LOCAL CONVERGENCE OF ADAPTIVELY REGULARIZED TENSOR METHODS*

KARL WELZEL†, YANG LIU†, RAPHAEL A. HAUSER†, AND CORALIA CARTIS†

Abstract. Optimization methods that make use of derivatives of the objective function up to order $p > 2$ are called tensor methods. Among them, ones that minimize a regularized pth-order Taylor expansion at each step have been shown to possess optimal global complexity which improves as p increases. The local convergence of such optimization algorithms on functions that have Lipschitz continuous pth derivatives and are uniformly convex of order q has been studied by Doikov and Nesterov [*Math. Program.*, 193 (2022), pp. 315–336]. We extend these local convergence results to locally uniformly convex functions and fully adaptive methods, which do not need knowledge of the Lipschitz constant, thus providing the first sharp local rates for ARp. We discuss the surprising new challenges encountered by nonconvex local models and non-unique model minimizers. For $p > 2$, our examples show that in particular when using the global minimizer of the subproblem, even asymptotically not all iterations need to be successful. Only if the "right" local model minimizer is used, the $p/(q-1)$th-order local convergence from the non-adaptive case is preserved for $p > q-1$, otherwise the superlinear rate can degrade. We thus confirm that adaptive higher-order methods achieve superlinear convergence for certain degenerate problems as long as p is large enough and provide sharp bounds on the order of convergence one can expect in the limit.

Key words. tensor methods, higher-order optimization, adaptive regularization, local convergence, uniform convexity

MSC codes. 90C26, 65K05

1. Introduction.

Tensor methods are a class of methods for unconstrained continuous optimization that can incorporate derivatives of order 1 to p of the objective function $f \in C^p(\mathbb{R}^n)$ in order to find a local minimizer of f. The most important method in this regard is the ARp method. It works by constructing the local model at each iterate $\boldsymbol{x}_k$ as the sum of the pth-order Taylor expansion at $\boldsymbol{x}_k$ and a regularization term $\sigma_k\|\boldsymbol{y}-\boldsymbol{x}_k\|^{p+1}$. This model is then minimized to find a potential new iterate $\boldsymbol{y}_k$. If the function decrease at the new iterate is sufficient, the iteration is called "successful", the iterate is accepted ($\boldsymbol{x}_{k+1} = \boldsymbol{y}_k$) and the regularization parameter σ_k is decreased; otherwise the iteration is called "unsuccessful", the iterate is rejected ($\boldsymbol{x}_{k+1} = \boldsymbol{x}_k$) and σ_k is increased. This **a**daptively chosen **r**egularization parameter gives the ARp method its name [2, 8].

The advantage of having access to higher derivatives can be quantified in two different ways on a theoretical level: in terms of *global* complexity guarantees, giving an upper bound on the number of iterations until an approximate stationary point is found, and in terms of *local* convergence rates when the algorithm is already close to the solution. For both cases it is standard to consider the class of p-times differentiable functions that have a Lipschitz continuous pth derivative and are lower bounded. On these functions the *global* complexity literature established that the ARp method has optimal worst-case global iteration complexity among all possible pth-order methods as it needs at most $O(\varepsilon^{-(p+1)/p})$ iterations to find an approximate stationary point with a gradient norm bounded by $\varepsilon > 0$ [2, 7, 8]. Clearly, this bound improves as p

*Received by the editors DATE.

Funding: This work is supported by the Hong Kong Innovation and Technology Commission (InnoHK Project CIMDA)

†Mathematical Institute, University of Oxford, Woodstock Road, Oxford OX2 6GG, United Kingdom (karl.welzel@maths.ox.ac.uk, yang.liu@maths.ox.ac.uk, raphael.hauser@maths.ox.ac.uk, coralia.cartis@maths.ox.ac.uk).

increases.

The *local* convergence for these tensor methods has been analysed for the subclass of uniformly convex functions by Doikov and Nesterov [11]. A differentiable function is uniformly convex of degree $q \geq 2$ if there exists a $\mu > 0$ such that $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) + \mu\|\boldsymbol{y} - \boldsymbol{x}\|^q$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. This means that f is strictly convex and grows at least like $O(\|\boldsymbol{x} - \boldsymbol{x}_*\|^q)$ around any stationary point $\boldsymbol{x}_*$. Doikov and Nesterov [11] additionally assume that the function's pth derivative is Lipschitz continuous with constant L_p and that the regularization parameter σ_k is constant and sufficiently large ($\sigma_k \geq \frac{pL_p}{(p+1)!}$). The latter condition allows them to analyse a non-adaptive version of ARp where $\boldsymbol{y}_k$ is the global minimizer of the local model at each step, and it is always accepted. For this method the authors obtain

$$f(\boldsymbol{x}_{k+1}) - f^* = O\Big((f(\boldsymbol{x}_k) - f^*)^{\frac{p}{q-1}}\Big), \tag{1.1a}$$

$$\|\nabla f(\boldsymbol{x}_{k+1})\| = O\Big(\|\nabla f(\boldsymbol{x}_k)\|^{\frac{p}{q-1}}\Big). \tag{1.1b}$$

where f^* is the global minimum of f. Equations (1.1a) and (1.1b) imply that, if $\boldsymbol{x}_k$ is close enough to the unique global minimizer $\boldsymbol{x}_*$ and $p > q-1$, then both the function value and the gradient norm converge superlinearly with order $\frac{p}{q-1}$.

In this paper, we generalize the results of Doikov and Nesterov [11] by

1. considering the ARp algorithm, with its *adaptively chosen regularization*,
2. requiring only *approximate local model minimizers* at each iteration, and
3. covering the case of *locally uniformly convex functions*.

In the local convergence setting where $\boldsymbol{x}_0$ is sufficiently close to a local minimizer $\boldsymbol{x}_*$, we obtain that the gradient norms converge with order $\frac{p}{q-1}$ when $p > q-1$ during successful iterations. By additionally assuming which model minimizer $\boldsymbol{y}_k$ is chosen when faced with ambiguity one can ensure that all iterations are successful, and we achieve the same local $\frac{p}{q-1}$th-order convergence of iterates, function values and gradient norms.

To put this into perspective, let us compare this convergence rate with the one of Newton's method. It is well-known that Newton's method converges quadratically when the Hessian at the limit point is nonsingular or, equivalently, when f is locally uniformly convex of order $q = 2$. In this case, the access to additional derivatives gives a speed-up from quadratic (Newton's method) to pth-order convergence (ARp). Therefore, to achieve a tolerance of $\varepsilon > 0$ on the norm of the gradient, both methods require $O(\log(\log(\varepsilon^{-1})))$ iterations, but pth-order methods improve the coefficient of the leading-order term from $1/\log(2)$ to $1/\log(p)$. The advantage is even greater when the Hessian at $\boldsymbol{x}_*$ is singular, as Newton's method only converges linearly in this case whereas the access to enough derivatives ($p > q-1$) allows superlinear convergence of order $\frac{p}{q-1}$. The maximum number of iterations are then bounded by $O(\log(\varepsilon^{-1}))$ for Newton and $O(\log(\log(\varepsilon^{-1})))$ for ARp.

There are other works that cover the local convergence of regularization methods: cubic regularization methods were introduced by Griewank [13] and later independently revisited by Nesterov and Polyak [16]. Cartis, Gould and Toint [4, 5] added adaptivity of the regularization parameter and arrived at the ARC method (AR2 in our notation). Nesterov and Polyak [16, Theorem 3] and Cartis, Gould and Toint [4, Corollary 4.10] show quadratic local convergence of these second-order methods when the Hessian is positive definite at the minimizer, both when σ_k converges to zero and when it stays bounded away from zero.

Yue et al. [19] go even further and are able to show that positive definiteness

of the Hessian at the minimizer is not necessary for quadratic convergence of AR2, but rather that a local error bound condition (EB) suffices. This agrees with earlier work by Fan and Yuan [12] on quadratic convergence of regularized second-order methods and work by Bellavia and Morini [1] on quadratic convergence of adaptively regularized methods for nonlinear least squares. Rebjock and Boumal [18] prove the equivalence of different local properties including the error bound assumption. With it, they conclude that in the case where the set of local minimizers with a certain function value is a submanifold $\mathcal{S}$ and only the eigenvalues corresponding to normal directions are bounded away from zero, quadratic convergence of the AR2 method still holds. In fact, for any neighbourhood $\mathcal{U}$ of a local minimizer $\bar{\boldsymbol{x}} \in \mathcal{S}$ there is a neighbourhood $\mathcal{V}$ of $\bar{\boldsymbol{x}}$ such that if an iterate enters $\mathcal{V}$ then the sequence of AR2 iterates converges to some $\boldsymbol{x}_* \in \mathcal{S} \cap \mathcal{U}$ Q-quadratically[1]. To achieve this, the only assumption on how the regularization parameter σ_k is updated is that it does not increase during successful iterations and that it is always lower bounded by some $\sigma_{\min} > 0$.

Cartis, Gould and Toint [8] prove complexity bounds for the ARp method on nonconvex objective functions under various assumptions. In [8, Section 5.3] they discuss measure-dominated problems. As we will show in Lemma 3.2, functions that are uniformly convex of order $q \geq 2$ are also gradient-dominated of level $\frac{q}{q-1}$. Using the fact that the function decrease of ARp methods can be lower bounded in a certain way, the results of [8] imply that ARp methods are guaranteed to converge at least superlinearly if $p > q-1$, at least linearly if $p = q-1$ and at least sublinearly if $p < q-1$. More precisely, the corresponding upper bounds are $O(\log(\log(\varepsilon^{-1})))$, $O(\log(\varepsilon^{-1}))$ and $O(\varepsilon^{-\frac{p+1}{p}+\frac{q}{q-1}})$ iterations, respectively, to achieve an error tolerance $\|\nabla f(\boldsymbol{x})\| < \varepsilon$ in the worst-case. The $p > q-1$ case matches the discussion above. Note though, that by assuming different properties of f (local uniform convexity instead of global gradient domination) and focusing on the gradient norm convergence, we can derive sharp bounds for the order of convergence missing from the results in [8].

The remainder of the paper is organized as follows: in section 2, we describe the ARp algorithm and state its properties under basic assumptions. Afterwards, an example, which motivates our paper, shows that even asymptotically, not all iterations need to be successful for higher-order methods unlike in the second-order case. Section 3 contains local convergence results for ARp under local uniform convexity assumptions on f, both when not all iterations are successful and when this property is restored by additional assumptions on the choice of the model minimizer. The sharpness of the derived convergence rates is examined in section 4 and the rates are illustrated with experiments in section 5. We finally discuss the differences between our own analysis and other works in the literature and make some concluding remarks in section 6.

