being $\Omega(d_t^2)$. By comparing with the inequality above, we conclude that the unit step size is accepted for t sufficiently large. The claim of superlinear convergence then follows by the same reasoning as in the remainder of the argument in Theorem 3.3, with d_t^{1+s}, $(F(x_t) - F^*)^{1+s}$), and r_t^{1+s} replaced by $o(d_t)$, $o(F(x_t) - F^*)$, and $o(r_t)$, respectively. ∎

From the proofs of Theorem 3.3 and Corollary 3.4, we can see that the analysis relies solely on the convergence rate of the update direction rather than the existence or continuity of the Hessian $\nabla^2 f$. Therefore, our novel line search serves as a general framework that is also compatible with any "fast directions" obtained beyond the proximal Newton approach analyzed in Theorem 2.3 and Corollary A.1, such as through a (proximal) semismooth Newton or a proximal quasi-Newton approach, to simultaneously ensure strict objective decrease and eventual unit step size acceptance for fast convergence.

4. Simplification For Smooth Problems. Our algorithm can be simplified for the case of convex smooth optimization — the setting of Section 3 with $\Psi \equiv 0$, that is, $F(x) = f(x)$. We have in this scenario that $R(x) = \nabla f(x)$ and $G_L(x) = \nabla f(x)$ for any $L > 0$, and the bound (3.34) simplifies to

$$F(x) - F^* \le \|\nabla f(x)\| \text{dist}(x, \mathcal{S}) \tag{4.1}$$

Algorithm 4.1 A Simple Newton Method for Degenerate Problems

input : $x_0 \in \mathcal{H}$, $\beta, \gamma \in (0,1)$, $\nu \in [0,1)$, $c > 0, \rho \in (0,1]$, $\delta \geq 0$

for $t = 0, 1, \ldots$ **do**

  Select H_t satisfying (3.3) and (3.5) and find an approximate solution $\tilde{x}_{t+1}$ of (3.2) (with $\Psi = 0$) satisfying (3.7)

  $\tilde{p}_t \leftarrow \tilde{x}_{t+1} - x_t$, $\alpha_t \leftarrow 1$

  while $F(x_t + \alpha_t \tilde{p}_t) > F(x_t) - \gamma \alpha_t^2 \|\tilde{p}_t\|^{2+\delta}$ **do**

    $\alpha_t \leftarrow \beta \alpha_t$

  $x_{t+1} \leftarrow x_t + \alpha_t \tilde{p}_t$

from convexity. Therefore, the proximal gradient step can be removed from Algorithm 3.2 without affecting the bounds on the objective value. The simplified method is shown as Algorithm 4.1.

Clearly, this is identical to the classical truncated Newton method except for the addition of a damping term to the quadratic approximation and the slightly unconventional step size acceptance criterion. As the proofs of Lemmas 3.1 and 3.2 do not involve any specific properties of the proximal gradient step, we see that they are still applicable to Algorithm 4.1. The local convergence result is as follows.

COROLLARY 4.1. *Assume that $\Psi \equiv 0$ and $f \in \mathcal{C}^2$ is convex and L-Lipschitz-continuously differentiable in (1.7) with $\nabla^2 f$ p-Hölder continuous in a neighborhood U of $\mathcal{S}$, and within U, (1.4) holds for some $\kappa > 0$ and some $q \in (0,1]$. Then for Algorithm 4.1, we have that (3.15) and (3.17) both hold. Further, if (2.16) is satisfied, then there is $t_0 \geq 0$ such that $\alpha_t = 1$ for all $t \geq t_0$, and (3.33) holds.*

Proof. The claims for (3.15) and (3.17) follow directly from the same reasoning as before, noting that $x_{t+1} = y_{t+1}$. For the superlinear convergence claim, we see from (1.4) (recalling that $R(x) = \nabla f(x)$ in this case) and (4.1) that the following condition holds.

$$F(x) - F^* \leq \kappa \|\nabla f(x)\|^{1+q}. \tag{4.2}$$

Theorem 3.4 of [10] then indicates that (3.39) holds for some $\kappa_2 > 0$ near $\mathcal{S}$. Finally, using the argument in the proof of Theorem 3.3, with (3.34) replaced by (4.1) for $x_{t+1} = y_{t+1}$, the desired results in (3.33) follow. ∎

Recent works [3, 9] consider a specific damping in the form of (3.3) with $\rho = 1/2$ and $J_t = \nabla^2 f(x_t)$ for unconstrained convex optimization. Under the assumption that $\nabla^2 f$ is globally H-Lipschitz continuous, these two works showed that by fixing c to a specific value in (3.3) and solving the subproblem exactly, $\alpha_t \equiv 1$ can be used and global convergence of the objective value to the optimum is guaranteed, with a speed of $O(t^{-2})$. They also showed local superlinear convergence under an additional global strong convexity condition (global strong convexity, continuous Hessian, and a descent algorithm imply that the gradient is Lipschitz continuous in the region of interest). However, we notice that their analyses utilized the second-order Taylor approximation of f and the Lipschitz continuity of the Hessian, and thus are not applicable to the problem we consider here, where we do not assume Lipschitz continuity of $\nabla^2 f$. Therefore, Corollary 4.1 extends the problem class on which their damped truncated Newton leads to superlinear convergence to non-strongly-convex ones with non-Lipschitz Hessians.

5. Conclusions. We have examined inexact, damped, proximal Newton-like methods for solving degenerate regularized optimization problems, and their extension to the generalized equations setting. We have described algorithms that achieved superlinear convergence even in the presence of nonconvexity of the smooth term and singularity of its Hessian (or nonmonotonicity, and singularity of the Jacobian, in the case of generalized equations). Moreover, we show that the standard assumptions of Lipschitz continuity of the Hessian (or Jacobian) and the Lipschitz error bound can be relaxed to Hölderian ones, and we can further relax the Hölder continuity assumption of the Hessian (or Jacobian) to uniform continuity. These results require careful choices of the parameters that govern the damping and the measure of inexactness in the solution of each subproblem.

Acknowledgements. We thank Defeng Sun for fruitful discussion and pointing out an error in an early draft.

References.

[1] A. BECK AND M. TEBOULLE, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM Journal on Imaging Sciences, 2 (2009), pp. 183–202.

[2] R. A. DEVORE AND G. G. LORENTZ, *Constructive Approximation*, vol. 303 of Grundlehren der mathematischen Wissenschaften, Springer-Verlag, Berlin, Heidelberg, 1993.

[3] N. DOIKOV AND Y. NESTEROV, *Gradient regularization of Newton method with Bregman distances*, Mathematical Programming, 204 (2024), pp. 1–25.

[4] D. Drusvyatskiy and A. S. Lewis, *Error bounds, quadratic growth, and linear convergence of proximal methods*, Mathematics of Operations Research, 43 (2018), pp. 919–948.
[5] C.-p. Lee, *Accelerating inexact successive quadratic approximation for regularized optimization through manifold identification*, Mathematical Programming, 201 (2023), pp. 599–633.
[6] C.-p. Lee and S. J. Wright, *First-order algorithms converge faster than $O(1/k)$ on convex problems*, in Proceedings of the 36th International Conference on Machine Learning, 2019.
[7] J. D. Lee, Y. Sun, and M. A. Saunders, *Proximal Newton-type methods for minimizing composite functions*, SIAM Journal on Optimization, 24 (2014), pp. 1420–1443.
[8] R. Liu, S. Pan, Y. Wu, and X. Yang, *An inexact regularized proximal Newton method for nonconvex and nonsmooth optimization*, Computational Optimization and Applications, (2024), pp. 1–39.
[9] K. Mishchenko, *Regularized Newton method with global $O(1/k^2)$ convergence*, SIAM Journal on Optimization, 33 (2023), pp. 1440 – 1462.
[10] B. S. Mordukhovich and W. Ouyang, *Higher-order metric subregularity and its applications*, Journal of Global Optimization, 63 (2015), pp. 777–795.
[11] B. S. Mordukhovich, X. Yuan, S. Zeng, and J. Zhang, *A globally convergent proximal Newton-type method in nonsmooth convex optimization*, Mathematical Programming, 198 (2023), pp. 899–936.
[12] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, second ed., 2006.
[13] P. Tseng and S. Yun, *A coordinate gradient descent method for nonsmooth separable minimization*, Mathematical Programming, 117 (2009), pp. 387–423.
[14] S. vom Dahl and C. Kanzow, *An inexact regularized proximal Newton method without line search*, Computational Optimization and Applications, (2024), pp. 1–40.
[15] X. Xiao, Y. Li, Z. Wen, and L. Zhang, *A regularized semi-smooth Newton method with projection steps for composite convex programs*, Journal of Scientific Computing, 76 (2018), pp. 364–389.
[16] M.-C. Yue, Z. Zhou, and A. M.-C. So, *A family of inexact SQA methods for non-smooth convex minimization with provable convergence guarantees based on the Luo–Tseng error bound property*, Mathematical Programming, 174 (2019), pp. 327–358.