2. The ARp algorithm. In this paper, we aim to understand the asymptotic behaviour of the ARp algorithm, which is given in Algorithm 2.1. To this end, we assume that Algorithm 2.1 generates an infinite sequence $\boldsymbol{x}_k$, i.e., that no iterate ever satisfies the termination condition $\nabla f(\boldsymbol{x}_k) = \boldsymbol{0}$ exactly.

At iteration k the local Taylor expansion t_k and the local regularized model m_k

[1]Throughout the paper we use the terms Q-convergence and R-convergence as in [17, section A.2]. In particular, we say that a sequence converges linearly if it converges *at least* linearly, without claiming that it does not converge at a higher-order rate. In this parlance, even a quadratically converging sequence also converges linearly.

Algorithm 2.1: ARp algorithm

Parameters: $0 < \eta_1 \leq \eta_2 < 1$, $0 < \gamma_1 \leq 1 < \gamma_2$, $\theta > 0$
Input: $\boldsymbol{x}_0 \in \mathbb{R}^n$, $\sigma_0 > 0$
for $k = 0, 1, \ldots$ **do**
  if $\nabla f(\boldsymbol{x}_k) = \boldsymbol{0}$ **then**
    return
  end
  Find a local minimizer $\boldsymbol{y}_k$ of m_k defined according to (2.2b) that satisfies

$$m_k(\boldsymbol{y}_k) < m_k(\boldsymbol{x}_k) \quad \text{and} \quad \|\nabla m_k(\boldsymbol{x}_k)\| \leq \theta \|\boldsymbol{x}_k - \boldsymbol{y}_k\|^p \tag{2.1}$$

  Set $\rho_k = \dfrac{f(\boldsymbol{x}_k) - f(\boldsymbol{y}_k)}{t_k(\boldsymbol{x}_k) - t_k(\boldsymbol{y}_k)}$
  if $\rho_k \geq \eta_2$ **then**
    Set $\boldsymbol{x}_{k+1} = \boldsymbol{y}_k$ and $\sigma_{k+1} = \gamma_1 \sigma_k$ `// very successful iteration`
  else if $\rho_k \geq \eta_1$ **then**
    Set $\boldsymbol{x}_{k+1} = \boldsymbol{y}_k$ and $\sigma_{k+1} = \sigma_k$ `// successful iteration`
  else
    Set $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k$ and $\sigma_{k+1} = \gamma_2 \sigma_k$ `// unsuccessful iteration`
  end
end

at $\boldsymbol{x}_k$ are defined as

$$t_k(\boldsymbol{y}) = t^p_{\boldsymbol{x}_k}(\boldsymbol{y}) = f(\boldsymbol{x}_k) + \sum_{j=1}^{p} \frac{1}{j!} \nabla^j f(\boldsymbol{x}_k)[\boldsymbol{y} - \boldsymbol{x}_k]^j \tag{2.2a}$$

$$m_k(\boldsymbol{y}) = m^p_{\boldsymbol{x}_k, \sigma_k}(\boldsymbol{y}) = t^p_{\boldsymbol{x}_k}(\boldsymbol{y}) + \sigma_k \|\boldsymbol{y} - \boldsymbol{x}_k\|^{p+1} \tag{2.2b}$$

for $\boldsymbol{y} \in \mathbb{R}^n$. The tentative new iterate $\boldsymbol{y}_k$ is computed such that it satisfies (2.1). We call any such point satisfying (2.1) an *approximate minimizer* of m_k. After finding $\boldsymbol{y}_k$, the algorithm checks whether $\rho_k \geq \eta_1$ for this point, i.e., whether the decrease in the objective function is at least a specified fraction η_1 of the decrease predicted by the Taylor expansion t_k. The predicted decrease is always positive since $t_k(\boldsymbol{y}_k) \leq m_k(\boldsymbol{y}_k) < m_k(\boldsymbol{x}_k) = t_k(\boldsymbol{x}_k)$. If the sufficient decrease condition $\rho_k \geq \eta$ is satisfied, the iteration is called "successful", and the iterate is accepted ($\boldsymbol{x}_{k+1} = \boldsymbol{y}_k$). If moreover $\rho_k \geq \eta_2$, then regularization parameter is decreased to $\sigma_{k+1} = \gamma_1 \sigma_k$. Otherwise, the iteration is called "unsuccessful", the iterate is rejected ($\boldsymbol{x}_{k+1} = \boldsymbol{x}_k$), and the regularization is increased to $\sigma_{k+1} = \gamma_2 \sigma_k$.

In the remainder of this paper, we will always assume that $\eta_1 = \eta_2 = \eta$ for convenience. The presented convergence results also apply for the case when $\eta_1 < \eta_2$, but the assumption simplifies the statements and proofs. Compared to the ARp algorithms in [2, 7, 9] we have not included a $\sigma_{\min}$ parameter that provides a lower bound for the regularization parameter. Since we are analysing local convergence, there is no need to keep σ_k artificially away from zero. As we will discuss in subsection 3.2, under the right circumstances convergence at the optimal rate is guaranteed even when setting $\sigma_k = 0$.

The regularization in (2.2b) is chosen such that it compensates for the error in the Taylor approximation for the following function class:

DEFINITION 2.1. *A function is pth-order Lipschitz continuous for some $p \geq 2$ if it is p times continuous differentiable and there exists a constant $L_p > 0$ such that*

$$\|\nabla^p f(\boldsymbol{x}) - \nabla^p f(\boldsymbol{y})\| \leq L_p \|\boldsymbol{x} - \boldsymbol{y}\|$$

holds for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. For notational convenience, we define C^p_{Lip} as the set of pth-order Lipschitz continuous functions.

Under this assumption on f, the error in the pth-order Taylor expansion and its derivatives can be bounded in the following way by Lemma 2.1 in [7]:

$$|f(\boldsymbol{y}) - t^p_{\boldsymbol{x}}(\boldsymbol{y})| \leq \frac{L_p}{(p+1)!} \|\boldsymbol{x} - \boldsymbol{y}\|^{p+1} \tag{2.3a}$$

$$\|\nabla f(\boldsymbol{y}) - \nabla t^p_{\boldsymbol{x}}(\boldsymbol{y})\| \leq \frac{L_p}{p!} \|\boldsymbol{x} - \boldsymbol{y}\|^{p} \tag{2.3b}$$

$$\|\nabla^2 f(\boldsymbol{y}) - \nabla^2 t^p_{\boldsymbol{x}}(\boldsymbol{y})\| \leq \frac{L_p}{(p-1)!} \|\boldsymbol{x} - \boldsymbol{y}\|^{p-1} \tag{2.3c}$$

In particular, when $\sigma_k \geq L_p/(p+1)!$, then $m_k(\boldsymbol{y}) \geq f(\boldsymbol{y})$ for all $\boldsymbol{y} \in \mathbb{R}^n$.[2]

The following fundamental result shows that σ_k cannot become arbitrarily large throughout the course of the algorithm.

LEMMA 2.2 (Lemma 2.2 in [2], Lemma 3.2 in [9]). *For $f \in C^p_{\text{Lip}}$ the kth iteration is guaranteed to be successful ($\rho_k \geq \eta$) whenever $\sigma_k \geq \frac{1}{1-\eta}\frac{L_p}{(p+1)!}$. As a consequence, the regularization parameter stays upper bounded:*

$$\sigma_k \leq \sigma_{\max} \coloneqq \max\left\{\sigma_0,\ \frac{\gamma_2}{1-\eta}\frac{L_p}{(p+1)!}\right\}$$

for all $k \in \mathbb{N}$.

Proof. It suffices to show that $|\rho_k - 1| \leq 1 - \eta$ whenever $\sigma_k \geq \frac{1}{1-\eta}\frac{L_p}{(p+1)!}$. If σ_k is this large, we require

$$\begin{aligned}
|\rho_k - 1| &= \left|\frac{f(\boldsymbol{x}_k) - f(\boldsymbol{y}_k)}{t_k(\boldsymbol{x}_k) - t_k(\boldsymbol{y}_k)} - 1\right| = \frac{|t_k(\boldsymbol{y}_k) - f(\boldsymbol{y}_k)|}{|t_k(\boldsymbol{x}_k) - t_k(\boldsymbol{y}_k)|} \\
&= \frac{|t_k(\boldsymbol{y}_k) - f(\boldsymbol{y}_k)|}{|m_k(\boldsymbol{x}_k) - m_k(\boldsymbol{y}_k) + \sigma_k\|\boldsymbol{y}_k - \boldsymbol{x}_k\|^{p+1}|} \\
&\leq \frac{(L_p/(p+1)!)\|\boldsymbol{y}_k - \boldsymbol{x}_k\|^{p+1}}{\sigma_k\|\boldsymbol{y}_k - \boldsymbol{x}_k\|^{p+1}} = \frac{L_p}{(p+1)!}\frac{1}{\sigma_k} \leq 1 - \eta
\end{aligned}$$

using $t_k(\boldsymbol{x}_k) = m_k(\boldsymbol{x}_k) = f(\boldsymbol{x}_k)$, (2.1), (2.2b), and (2.3a). The update mechanism of Algorithm 2.1 then ensures that the regularization parameter is never increased beyond $\sigma_{\max}$. ∎

Note that if $\gamma_1 < 1$, then starting at $\sigma_0 \geq \frac{1}{1-\eta}\frac{L_p}{(p+1)!}$ the regularization parameter is decreased until it becomes smaller than this value. Afterwards, it never increases beyond $\frac{\gamma_2}{1-\eta}\frac{L_p}{(p+1)!}$ as shown above. Therefore, asymptotically the regularization parameter can be bounded independently of σ_0:

$$\limsup_{k\to\infty} \sigma_k \leq \frac{\gamma_2}{1-\eta}\frac{L_p}{(p+1)!}$$

[2]Note that the denominators in (2.3) differ from those found in the corresponding bounds in [2, 6, 9], by a factor of $p+1$, p and $p-1$ respectively. This is because the bounds derived in these papers are not sharp in terms of the constants. Therefore, some fundamental lemmas in our paper differ in the constants involved compared to their counterparts in [2, 6, 9].