Appendix A. Relaxation from p-Hölder Continuity of the Hessian to Uniformly Continuous. This appendix shows that the p-Hölder assumption (1.5) on ∇A in section 2 can be further relaxed. In particular, for the case of $q = 1$, superlinear convergence can be preserved even when the Jacobian is merely *uniformly continuous*, provided that the damping and the stopping tolerance decay slowly enough. For this result, we note that for any function f uniformly continuous in a convex set V, it admits a modulus of continuity $\omega : [0, \infty) \to [0, \infty)$ such that

$$\lim_{s \downarrow 0} \omega(s) = \omega(0) = 0, \quad \|f(x) - f(y)\| \le \omega(\|x - y\|), \quad \forall x, y \in V. \tag{A.1}$$

It is known that we can always select an ω that is continuous, monotonically increasing, and subadditive. See, for example, [2, Chapter 2, Section 6].

Subadditivity and monotonicity of ω play a crucial role in our analysis below. More specifically, subadditivity and monotonicity of ω imply that

$$f(t) = O(g(t)) \quad \Rightarrow \quad \omega(f(t)) = O(\omega(g(t)). \tag{A.2}$$

Indeed, by definition, $f(t) = O(g(t))$ means there is $\beta > 0$ such that $f(t) \le \beta g(t)$ for all t large, and thus monotonicity of ω implies that

$$\omega(f(t)) \le \omega(\beta g(t)) \le \omega(\lceil \beta \rceil g(t)),$$

while subadditivity of ω indicates

$$\omega(\lceil \beta \rceil g(t)) \le \lceil \beta \rceil \omega(g(t)).$$

These two inequalities in combination then lead to

$$\omega(f(t)) \le \lceil \beta \rceil \omega(g(t)) = O(\omega(g(t)),$$

which is exactly (A.2).

Corollary A.1. *Consider (1.1) with the same assumptions as Theorem 2.3, except that (1.4) holds with $q = 1$ and ∇A is only uniformly continuous in V. Let ω_1 denote the nondecreasing and subadditive modulus of continuity of ∇A. Consider the update scheme (1.8), but with μ_t, J_t, and the stopping condition (1.12) replaced by*

$$\mu_t = c\,\omega_2(r_t), \quad \|J_t - \nabla A(x_t)\| = O(\omega_2(r_t)), \quad \hat{r}_t(x_{t+1}) \le \nu \omega_2(r_t) r_t \tag{A.3}$$

for some $\nu \geq 0$ and some given nondecreasing, subadditive, and continuous function $\omega_2 : [0,\infty) \to [0,\infty)$ that vanishes at zero. If the problem satisfies

$$\omega_1(s) \leq \beta_1 \omega_2(s) \tag{A.4}$$

for some $\beta_1 > 0$ and all $s \geq 0$, then, provided r_0 is sufficiently small, we have $r_{t+1} = o(r_t)$, $d_{t+1} = o(d_t)$, and $\{x_t\}$ converges strongly to a point in the solution set $\mathcal{S}$.