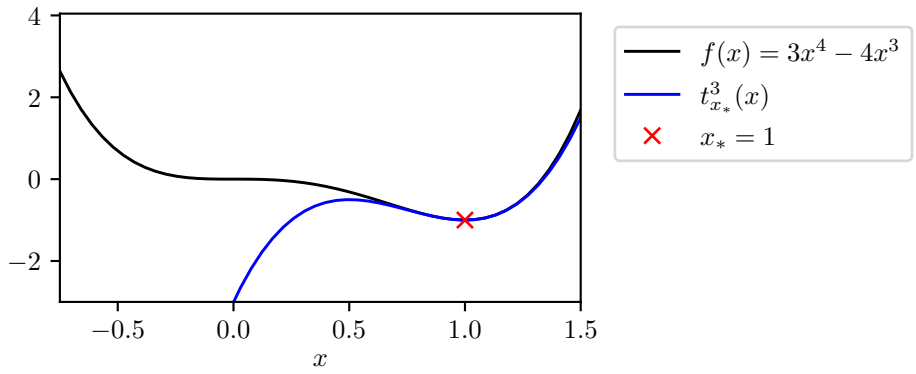

FIG. 2.1. *Example function and its third-order Taylor expansion at the global minimizer* $x_* = 1$.

2.1. Regularization parameter oscillations. In order to establish the difficulties inherent in proving local convergence of pth-order tensor methods for $p \geq 3$ it is helpful to consider a simple one-dimensional example. It shows that when $\boldsymbol{x}_k$ is close to a non-degenerate global minimizer of the objective function, the pth-order Taylor expansion is locally strongly convex but does not have to be globally convex or even lower bounded, unlike for the second-order case. This impacts the local convergence when $\boldsymbol{y}_k$ is chosen as the global minimizer of m_k.

Example 2.3. Consider the objective function $f(x) = 3x^4 - 4x^3$ shown in Figure 2.1. Its only stationary points are $x = 0$ and $x = 1$, of which the first is a saddle point and the second is the global minimizer x_*. The objective function and its derivatives at x_* are $f(x_*) = -1$, $f'(x_*) = 0$, $f''(x_*) = 12$, and $f'''(x_*) = 48$. Therefore, x_* is a nondegenerate local minimizer in the sense that the Hessian is positive definite at x_* and the function is locally strongly convex around that point. Nonetheless, because $f'''(x_*) \neq 0$ the third-order Taylor expansion at the minimizer is neither convex nor lower bounded. We will show that when y_k is chosen as the global minimizer of m_k, then there exist parameters σ_0, η, γ_1, and γ_2 such that only every second iteration is successful for *any* x_0 close enough to x_*. This is because on every other iteration, the regularization parameter is so small that the global minimizer of the model lies outside the convex region around the expansion point.

Note that when $\sigma = 3$ the local model is exact: $m^3_{x,3}(y) = t^3_x(y) + 3(y-x)^4 = f(y)$. The last equality is easy to see by considering that both sides are quartic polynomials and comparing the first four derivatives at $y = x$. It implies

$$m_k(y) = f(y) + (\sigma_k - 3)(y - x)^4.$$

For the given objective function we have $L_3 = 4! \cdot 3$. By virtue of Lemma 2.2 we know that for $\sigma_k \geq \frac{L_3}{4!(1-\eta)}$ every iteration will be successful ($\rho \geq \eta$) regardless of which minimizer of m_k is chosen. Let $\eta = 1/2$, then we obtain that the iteration is successful whenever $\sigma_k \geq 6$. Next, we show that the global minimizer y_k of the model m_k leads to an unsuccessful iteration whenever $\sigma_k \leq 2$ and $x_k \in [\frac{4}{5}, \frac{6}{5}]$. In this case, we focus on the interval $[0, \frac{4}{3}]$ where $f(x_k) < 0$. Notice that

$$\min_{y \in [0, \frac{4}{3}]} m_k(y) = \min_{y \in [0, \frac{4}{3}]} f(y) + (\sigma_k - 3)(y - x_k)^4 \geq -1 + (\sigma_k - 3)\left(\frac{6}{5}\right)^4$$

and

$$m_k(-1) = f(-1) + (\sigma_k - 3)(x_k + 1)^4 \leq 7 + (\sigma_k - 3)\left(\frac{9}{5}\right)^4.$$

The difference between these bounds is

$$\left(\min_{y\in[0,\frac{4}{3}]} m_k(y)\right) - m_k(-1) \geq -8 + (\sigma_k - 3)\left(\left(\frac{6}{5}\right)^4 - \left(\frac{9}{5}\right)^4\right) > 0.$$

This implies that any global minimizer y_k of m_k is outside the interval $[0, \frac{4}{3}]$, which means $f(y_k) > 0 > f(x_k)$. Therefore, $\rho_k < 0$ and the iteration is unsuccessful.

Now, take $\sigma_0 = 6$, $\eta = \frac{1}{2}$, $\gamma_1 = \frac{1}{3}$, $\gamma_2 = 3$, and x_0 with $f(x_0) \leq -0.9$. Because $\{x \in \mathbb{R} \mid f(x) \leq -0.9\} \subset [\frac{4}{5}, \frac{6}{5}]$ and iterates of Algorithm 2.1 have monotonically decreasing function values, all iterates lie in this interval. The first iteration with $\sigma_0 = 6$ is successful, so the regularization parameter is decreased to $\sigma_1 = 2$. The parameter is now too small ($\sigma_1 \leq 2$), the iteration is unsuccessful, and the parameter is increased to $\sigma_2 = 6$. At this point the same cycle repeats ad infinitum. The points x_k converge to x_*, but only every second iteration is successful and makes progress.

This example gives an important insight into the ARp method: when choosing a global minimizer of m_k at each iteration, the regularization parameter σ_k will not in general go to zero asymptotically even when the objective function is locally strongly convex around the minimizer. Since the global minimizer satisfies (2.1), we conclude that it is impossible to show for the fully adaptive[3] ARp method in Algorithm 2.1 that asymptotically all iterations will be successful without specifying which minimizer of the local model needs to be chosen.

3. Local convergence for locally uniformly convex functions. We follow Doikov and Nesterov [11] by considering the property of uniform convexity of order q. For our purposes though, we only require the function to be *locally* uniformly convex in a closed convex neighbourhood around a local minimizer $\boldsymbol{x}_*$:

DEFINITION 3.1. *A function is locally uniformly convex of order $q \geq 2$ around a local minimizer $\boldsymbol{x}_*$ if there exist parameters $\mu_q, r_q > 0$ such that*

$$(\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}))^T(\boldsymbol{x} - \boldsymbol{y}) \geq \mu_q \|\boldsymbol{x} - \boldsymbol{y}\|^q \quad \forall \boldsymbol{x}, \boldsymbol{y} \in B(\boldsymbol{x}_*, r_q), \tag{3.1}$$

where $B(\boldsymbol{x}_*, r_q)$ is a closed ball of radius r_q around $\boldsymbol{x}_*$. For notational convenience, we write $M^q(f)$ for the set of local minimizers $\boldsymbol{x}_*$ such that (3.1) holds.

Note that by virtue of [15, Lemma 4.2.1], (3.1) implies

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) + \frac{\mu_q}{q}\|\boldsymbol{y} - \boldsymbol{x}\|^q \quad \forall \boldsymbol{x}, \boldsymbol{y} \in B(\boldsymbol{x}_*, r_q), \tag{3.2}$$

and simultaneously (3.2) implies (3.1) (albeit with a different constant μ_q). Therefore, the two inequalities define the same notion of uniform convexity.

Local uniform convexity of order q implies local strict convexity and therefore that $\boldsymbol{x}_*$ is an isolated minimizer, but for $q > 2$ the assumption is weaker than local strong convexity. Indeed, local strong convexity, namely $\nabla^2 f(\boldsymbol{x}) \succeq \mu_2 \boldsymbol{I}$ for all $\boldsymbol{x} \in B(\boldsymbol{x}_*, r_q)$, is equivalent to (3.1) with $q = 2$. By analysing uniformly convex functions instead of strongly convex functions we are able to show superlinear convergence even for degenerate minimizers in certain cases.

[3] By fully adaptive ARp we mean the case $\gamma_1 < 1$, i.e., the regularization parameter is not just increased during unsuccessful iterations but also strictly decreased during successful iterations. In this terminology, the $\gamma_1 = 1$ case might be called semi-adaptive.

LEMMA 3.2. *Let $\boldsymbol{x}_* \in M^q(f)$. Then f satisfies*

$$f(\boldsymbol{x}) - f(\boldsymbol{x}_*) \geq \frac{\mu_q}{q}\|\boldsymbol{x} - \boldsymbol{x}_*\|^q, \tag{3.3}$$

$$\|\nabla f(\boldsymbol{x})\| \geq \mu_q \|\boldsymbol{x} - \boldsymbol{x}_*\|^{q-1}, \text{ and} \tag{3.4}$$

$$f(\boldsymbol{x}) - f(\boldsymbol{x}_*) \leq \frac{q-1}{q}\left(\frac{1}{\mu_q}\right)^{\frac{1}{q-1}} \|\nabla f(\boldsymbol{x})\|^{\frac{q}{q-1}} \tag{3.5}$$

for all $\boldsymbol{x} \in B(\boldsymbol{x}_*, r_q)$. Moreover, for $f \in \mathcal{C}^2$ there exists some $\nu > 0$ such that

$$f(\boldsymbol{x}) - f(\boldsymbol{x}_*) \leq \frac{\nu}{2}\|\boldsymbol{x} - \boldsymbol{x}_*\|^2 \tag{3.6}$$

$$\|\nabla f(\boldsymbol{x})\| \leq \nu \|\boldsymbol{x} - \boldsymbol{x}_*\| \tag{3.7}$$

hold for all $\boldsymbol{x} \in B(\boldsymbol{x}_*, r_q)$.

Proof. The first inequality follows from (3.2) by substituting $\boldsymbol{x} = \boldsymbol{x}_*$ and $\boldsymbol{y} = \boldsymbol{x}$. The second follows from (3.1) by considering

$$\|\nabla f(\boldsymbol{x})\|\|\boldsymbol{x} - \boldsymbol{x}_*\| \geq (\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}_*))^T(\boldsymbol{x} - \boldsymbol{x}_*) \geq \mu_q \|\boldsymbol{x} - \boldsymbol{x}_*\|^q.$$

By [15, Lemma 4.2.2], we obtain (3.5).

For the second part, the 2-norm of $\nabla^2 f$ is bounded by some $\nu \geq 0$ on the compact set $B(\boldsymbol{x}_*, r_q)$. Equivalently, ∇f is locally Lipschitz continuous with $L_1 = \nu$ on that ball. Apply (2.3a) and (2.3b) to the Taylor expansion at $\boldsymbol{x}_*$ to obtain (3.6) and (3.7). ∎

3.1. Convergence without assumptions on the model minimizer. We first consider the best possible convergence results that can be derived for Algorithm 2.1 as stated. In particular, there are no further restrictions on $\boldsymbol{y}_k$ other than (2.1). Since the Taylor expansions are not necessarily convex, even when $\boldsymbol{x}_k$ is close to $\boldsymbol{x}_*$, there can be many local minimizers of m_k as long as σ_k is small enough.

The common assumptions for all the results in this section and the next are that f is pth-order Lipschitz continuous (abbreviated as $f \in C^p_{\text{Lip}}$, see Definition 2.1), that f is locally uniformly convex of order q around the local minimizer $\boldsymbol{x}_*$ (abbreviated as $\boldsymbol{x}_* \in M^q(f)$, see Definition 3.1) and that $p > q - 1$. These will be stated explicitly at the start of each statement. Additionally, we silently assume that $\boldsymbol{x}_k$ and $\boldsymbol{y}_k$ are generated by Algorithm 2.1 using all p derivatives of the objective function and that the algorithm never terminates, i.e. that $\nabla f(\boldsymbol{x}_k) \neq \boldsymbol{0}$ for all k.

The following two lemmas form the basis of our convergence analysis. The first result bounds the gradient norm at the tentative new iterate $\boldsymbol{y}_k$ in terms of the gradient norm at the current iterate $\boldsymbol{x}_k$ during successful iterations, while the second provides a corresponding bound for the function value gap $f(\boldsymbol{x}_k) - f(\boldsymbol{x}_*)$.

LEMMA 3.3. *Let $f \in C^p_{\text{Lip}}$ and $\boldsymbol{x}_* \in M^q(f)$. Assume that $\boldsymbol{x}_k, \boldsymbol{y}_k \in B(\boldsymbol{x}_*, r_q)$ and $f(\boldsymbol{y}_k) \leq f(\boldsymbol{x}_k)$, then the gradients satisfy*

$$\|\nabla f(\boldsymbol{y}_k)\| \leq R_k \|\boldsymbol{y}_k - \boldsymbol{x}_k\|^p \leq R_k \left(\frac{q}{\mu_q}\right)^{\frac{p}{q-1}} \|\nabla f(\boldsymbol{x}_k)\|^{\frac{p}{q-1}} \tag{3.8}$$

where $R_k := \big(L_p/p! + \theta + (p+1)\sigma_k\big)$. If all iterates stay within $B(\boldsymbol{x}_*, r_q)$, then

$$\|\nabla f(\boldsymbol{x}_{k+1})\| \leq R_{\max} \left(\frac{q}{\mu_q}\right)^{\frac{p}{q-1}} \|\nabla f(\boldsymbol{x}_k)\|^{\frac{p}{q-1}} \tag{3.9}$$

holds at every successful iteration for $R_{\max} := \big(L_p/p! + \theta + (p+1)\sigma_{\max}\big)$.

Proof. Following the argument in [2, Lemma 2.3], we have

$$\begin{aligned}\|\nabla f(\boldsymbol{y}_k)\| &\leq \|\nabla f(\boldsymbol{y}_k) - \nabla t_k(\boldsymbol{y}_k)\| + \|\nabla t_k(\boldsymbol{y}_k)\| \\ &\leq \|\nabla f(\boldsymbol{y}_k) - \nabla t_k(\boldsymbol{y}_k)\| + \|\nabla m_k(\boldsymbol{y}_k)\| + (p+1)\sigma_k\|\boldsymbol{y}_k - \boldsymbol{x}_k\|^p \\ &\leq \frac{L_p}{p!}\|\boldsymbol{y}_k - \boldsymbol{x}_k\|^p + \theta\|\boldsymbol{y}_k - \boldsymbol{x}_k\|^p + (p+1)\sigma_k\|\boldsymbol{y}_k - \boldsymbol{x}_k\|^p \\ &= \big(L_p/p! + \theta + (p+1)\sigma_k\big)\|\boldsymbol{y}_k - \boldsymbol{x}_k\|^p\end{aligned}$$

where we have used the triangle inequality, (2.3b), and (2.1).

Rearranging (3.2) with $\boldsymbol{x} = \boldsymbol{x}_k$ and $\boldsymbol{y} = \boldsymbol{y}_k$ gives

$$\nabla f(\boldsymbol{x}_k)^T(\boldsymbol{x}_k - \boldsymbol{y}_k) \geq f(\boldsymbol{x}_k) - f(\boldsymbol{y}_k) + \frac{\mu_q}{q}\|\boldsymbol{x}_k - \boldsymbol{y}_k\|^q \geq \frac{\mu_q}{q}\|\boldsymbol{x}_k - \boldsymbol{y}_k\|^q$$

and therefore $\|\nabla f(\boldsymbol{x}_k)\| \geq \frac{\mu_q}{q}\|\boldsymbol{x}_k - \boldsymbol{y}_k\|^{q-1}$ by $f(\boldsymbol{y}_k) \leq f(\boldsymbol{x}_k)$. Combine this result with the bound on $\|\nabla f(\boldsymbol{y}_k)\|$ to obtain (3.8).

By Lemma 2.2 we know $\sigma_k \leq \sigma_{\max}$, and during successful iterations $\boldsymbol{x}_{k+1} = \boldsymbol{y}_k$ and $f(\boldsymbol{x}_k) - f(\boldsymbol{y}_k) \geq \eta(t_k(\boldsymbol{x}_k) - t_k(\boldsymbol{y}_k)) \geq \eta(m_k(\boldsymbol{x}_k) - m_k(\boldsymbol{y}_k)) > 0$, so (3.9) follows immediately. ∎

LEMMA 3.4. *Let $f \in C^p_{\text{Lip}}$, $\boldsymbol{x}_* \in M^q(f)$ and $p > q - 1$. Assume that $\boldsymbol{x}_k, \boldsymbol{y}_k \in B(\boldsymbol{x}_*, r_q)$, $f(\boldsymbol{y}_k) \leq f(\boldsymbol{x}_k)$ and that the norm of the gradient at $\boldsymbol{x}_k$ is small enough. Then the function values satisfy*

$$f(\boldsymbol{y}_k) - f(\boldsymbol{x}_*) \leq \frac{q-1}{q}\left(\frac{1}{\mu_q}\right)^{\frac{1}{q-1}}\left(\frac{2q}{\mu_q}\right)^{\frac{p}{q-1}} R_k^{\frac{q}{q-1}}(f(\boldsymbol{x}_k) - f(\boldsymbol{x}_*))^{\frac{p}{q-1}}. \tag{3.10}$$

In particular, if all iterates stay within $B(\boldsymbol{x}_*, r_q)$ and $\|\nabla f(\boldsymbol{x}_k)\| \to 0$, then

$$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_*) \leq \frac{q-1}{q}\left(\frac{1}{\mu_q}\right)^{\frac{1}{q-1}}\left(\frac{2q}{\mu_q}\right)^{\frac{p}{q-1}} R_{\max}^{\frac{q}{q-1}}(f(\boldsymbol{x}_k) - f(\boldsymbol{x}_*))^{\frac{p}{q-1}} \tag{3.11}$$

holds at every successful iteration k for k large enough.

Proof. Whenever the norm of the gradient at $\boldsymbol{y}_k$ is small enough such that

$$\|\nabla f(\boldsymbol{y}_k)\|^{\frac{p-q+1}{p}} \leq \frac{\mu_q}{2q}\left(\frac{1}{R_{\max}}\right)^{\frac{q-1}{p}}, \tag{3.12}$$

then we can use (3.2) and the first inequality in (3.8) to obtain

$$\begin{aligned}f(\boldsymbol{x}_k) - f(\boldsymbol{x}_*) &\geq f(\boldsymbol{x}_k) - f(\boldsymbol{y}_k) \\ &\geq -\|\nabla f(\boldsymbol{y}_k)\|\|\boldsymbol{x}_k - \boldsymbol{y}_k\| + \frac{\mu_q}{q}\|\boldsymbol{x}_k - \boldsymbol{y}_k\|^q \\ &\geq \left(\frac{\mu_q}{q}\left(\frac{\|\nabla f(\boldsymbol{y}_k)\|}{R_k}\right)^{\frac{q-1}{p}} - \|\nabla f(\boldsymbol{y}_k)\|\right)\|\boldsymbol{x}_k - \boldsymbol{y}_k\| \\ &\geq \frac{\mu_q}{2q}\left(\frac{1}{R_k}\right)^{\frac{q}{p}}\|\nabla f(\boldsymbol{y}_k)\|^{\frac{q}{p}}.\end{aligned}$$

Equation (3.10) follows from this and (3.5) applied to $\boldsymbol{y}_k$. Note that from (3.9) we know that (3.12) holds as long as $\|\nabla f(\boldsymbol{x}_k)\|$ is small enough.

As before $\sigma_k \leq \sigma_{\max}$, the function value decreases in successful iterations and by assumption the gradient norms converge to zero, which means that the second claim follows from (3.10). ∎

Lemmas 3.3 and 3.4 show that asymptotically the norm of the gradients and the function value gaps converge at a Q-$\frac{p}{q-1}$th-order rate during successful iterations when $p > q-1$. Using local uniform convexity of f, the convergence of the iterates is also of order $\frac{p}{q-1}$ as shown by the following theorem. The necessity of the assumption that all minimizers stay in the uniform convexity region around $\boldsymbol{x}_*$ is clarified in Theorem 3.7 and the preceding discussion.

THEOREM 3.5. *Let $f \in C^p_{\text{Lip}}$, $\boldsymbol{x}_* \in M^q(f)$ and $p > q-1$. Assume that all iterates stay in $B(\boldsymbol{x}_*, r_q)$ and that $\boldsymbol{x}_0$ is close enough to $\boldsymbol{x}_*$. Then*

$$\begin{aligned} \|\nabla f(\boldsymbol{x}_{k_i})\| \to 0 \quad & \text{at a Q-}\tfrac{p}{q-1}\text{th-order rate} \\ f(\boldsymbol{x}_{k_i}) \to f(\boldsymbol{x}_*) \quad & \text{at a Q-}\tfrac{p}{q-1}\text{th-order rate} \\ \boldsymbol{x}_{k_i} \to \boldsymbol{x}_* \quad & \text{at an R-}\tfrac{p}{q-1}\text{th-order rate} \end{aligned}$$

as $i \to \infty$, where $\{k_1, k_2, \dots\}$ are the successful iterations.