Proof. We use the same notations as (2.1) such that

$$d_t := \operatorname{dist}(x_t, \mathcal{S}), \quad r_t := r(x_t), \quad p_t := x_{t+1} - x_t,$$

and $\bar{x}_t \in P_{\mathcal{S}}(x_t)$, where $\mathcal{S}$ is the set of solutions. We observe first from (A.1) that uniform continuity of ∇A implies that

$$\begin{aligned} \|A(x) - A(y) - \nabla A(y)(x-y)\| &= \left\| \int_0^1 \nabla A(y + t(x-y))(x-y)\, dt - \nabla A(y)(x-y) \right\| \\ &\leq \int_0^1 \omega_1(t\|x-y\|)\, dt\, \|x-y\| \\ &\leq \omega_1(\|x-y\|)\|x-y\|. \end{aligned} \tag{A.5}$$

Second, since $q = 1$, (1.4) and (2.13) indicate $r_t = O(d_t)$ and $d_t = O(r_t)$. From the definitions (1.11) and (2.4), we have $\|\xi_t\| = \hat{r}(x_{t+1})$ and so (A.3) implies that $\|\xi_t\| \leq \nu\omega_2(r_t)r_t$. Since $\omega_2(r_t) \to 0$ as $r_t \to 0$ and $r_t = O(d_t)$, we have

$$\|\xi_t\| = o(r_t) = o(d_t). \tag{A.6}$$

We now find a bound on $\|\bar{x}_t - x_{t+1} + \xi_t\|$, using (2.8) again in a similar way to how we derived (2.9). Substituting the parameter choices from (A.3) into (2.8), and using (A.5) with $x = \bar{x}_t$ and $y = x_t$, we obtain:

$$\begin{aligned} &\|\bar{x}_t - x_{t+1} + \xi_t\| \\ &\leq \mu_t^{-1}\Big(\|(J_t - \nabla A(x_t))(\bar{x}_t - x_t)\| + \|A(x_t) - A(\bar{x}_t) - \nabla A(x_t)(x_t - \bar{x}_t)\| \\ &\qquad + \mu_t\|\bar{x}_t - x_t\| + (1 + \|H_t\|)\|\xi_t\| \Big) \\ &\leq \frac{1}{c\omega_2(r_t)} \Big(O(\omega_2(r_t))d_t + \omega_1(d_t)d_t + c\omega_2(r_t)d_t + O(1)\cdot \nu\omega_2(r_t)r_t \Big). \end{aligned}$$

In the last inequality, we used $\|H_t\| = O(1)$, which is from

$$\begin{aligned} \|H_t\| &\overset{(1.8),(A.3)}{=} \|\mu_t I + \nabla A(x_t) + O(\omega_2(r_t))\| \\ &\overset{(A.3)}{\leq} O(\omega_2(r_t)) + \|\nabla A(x_t)\| \\ &\overset{(2.3)}{\leq} O(\omega_2(r_t)) + L = O(1). \end{aligned} \tag{A.7}$$

Using (A.4) and the fact that $d_t = O(r_t)$, we have from (A.2) and monotonicity and subadditivity of ω_2 that $\omega_2(d_t) = O(\omega_2(r_t))$. Thus, the numerator is dominated by $O(\omega_2(r_t)d_t)$ since $r_t = O(d_t)$ according to (2.13). Consequently, the damping term $\omega_2(r_t)$ in the denominator cancels out, yielding:

$$\|\bar{x}_t - x_{t+1} + \xi_t\| \leq O(d_t).$$

By combining the above inequality with (A.6), and using the definition $\bar{x}_t = P_{\mathcal{S}}(x_t)$ and (2.1), we obtain

$$\|p_t\| \leq \|\bar{x}_t - x_{t+1} + \xi_t\| + \|x_t - \bar{x}_t\| + \|\xi_t\| = O(d_t) + d_t + o(d_t) = O(d_t). \tag{A.8}$$