Proof. Note that the iterates stay constant in unsuccessful iterations, so it suffices to analyse how gradient norms, function values and iterates change from $\boldsymbol{x}_k$ to $\boldsymbol{x}_{k+1}$ in successful iterations.

Since ∇f is continuous, we know that for $\boldsymbol{x}_0$ close enough to $\boldsymbol{x}_*$, $\|\nabla f(\boldsymbol{x}_0)\|$ becomes small enough such that (3.9) ensures at least linear decrease to zero at every successful iteration. This shows that $\|\nabla f(\boldsymbol{x}_{k_i})\|$ converges to zero, which in turn means that (3.9) also implies the Q-superlinear convergence rate and that (3.11) shows the same Q-superlinear rate for the function value gaps $f(\boldsymbol{x}_{k_i}) - f(\boldsymbol{x}_*)$.

The R-convergence of $\boldsymbol{x}_{k_i}$ follows by Lemma 3.2, because

$$\|\boldsymbol{x}_{k_i} - \boldsymbol{x}_*\| \leq \mu_q^{-\frac{1}{q-1}} \|\nabla f(\boldsymbol{x}_{k_i})\|^{\frac{1}{q-1}}.$$

Even though the bound includes an exponent on $\|\nabla f(\boldsymbol{x}_{k_i})\|$ this does not change the order of convergence. ∎

Now that we know how fast gradient norms, function values and iterates converge during successful iterates, we can give a lower bound on the convergence rate of all iterates by exploiting that we can bound the number of unsuccessful iterations in terms of the number of successful ones. The ratio of unsuccessful to successful iterations depends on the choice of γ_1 and γ_2. In particular, when $\gamma_1 = 1$ (σ_k is never decreased) then asymptotically all iterations are successful and we recover the full $\frac{p}{q-1}$th-order rates from Theorem 3.5.

THEOREM 3.6. *Let $f \in C^p_{\text{Lip}}$, $\boldsymbol{x}_* \in M^q(f)$ and $p > q-1$. Assume that all iterates stay in $B(\boldsymbol{x}_*, r_q)$ and that $\boldsymbol{x}_0$ is close enough to $\boldsymbol{x}_*$. Then*

$$\begin{aligned} \|\nabla f(\boldsymbol{x}_k)\| \to 0 \quad & \text{at an R-}\sqrt[\alpha+1]{\tfrac{p}{q-1}}\text{th-order rate} \\ f(\boldsymbol{x}_k) \to f(\boldsymbol{x}_*) \quad & \text{at an R-}\sqrt[\alpha+1]{\tfrac{p}{q-1}}\text{th-order rate} \\ \boldsymbol{x}_k \to \boldsymbol{x}_* \quad & \text{at an R-}\sqrt[\alpha+1]{\tfrac{p}{q-1}}\text{th-order rate} \end{aligned}$$

as $k \to \infty$, where $\gamma_1 = \gamma_2^{-\alpha}$.

Proof. The assumptions are the same as those of Theorem 3.5, so we know the convergence rates when considering only the successful iterations. Let s_k and u_k be the number of successful and unsuccessful iterations from 0 to $k-1$ respectively and

let $\{k_1, k_2, \dots\}$ be the indices of the successful iterations. Note that since the iterates do not change in unsuccessful iterations, we have that $\boldsymbol{x}_k = \boldsymbol{x}_{k_i}$ for $i = s_k$.

Applying the same reasoning as [5, Lemma 2.1] and [2, Lemma 2.4], we see that

$$\sigma_0 \gamma_1^{s_k} \gamma_2^{u_k} = \sigma_k \leq \sigma_{\max}.$$

Therefore, substituting $\gamma_1 = \gamma_2^{-\alpha}$, taking logs on both sides and rearranging we get that $u_k - \alpha s_k$ is upper bounded by the constant $\log(\sigma_{\max}/\sigma_0)/\log(\gamma_2)$. Expressed in another way, using $s_k + u_k = k$, we get

$$s_k \geq \frac{k}{\alpha+1} - \frac{\log(\sigma_{\max}/\sigma_0)}{(\alpha+1)\log(\gamma_2)}.$$

To establish rth-order rates of R-convergence for a sequence of values $(\varepsilon_k)_{k\in\mathbb{N}}$ that converge to zero, we need to show that asymptotically $\log(\varepsilon_k^{-1})$ grows at least as fast as $\Theta(r^k)$. By Theorem 3.5 we already know that there exist $c > 0$ such that

$$\log(\|\nabla f(\boldsymbol{x}_{k_i})\|^{-1}) \geq c\Big(\tfrac{p}{q-1}\Big)^i$$

holds for all i large enough. This implies

$$\begin{aligned}\log(\|\nabla f(\boldsymbol{x}_k)\|^{-1}) = \log(\|\nabla f(\boldsymbol{x}_{k_{s_k}})\|^{-1}) &\geq c\Big(\tfrac{p}{q-1}\Big)^{s_k} \\ &\geq c\Big(\tfrac{p}{q-1}\Big)^{\frac{k}{\alpha+1} - \frac{\log(\sigma_{\max}/\sigma_0)}{(\alpha+1)\log(\gamma_2)}} = c\Big(\tfrac{p}{q-1}\Big)^{-\frac{\log(\sigma_{\max}/\sigma_0)}{(\alpha+1)\log(\gamma_2)}} \Big(\sqrt[\alpha+1]{\tfrac{p}{q-1}}\Big)^k\end{aligned}$$

for all k large enough, which means $\|\nabla f(\boldsymbol{x}_k)\| \to 0$ at the claimed rate. Following the same arguments, we also get the correct convergence rates for $f(\boldsymbol{x}_k) \to f(\boldsymbol{x}_*)$ and $\boldsymbol{x}_k \to \boldsymbol{x}_*$ from Theorem 3.5. ∎

Note that even though the order of convergence decreased between Theorems 3.5 and 3.6, it is still greater than one, since $\alpha \geq 0$ and $p > q - 1$. Therefore, the convergence stays superlinear.

A common assumption in the previous results is that all iterates stay in the neighbourhood of uniform convexity around $\boldsymbol{x}_*$. Clearly, this is unavoidable as we made very few assumptions about f outside this region. In particular, as $\boldsymbol{x}_*$ is only a local minimizer, a tentative iterate $\boldsymbol{y}_k$ outside the neighbourhood might achieve an even lower function value than $f(\boldsymbol{x}_*)$ and be accepted as $\boldsymbol{x}_{k+1}$. In such case, the algorithm does not converge to $\boldsymbol{x}_*$. Indeed, it will never return to the neighbourhood of $\boldsymbol{x}_*$, since $f(\boldsymbol{x}_k)$ is monotonically decreasing. The next result shows that, if $\boldsymbol{x}_0$ is close enough to $\boldsymbol{x}_*$, then the case described above is the only case in which the algorithm leaves the neighbourhood of $\boldsymbol{x}_*$.

THEOREM 3.7. *Let $f \in C^p_{\text{Lip}}$, $\boldsymbol{x}_* \in M^q(f)$ and $p > q-1$. If $\boldsymbol{x}_0$ is close enough to $\boldsymbol{x}_*$ (depending on σ_0), then either all iterates stay inside $B(\boldsymbol{x}_*, r_q)$ or the first iterate outside this ball satisfies $f(\boldsymbol{x}_k) < f(\boldsymbol{x}_*)$.*

Proof. The iterates only change when the iteration is successful, and it suffices to consider successful iterations. Let $\{k_1, k_2, \dots\}$ be the set of iterations k where $\rho_k \geq \eta$, then $\boldsymbol{x}_{k_{i+1}} = \boldsymbol{x}_{k_i+1}$. Moreover, let $I \in \mathbb{N}_{>0}$ be the first iteration such that $f(\boldsymbol{x}_{k_I}) < f(\boldsymbol{x}_*)$ or infinity if no such iteration exists. We have that

$$\begin{aligned}f(\boldsymbol{x}_{k_i}) - f(\boldsymbol{x}_*) &\geq f(\boldsymbol{x}_{k_i}) - f(\boldsymbol{x}_{k_i+1}) \geq \eta(t_{k_i}(\boldsymbol{x}_{k_i}) - t_{k_i}(\boldsymbol{x}_{k_i+1})) \\ &\geq \eta\sigma_{k_i}\|\boldsymbol{x}_{k_i} - \boldsymbol{x}_{k_i+1}\|^{p+1} \geq \eta\sigma_0\gamma_1^i\|\boldsymbol{x}_{k_i} - \boldsymbol{x}_{k_i+1}\|^{p+1}\end{aligned}$$

for all $i+1 < I$ using $\rho_{k_i} \geq \eta$, (2.1), and the fact that σ_k is only decreased on successful iterations. Therefore,

$$\|\boldsymbol{x}_{k_{i+1}} - \boldsymbol{x}_*\| = \|\boldsymbol{x}_{k_i+1} - \boldsymbol{x}_*\| \leq \|\boldsymbol{x}_{k_i+1} - \boldsymbol{x}_{k_i}\| + \|\boldsymbol{x}_{k_i} - \boldsymbol{x}_*\| \tag{3.13a}$$

$$\leq \left(\frac{f(\boldsymbol{x}_{k_i}) - f(\boldsymbol{x}_*)}{\eta\sigma_0\gamma_1^i}\right)^{\frac{1}{p+1}} + \|\boldsymbol{x}_{k_i} - \boldsymbol{x}_*\| \tag{3.13b}$$

$$\leq \left(\frac{\frac{q-1}{q}\left(\frac{1}{\mu_q}\right)^{\frac{1}{q-1}}}{\eta\sigma_0} \frac{\|\nabla f(\boldsymbol{x}_{k_i})\|^{\frac{q}{q-1}}}{\gamma_1^i}\right)^{\frac{1}{p+1}} + \left(\frac{1}{\mu_q}\right)^{\frac{1}{q-1}} \|\nabla f(\boldsymbol{x}_{k_i})\|^{\frac{1}{q-1}}, \tag{3.13c}$$

where the last inequality follows from Lemma 3.2 assuming $\boldsymbol{x}_{k_i}$ is inside $B(\boldsymbol{x}_*, r_q)$.

At this point, we cannot yet assert that either (3.9) or (3.13) holds, since the former requires $\boldsymbol{x}_{k_i+1}, \boldsymbol{x}_{k_i} \in B(\boldsymbol{x}_*, r_q)$ and the latter $\boldsymbol{x}_{k_i} \in B(\boldsymbol{x}_*, r_q)$. Still, choosing $\boldsymbol{x}_0$ close enough to $\boldsymbol{x}_*$ we have that $\|\nabla f(\boldsymbol{x}_0)\|$ gets arbitrarily small. We can select $\boldsymbol{x}_0$ in such a way that it simultaneously satisfies

1. $\boldsymbol{x}_0 \in B(\boldsymbol{x}_*, r_q)$,
2. the right-hand side of (3.13) is smaller than r_q for $i = 0$, and
3. the right-hand side of (3.9) is bounded by $\gamma_1^{(q-1)/q}\|\nabla f(\boldsymbol{x}_0)\|$ for $k = 0$.

By choosing $\boldsymbol{x}_{k_0} = \boldsymbol{x}_0$ in such a way, it follows from (3.13) that $\boldsymbol{x}_{k_1}$ lies inside $B(\boldsymbol{x}_*, r_q)$ and therefore its gradient norm is bounded by (3.9). This implies that the gradient norm decreases at least by a factor of $\gamma_1^{(q-1)/q}$. Then the right-hand side of (3.13) for $i = 1$ is smaller than the bound for $i = 0$, which was already smaller than r_q, and so $\boldsymbol{x}_{k_2} \in B(\boldsymbol{x}_*, r_q)$. By induction, all successful iterates $\boldsymbol{x}_{k_i}$ with $i < I$ lie inside a circle of radius r_q around $\boldsymbol{x}_*$. This proves the claim. ∎

3.2. Choosing the right minimizer. The results of the previous subsection capture the worst-case convergence rate when all iterates stay in the region of uniform convexity. They explicitly deal with the fact that even asymptotically not all iterations need to be successful. However, this can be amended by specifying which local minimizer of m_k should be chosen, if there are multiple options. In the following, we show that there always exists a local minimizer in the vicinity of $\boldsymbol{x}_*$ that leads to a successful iteration (Lemmas 3.10 and 3.11). Recall that in this paper an approximate model minimizer is simply one that satisfies (2.1). By selecting any such minimizer close to $\boldsymbol{x}_*$, we can restore Q-$\frac{p}{q-1}$th-order convergence of the gradient norms regardless of the values of γ_1 and γ_2 (Theorem 3.12). As a stepping stone, we use that even though t_k is not necessarily convex around $\boldsymbol{x}_*$, even when the expansion point is close to $\boldsymbol{x}_*$,[4] it does satisfy (3.2) if one of $\boldsymbol{x}$ and $\boldsymbol{y}$ is the expansion point.

LEMMA 3.8. *Let $f \in C^p_{\text{Lip}}$, $\boldsymbol{x}_* \in M^q(f)$ and $p > q-1$. There exist $\mu'_q < \mu_q$ and $r'_q \leq r_q$ such that*

$$t^p_{\boldsymbol{x}}(\boldsymbol{x}) \geq t^p_{\boldsymbol{x}}(\boldsymbol{y}) + \nabla t^p_{\boldsymbol{x}}(\boldsymbol{y})^T(\boldsymbol{x} - \boldsymbol{y}) + \frac{\mu'_q}{q}\|\boldsymbol{x} - \boldsymbol{y}\|^q \tag{3.14}$$

$$t^p_{\boldsymbol{x}}(\boldsymbol{y}) \geq t^p_{\boldsymbol{x}}(\boldsymbol{x}) + \nabla t^p_{\boldsymbol{x}}(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) + \frac{\mu'_q}{q}\|\boldsymbol{y} - \boldsymbol{x}\|^q \tag{3.15}$$

hold for all $\boldsymbol{x}, \boldsymbol{y} \in B(\boldsymbol{x}_*, r'_q)$.

[4] As a counterexample consider $f(x) = x^4 + x^5$ (which is locally uniformly convex of order $q = 4$ around $x_* = 0$) and its fourth-order Taylor expansion t_k at a point $x_k < 0$. Analytically, we can deduce that $t''_k(0) = 20x_k^3 < 0$ which means t_k is not locally convex around x_*.

Proof. Let $\boldsymbol{x}, \boldsymbol{y} \in B(\boldsymbol{x}_*, r_q')$ for $r_q' \leq r_q$ small enough such that

$$\|\boldsymbol{x} - \boldsymbol{y}\| \leq \|\boldsymbol{x} - \boldsymbol{x}_*\| + \|\boldsymbol{y} - \boldsymbol{x}_*\| \leq \sqrt[p-q+1]{\frac{\mu_q}{2q} \Big/ \left(\frac{L_p}{(p+1)!} + \frac{L_p}{p!} \right)}.$$

In the first case we use (2.3a), (2.3b), and (3.2) to obtain

$$t_{\boldsymbol{x}}^p(\boldsymbol{x}) - t_{\boldsymbol{x}}^p(\boldsymbol{y}) - \nabla t_{\boldsymbol{x}}^p(\boldsymbol{y})^T(\boldsymbol{x} - \boldsymbol{y}) \geq \frac{\mu_q}{q}\|\boldsymbol{x} - \boldsymbol{y}\|^q - \left(\frac{L_p}{(p+1)!} + \frac{L_p}{p!} \right) \|\boldsymbol{x} - \boldsymbol{y}\|^{p+1}$$

and similarly

$$t_{\boldsymbol{x}}^p(\boldsymbol{y}) - t_{\boldsymbol{x}}^p(\boldsymbol{x}) - \nabla t_{\boldsymbol{x}}^p(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) \geq \frac{\mu_q}{q}\|\boldsymbol{x} - \boldsymbol{y}\|^q - \frac{L_p}{(p+1)!}\|\boldsymbol{x} - \boldsymbol{y}\|^{p+1}.$$

With above bound on $\|\boldsymbol{x} - \boldsymbol{y}\|$ we conclude that (3.14) and (3.15) hold for $\mu_q' = \mu_q/2$. ∎

LEMMA 3.9. *Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be continuous and $\mathcal{L}_f = \{\boldsymbol{y} \in \mathbb{R}^n \mid f(\boldsymbol{y}) \leq c\}$ a sublevel set for some $c \in \mathbb{R}$. Every point $\boldsymbol{y}$ on the boundary[5] of $\mathcal{L}_f$ or on the boundary of any connected component[6] of $\mathcal{L}_f$ satisfies $f(\boldsymbol{y}) = c$.*

Proof. First, we show that $f(\boldsymbol{y}) = c$ for all $\boldsymbol{y} \in \partial\mathcal{L}_f$. Let $\boldsymbol{z} \in \mathbb{R}^n$ be such that $f(\boldsymbol{z}) > c$. By continuity, there exists a ball around $\boldsymbol{z}$ with the property that the value of f is larger than c for every element of this ball. Therefore, $\boldsymbol{z}$ lies in the interior of $\mathbb{R}^n \setminus \mathcal{L}_f$ and not on the boundary of $\mathcal{L}_f$. Similarly, every $\boldsymbol{z} \in \mathbb{R}^n$ with $f(\boldsymbol{z}) < c$ lies in the interior of $\mathcal{L}_f$ and not on the boundary. This proves the first part.

Now, consider any connected component $\mathcal{C}$ of $\mathcal{L}_f$. It suffices to show $\partial\mathcal{C} \subseteq \partial\mathcal{L}_f$. For any point $\boldsymbol{z} \notin \partial\mathcal{L}_f$ there exists a ball around $\boldsymbol{z}$ such that this ball is either entirely contained in $\mathcal{L}_f$ or entirely contained in $\mathbb{R}^n \setminus \mathcal{L}_f$. In the latter case, the ball is also contained in $\mathbb{R}^n \setminus \mathcal{C}$ and $\boldsymbol{z}$ does not lie on the boundary ∂C. In the former case, since every ball is connected, it is contained in one connected component of $\mathcal{L}_f$. It is therefore either contained in $\mathcal{C}$ or contained in $\mathbb{R}^n \setminus \mathcal{C}$ and again $\boldsymbol{z} \notin \partial C$. ∎

With these prerequisites, we can now prove that in particular local model minimizers close to $\boldsymbol{x}_*$ are minimizers that will lead to a successful iteration.

LEMMA 3.10. *Let $f \in C^p_{\text{Lip}}$, $\boldsymbol{x}_* \in M^q(f)$ and $p > q - 1$. There exists radii $0 < r_x \leq r_y \leq r_q$ such that for any $\boldsymbol{x}_k \in B(\boldsymbol{x}_*, r_x)$ there exists a local minimizer of m_k inside $B(\boldsymbol{x}_*, r_y)$ and any approximate minimizer $\boldsymbol{y}_k$ of m_k in the sense of (2.1) inside $B(\boldsymbol{x}_*, r_y)$ has a success ratio $\rho_k \geq \eta$.*

Proof. We first derive bounds for $\boldsymbol{x}_k$ and $\boldsymbol{y}_k$ such that $\rho_k \geq \eta$, assuming that $\boldsymbol{y}_k$ is an approximate minimizer of m_k satisfying (2.1). It suffices to show that $|\rho_k - 1| \leq 1 - \eta$. Since

$$|\rho_k - 1| = \left| \frac{f(\boldsymbol{x}_k) - f(\boldsymbol{y}_k)}{t_k(\boldsymbol{x}_k) - t_k(\boldsymbol{y}_k)} - 1 \right| = \frac{|t_k(\boldsymbol{y}_k) - f(\boldsymbol{y}_k)|}{|t_k(\boldsymbol{x}_k) - t_k(\boldsymbol{y}_k)|} \tag{3.16}$$

we need to find an upper bound for the numerator and a lower bound for the denominator. The upper bound follows directly from (2.3a).

[5] The topological boundary of a set are all points such that every open neighbourhood of the point contains at least one point in the set and at least one point not in the set.

[6] The term *connected component* refers to maximally connected subsets in the topological sense, see for example [14, Section 4.5].