We now proceed to obtain an upper bound for r_{t+1} similar to (2.10) by following the proof of Lemma 2.2. We know that (2.11) and (2.12) still hold as they do not involve elements changed in this corollary, so they indicate

$$\begin{aligned} r_{t+1} &\leq \|A(x_t) - A(x_{t+1}) - H_t(x_t - x_{t+1}\|) + \hat{r}_t(x_{t+1}) \\ &\overset{(A.7),(A.3)}{\leq} \|A(x_t) - A(x_{t+1}) - \nabla A(x_t)(x_{t+1} - x_t)\| + O(\omega_2(r_t))\|p_t\| + O(\omega_2(r_t)r_t) \\ &\overset{(A.5)}{\leq} \omega_1(\|p_t\|)\|p_t\| + O(\omega_2(r_t)(\|p_t\| + r_t)). \end{aligned}$$

From (1.4) and (A.8), we know that $\|p_t\| = O(r_t)$, and from (A.4) and (A.2) resulted from subadditivity and monotonicity of ω_1 and ω_2, we conclude that $\omega_1(\|p_t\|)\|p_t\| = O(\omega_2(r_t)r_t)$. Therefore, the inequality above simplifies to

$$r_{t+1} \leq O(\omega_2(r_t)r_t) = o(r_t)$$

as ω_2 vanishes at zero. Since $d_{t+1} = O(r_{t+1})$ and $r_t = O(d_t)$, the inequality above also shows that $d_{t+1} = o(d_t)$, proving the claimed superlinear convergence. Convergence of the sequence $\{x_t\}$ then follows from the argument of Theorem 2.5 using $\|p_t\| = O(r_t)$. ∎

The case of p-Hölder continuity corresponds to Corollary A.1 with $\omega_1(t) = \zeta t^p$. However, for this special case, Theorems 2.3 and 2.4 provide more refined results with explicit rates.

Comparing Corollary A.1 with Theorems 2.3 and 2.4 reveals an inherent trade-off in the decay rate of the damping term and the stopping tolerance. To ensure robustness against the lack of smoothness (small p or uniform continuity), these parameters must decay slowly. Conversely, to accommodate a wider range of the error bound exponent, they must vanish rapidly.

Appendix B. A Technical Lemma.

LEMMA B.1. *Consider the setting of Lemma 3.2. For all $t \geq 0$, we have*

$$(1-\nu)r(x_t) \leq (\|H_t\| + 2)\|\tilde{p}_t\|.$$

Proof. Since $\nu < 1$ and $\hat{r}_t(\tilde{x}_{t+1}) \leq \nu r(x_t)$ in (3.7), we obtain from (3.4), (3.7), and (3.8), the triangle inequality, and the nonexpansiveness of prox_Ψ due to convexity of Ψ in the fourth inequality that

$$\begin{aligned}
(1-\nu)r_t &\leq r_t - \hat{r}_t(\tilde{x}_{t+1}) \\
&\leq \left\| R_t(x_t) - \hat{R}_t(\tilde{x}_{t+1}) \right\| \\
&\leq \|x_t - \tilde{x}_{t+1}\| + \|\text{prox}_\Psi(\tilde{x}_{t+1} - g_t - H_t(\tilde{x}_{t+1} - x_t)) - \text{prox}_\Psi(x_t - g_t)\| \\
&\leq \|x_t - \tilde{x}_{t+1}\| + \|(\tilde{x}_{t+1} - g_t - H_t(\tilde{x}_{t+1} - x_t)) - (x_t - g_t)\| \\
&= \|x_t - \tilde{x}_{t+1}\| + \|(\tilde{x}_{t+1} - x_t - H_t(\tilde{x}_{t+1} - x_t))\| \\
&\leq (2 + \|H_t\|)\|\tilde{x}_{t+1} - x_t\| \\
&= (2 + \|H_t\|)\|\tilde{p}_t\|,
\end{aligned}$$

proving the stated result. ∎