Take the derivative of (2.2b), multiply with $\boldsymbol{x}_k - \boldsymbol{y}_k$ and rearrange to obtain

$$\begin{aligned}\nabla t_k(\boldsymbol{y}_k)^T(\boldsymbol{x}_k - \boldsymbol{y}_k) &= \nabla m_k(\boldsymbol{y}_k)^T(\boldsymbol{x}_k - \boldsymbol{y}_k) + (p+1)\sigma_k\|\boldsymbol{y}_k - \boldsymbol{x}_k\|^{p+1} \\ &\geq -\|\nabla m_k(\boldsymbol{y}_k)\|\|\boldsymbol{x}_k - \boldsymbol{y}_k\| \geq -\theta\|\boldsymbol{x}_k - \boldsymbol{y}_k\|^{p+1}.\end{aligned}$$

Here, the last inequality used that $\boldsymbol{y}_k$ is an approximate minimizer in the sense of (2.1). By Lemma 3.8 there is a ball of radius $r_q' > 0$ around $\boldsymbol{x}_*$ such that (3.14) and (3.15) hold inside that ball. Assume $\|\boldsymbol{x}_k - \boldsymbol{y}_k\| \leq \sqrt[p-q+1]{\frac{\mu_q'}{2q\theta}}$ and $\boldsymbol{x}_k, \boldsymbol{y}_k \in B(\boldsymbol{x}_*, r_q')$, then combining (3.14) and hte previous inequality we get

$$\begin{aligned}t_k(\boldsymbol{x}_k) - t_k(\boldsymbol{y}_k) &\geq \nabla t_k(\boldsymbol{y}_k)^T(\boldsymbol{x}_k - \boldsymbol{y}_k) + \frac{\mu_q'}{q}\|\boldsymbol{x}_k - \boldsymbol{y}_k\|^q \\ &\geq \left(\frac{\mu_q'}{q} - \theta\|\boldsymbol{x}_k - \boldsymbol{y}_k\|^{p-q+1}\right)\|\boldsymbol{x}_k - \boldsymbol{y}_k\|^q \geq \frac{\mu_q'}{2q}\|\boldsymbol{x}_k - \boldsymbol{y}_k\|^q.\end{aligned}$$

Therefore, the numerator in (3.16) is bounded by an $O(\|\boldsymbol{x}_k - \boldsymbol{y}_k\|^{p+1})$ term and the denominator by an $O(\|\boldsymbol{x}_k - \boldsymbol{y}_k\|^q)$ term. This means

$$|\rho_k - 1| = \frac{|t_k(\boldsymbol{y}_k) - f(\boldsymbol{y}_k)|}{|t_k(\boldsymbol{x}_k) - t_k(\boldsymbol{y}_k)|} \leq \frac{L_p}{(p+1)!}\frac{2q}{\mu_q'}\|\boldsymbol{x}_k - \boldsymbol{y}_k\|^{p-q+1} \leq 1 - \eta$$

if additionally $\|\boldsymbol{x}_k - \boldsymbol{y}_k\| \leq \sqrt[p-q+1]{(1-\eta)\frac{(p+1)!}{L_p}\frac{\mu_q'}{2q}}$. Overall, any approximate minimizer $\boldsymbol{y}_k$ leads to a successful iteration whenever $\boldsymbol{x}_k, \boldsymbol{y}_k \in B(\boldsymbol{x}_*, r_y)$ for

$$r_y = \min\left\{r_q',\ \frac{1}{2}\sqrt[p-q+1]{\frac{\mu_q'}{2q\theta}},\ \frac{1}{2}\sqrt[p-q+1]{(1-\eta)\frac{(p+1)!}{L_p}\frac{\mu_q'}{2q}}\right\}.$$

What is left to show is that for $\boldsymbol{x}_k \in B(\boldsymbol{x}_*, r_x)$ such a minimizer exists in the first place. When $\boldsymbol{x}_k$ and $\boldsymbol{y}$ are close enough to $\boldsymbol{x}_*$ then, using (3.15), we can derive a lower bound for m_k as

$$m_k(\boldsymbol{y}) \geq t_k(\boldsymbol{y}) \geq t_k(\boldsymbol{x}_k) + \nabla t_k(\boldsymbol{x}_k)^T(\boldsymbol{y} - \boldsymbol{x}_k) + \frac{\mu_q'}{q}\|\boldsymbol{y} - \boldsymbol{x}_k\|^q =: \underline{m}_k(\boldsymbol{y}). \tag{3.17}$$

Independently of whether the bound holds for any local minimizer of m_k, observe that $\underline{m}_k(\boldsymbol{x}_k) = m_k(\boldsymbol{x}_k)$, that $\underline{m}_k$ is convex, and that the sublevel set

$$\underline{\mathcal{L}}_k = \{\boldsymbol{y} \in \mathbb{R}^n \mid \underline{m}_k(\boldsymbol{y}) \leq m_k(\boldsymbol{x}_k)\}$$

is contained in a closed ball of radius $\sqrt[q-1]{\frac{q}{\mu_q'}\|\nabla t_k(\boldsymbol{x}_k)\|}$ around $\boldsymbol{x}_k$, because

$$\underline{m}_k(\boldsymbol{y}) \geq \underbrace{t_k(\boldsymbol{x}_k)}_{=m_k(\boldsymbol{x}_k)} + \underbrace{\left(-\|\nabla t_k(\boldsymbol{x}_k)\| + \frac{\mu_q'}{q}\|\boldsymbol{y} - \boldsymbol{x}_k\|^{q-1}\right)}_{>0}\|\boldsymbol{y} - \boldsymbol{x}_k\| > m_k(\boldsymbol{x}_k). \tag{3.18}$$

for all $\boldsymbol{y}$ outside of that ball. We need to make sure that $\underline{\mathcal{L}}_k$ is contained in $B(\boldsymbol{x}_*, r_q')$ such that (3.17) holds for points in the sublevel set. Let $\boldsymbol{x}_k \in B(\boldsymbol{x}_*, r_q')$ and $\boldsymbol{y} \in \underline{\mathcal{L}}_k$, then the distance of $\boldsymbol{y}$ to $\boldsymbol{x}_*$ can be bounded by the triangle inequality and

$$\|\boldsymbol{y} - \boldsymbol{x}_k\| \leq \sqrt[q-1]{\frac{q}{\mu_q'}\|\nabla t_k(\boldsymbol{x}_k)\|} \leq \sqrt[q-1]{\frac{q\nu}{\mu_q'}\|\boldsymbol{x}_k - \boldsymbol{x}_*\|}$$

since $\|\nabla t_k(\boldsymbol{x}_k)\| = \|\nabla f(\boldsymbol{x}_k)\| \leq \nu\|\boldsymbol{x}_k - \boldsymbol{x}_*\|$ by Lemma 3.2. For

$$r_x = \min\left\{r'_q,\ \frac{\mu'_q}{q\nu}\left(\frac{r_y}{2}\right)^{q-1},\ \frac{r_y}{2}\right\} \tag{3.19}$$

and any expansion point $\boldsymbol{x}_k \in B(\boldsymbol{x}_*, r_x)$ we know that any point $\boldsymbol{y} \in \underline{\mathcal{L}}_k$ is contained in $B(\boldsymbol{x}_*, r_y) \subseteq B(\boldsymbol{x}_*, r'_q)$ which implies that (3.17) holds for all such points. Since $\underline{m}_k$ is convex, $\underline{\mathcal{L}}_k$ is a closed convex set and $\boldsymbol{x}_k$ lies on its boundary. We always assume that the gradient of f at any iterate $\boldsymbol{x}_k$ is nonzero, i.e. that $\nabla \underline{m}_k(\boldsymbol{x}_k) = \nabla t_k(\boldsymbol{x}_k) = \nabla m_k(\boldsymbol{x}_k) = \nabla f(\boldsymbol{x}_k) \neq \boldsymbol{0}$. This implies that there are points $\boldsymbol{y}$ in the interior of $\underline{\mathcal{L}}_k$ where $m_k(\boldsymbol{y}) < m_k(\boldsymbol{x}_k)$. By compactness, m_k attains its minimum on $\underline{\mathcal{L}}_k$, and the minimizer is not on the boundary, since by Lemma 3.9 any point $\boldsymbol{y}$ on the boundary satisfies $m_k(\boldsymbol{y}) \geq \underline{m}_k(\boldsymbol{y}) = m_k(\boldsymbol{x}_k)$. The point at which the minimum is attained must therefore lie in the interior of $\underline{\mathcal{L}}_k$ and be a local minimizer of m_k which is contained in $B(\boldsymbol{x}_*, r_y)$. This concludes the proof. ∎

The way of characterizing successful model minimizers in Lemma 3.10 has a direct dependence on the function minimizer $\boldsymbol{x}_*$, which the algorithm aims to find and is thus a priori unknown. Alternatively, it is also possible to show that (asymptotically) it suffices to choose a local model minimizer from the correct connected component of the model sublevel set. Clearly, by (2.1) the tentative iterate $\boldsymbol{y}_k$ lies in the sublevel set $\{\boldsymbol{y} \in \mathbb{R}^n \mid m_k(\boldsymbol{y}) \leq m_k(\boldsymbol{x}_k)\}$. Since the local model does not have to be convex, there can be multiple connected components of this sublevel set. Selecting a minimizer from the connected component that contains $\boldsymbol{x}_k$ avoids spurious model minimizers far away from $\boldsymbol{x}_*$ and ensures a successful step.

LEMMA 3.11. *Let $f \in C^p_{\text{Lip}}$, $\boldsymbol{x}_* \in M^q(f)$ and $p > q - 1$. Moreover, let $\mathcal{C}_k$ be the connected component of the sublevel set $\mathcal{L}_k := \{\boldsymbol{y} \in \mathbb{R}^n \mid m_k(\boldsymbol{y}) \leq m_k(\boldsymbol{x}_k)\}$ that contains $\boldsymbol{x}_k$. There exists a radius $r_c > 0$ such that for any $\boldsymbol{x}_k \in B(\boldsymbol{x}_*, r_c)$ there exists a local minimizer of m_k inside $\mathcal{C}_k$ and any approximate minimizer $\boldsymbol{y}_k$ of m_k in the sense of (2.1) inside $\mathcal{C}_k$ has a success ratio $\rho_k \geq \eta$.*

Proof. The assumptions are the same as for Lemma 3.10, so there exists a radii $r_x, r_y > 0$ such that all approximate minimizers $\boldsymbol{y}_k$ inside $B(\boldsymbol{x}_*, r_y)$ have a success ratio $\rho_k \geq \eta$ whenever $\boldsymbol{x}_k \in B(\boldsymbol{x}_*, r_x)$. Moreover, following the proof of that Lemma, we see that, on a ball of radius $r'_q > 0$ around $\boldsymbol{x}_*$, equation (3.17) holds and that $\underline{m}_k$ is a convex lower bound for m_k. By (3.18) this lower bound is strictly larger than $m_k(\boldsymbol{x}_k)$ outside the ball $B(\boldsymbol{x}_k,\ \sqrt[q-1]{\frac{q}{\mu'_q}\|\nabla t_k(\boldsymbol{x}_k)\|})$. In other words,

$$\mathcal{L}_k \cap \left(B(\boldsymbol{x}_*, r'_q) \setminus B\left(\boldsymbol{x}_k,\ \sqrt[q-1]{\frac{q}{\mu'_q}\|\nabla t_k(\boldsymbol{x}_k)\|}\right)\right) = \emptyset.$$

Take $r_c \in (0, r_x)$ with r_x as defined in (3.19) and assume $\boldsymbol{x}_k \in B(\boldsymbol{x}_*, r_c)$ from now on. The radius r_c is chosen small enough such that the previous equation implies $\mathcal{L}_k \cap \partial B(\boldsymbol{x}_*, r'_q) = \emptyset$. Thus, every connected component is either completely contained in the interior of $B(\boldsymbol{x}_*, r'_q)$ or completely contained in $\mathbb{R}^n \setminus B(\boldsymbol{x}_*, r'_q)$. Otherwise, the connected component could be partitioned into two sets, each the intersection of the component with an open set, which contradicts the connectedness of the connected component. Clearly, $\mathcal{C}_k$ is contained in $B(\boldsymbol{x}_*, r'_q)$ and thus

$$\mathcal{C}_k \subseteq B\left(\boldsymbol{x}_k,\ \sqrt[q-1]{\frac{q}{\mu'_q}\|\nabla t_k(\boldsymbol{x}_k)\|}\right) \subseteq B(\boldsymbol{x}_*, r_y). \tag{3.20}$$

Every approximate minimizer $\boldsymbol{y}_k \in \mathcal{C}_k$ must therefore satisfy $\rho_k \geq \eta$.

We now turn towards proving that there always exists a local minimizer of m_k in $\mathcal{C}_k$. By (3.20) we know that $\mathcal{C}_k$ is bounded and since every connected component of a closed set is closed [14, Corollary 5.6, Chapter 4], $\mathcal{C}_k$ must be compact. Therefore, the continuous function m_k attains its minimum on $\mathcal{C}_k$ and by Lemma 3.9 every point on the boundary of $\mathcal{C}_k$ has the same function value as $\boldsymbol{x}_k$. At the same time, the fact that $\nabla m_k(\boldsymbol{x}_k) = \nabla f(\boldsymbol{x}_k) \neq \boldsymbol{0}$ implies the existence of a point with a lower function value in the interior of $\mathcal{C}_k$. This means the minimum of m_k on $\mathcal{C}_k$ is attained in the interior of $\mathcal{C}_k$ and so this point is a local minimizer of m_k. ∎

THEOREM 3.12. *Let $f \in C^p_{\text{Lip}}$, $\boldsymbol{x}_* \in M^q(f)$ and $p > q-1$. Assume that $\boldsymbol{y}_k$ is always chosen such that $\boldsymbol{y}_k \in B(\boldsymbol{x}_*, r_y)$, with r_y as defined in Lemma 3.10, or always chosen in a way that $\boldsymbol{y}_k \in \mathcal{C}_k$, with $\mathcal{C}_k$ as defined in Lemma 3.11. If additionally, $\boldsymbol{x}_0$ is close enough to $\boldsymbol{x}_*$, then all iterations of Algorithm 2.1 are successful and*

$$\begin{aligned} \|\nabla f(\boldsymbol{x}_k)\| &\to 0 && \text{at a } Q\text{-}\tfrac{p}{q-1}\text{th-order rate} \\ f(\boldsymbol{x}_k) &\to f(\boldsymbol{x}_*) && \text{at a } Q\text{-}\tfrac{p}{q-1}\text{th-order rate} \\ \boldsymbol{x}_k &\to \boldsymbol{x}_* && \text{at an } R\text{-}\tfrac{p}{q-1}\text{th-order rate} \end{aligned}$$

as $k \to \infty$.

Proof. Note that in the proof of Lemma 3.11, the radius $r_c > 0$ is chosen small enough such that for any $\boldsymbol{x}_k \in B(\boldsymbol{x}_*, r_c)$ the component satisfies $\mathcal{C}_k \subseteq B(\boldsymbol{x}_*, r_y)$. Therefore, for $\boldsymbol{x}_k$ close enough to $\boldsymbol{x}_*$ the assumption that $\boldsymbol{y}_k \in B(\boldsymbol{x}_*, r_y)$ is the weaker one, and it suffices to prove the result for this assumption.

Let r_x and r_y be as defined in Lemma 3.10. We will first need to show inductively that, for $\boldsymbol{x}_0$ close enough to $\boldsymbol{x}_*$, all $\boldsymbol{x}_k \in B(\boldsymbol{x}_*, r_x)$.[7] Choose $\boldsymbol{x}_0 \in B(\boldsymbol{x}_*, r_x)$ close enough to $\boldsymbol{x}_*$ such that

$$\|\nabla f(\boldsymbol{x}_0)\| \leq \min\left\{ \left(\frac{1}{R_{\max}} \left(\frac{\mu_q}{q} \right)^{\frac{p}{q-1}} \right)^{\frac{q-1}{p-q+1}}, \ \mu_q r_x^{q-1} \right\}$$

For the induction step assume $\boldsymbol{x}_k \in B(\boldsymbol{x}_*, r_x)$ is such that $\|\nabla f(\boldsymbol{x}_k)\| \leq \|\nabla f(\boldsymbol{x}_0)\|$. By Lemma 3.10 there exists an approximate model minimizer $\boldsymbol{y}_k \in B(\boldsymbol{x}_*, r_y)$ and by assumption such a point is chosen. This means that the iteration is successful and $\boldsymbol{x}_k, \boldsymbol{x}_{k+1} = \boldsymbol{y}_k \in B(\boldsymbol{x}_*, r_q)$, i.e., both points lie in the region where f is uniformly convex. Apply Lemma 3.3 to obtain

$$\|\nabla f(\boldsymbol{x}_{k+1})\| \leq R_{\max} \left(\frac{q}{\mu_q} \right)^{\frac{p}{q-1}} \|\nabla f(\boldsymbol{x}_k)\|^{\frac{p-q+1}{q-1}} \|\nabla f(\boldsymbol{x}_k)\|,$$

which implies $\|\nabla f(\boldsymbol{x}_{k+1})\| \leq \|\nabla f(\boldsymbol{x}_k)\| \leq \|\nabla f(\boldsymbol{x}_0)\|$. Therefore, by (3.4)

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_*\| \leq \sqrt[q-1]{\frac{\|\nabla f(\boldsymbol{x}_{k+1})\|}{\mu_q}} \leq \sqrt[q-1]{\frac{\|\nabla f(\boldsymbol{x}_0)\|}{\mu_q}} \leq r_x.$$

We conclude that $\boldsymbol{x}_{k+1} \in B(\boldsymbol{x}_*, r_x)$ and satisfies $\|\nabla f(\boldsymbol{x}_{k+1})\| \leq \|\nabla f(\boldsymbol{x}_0)\|$, thus completing the induction. As mentioned before, all iterations are successful and the rates follow directly from Theorem 3.5. ∎

[7] For the $\boldsymbol{y}_k \in \mathcal{C}_k$ assumption based on Lemma 3.11 we would of course need to prove that all iterates stay in $B(\boldsymbol{x}_*, r_c)$. The argument is the same after replacing r_x with r_c.

To conclude this section, we comment on different aspects of the results above. First, Theorem 3.12 implies that for fully adaptive methods there is a way to ensure that asymptotically all iterations are successful. However, the way to do so is not, as one might expect, to choose the global model minimizer, as Example 2.3 shows. Instead, when faced with ambiguity one should select a local minimizer of the model that lies in the vicinity of $\boldsymbol{x}_*$. While neither Lemma 3.10 nor Lemma 3.11 give characterizations of the right model minimizer that lead to practical algorithms for finding such minimizers, Lemma 3.11 suggests that any monotonically decreasing local optimization algorithm applied to m_k and starting at $\boldsymbol{x}_k$ is unlikely compute a wrong minimizer. To do so such a local method would need to compute a large enough step in the right direction to jump over a hill in m_k in order to move from $\mathcal{C}_k$ to a different connected component of the sublevel set.

Second, note that neither in the statement nor in the proof of Lemma 3.10 there is an assumption or dependence on σ_k, other that $\sigma_k \geq 0$. The existence of this local model minimizer close to $\boldsymbol{x}_*$ is even true for $\sigma_k = 0$. Thus, when selecting $\boldsymbol{y}_k$ in the appropriate way, it is possible to drop the regularization parameter to zero once $\boldsymbol{x}_k$ is close enough to $\boldsymbol{x}_*$ and still achieve the same convergence guarantees. Unfortunately, without regularization the subproblem only reduces to solving a single linear system in the second-order case. For third- and higher-order algorithms even minimizing the unregularized Taylor expansion is often done with general local optimization methods and gets more complicated as p increases.

Lastly, the claims and proofs of this section can easily be extended from functions with Lipschitz continuous pth derivatives to ones with Hölder continuous pth derivatives, i.e. those that satisfy

$$\|\nabla^p f(\boldsymbol{x}) - \nabla^p f(\boldsymbol{y})\| \leq H_p \|\boldsymbol{x} - \boldsymbol{y}\|^{\beta_p}$$

for some constants $H_p > 0$, $\beta_p \in [0, 1]$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. Let AR$p, r$ be the variant of Algorithm 2.1 where the exponent of the regularization term is r and the exponent in the conditions for an approximate local minimizer (2.1) is $r-1$. Applying AR$p, p+\beta_p$ to functions that satisfy the conditions above gives $\frac{p+\beta_p-1}{q-1}$-order convergence (compared to $\frac{p}{q-1}$ in Theorem 3.5) whenever $p + \beta_p > q$. Theorems 3.6 and 3.12 change correspondingly.

From one perspective, the analysis under the assumption of Hölder continuity bridges the gaps and clarifies how smoothness ($p+\beta_p$) and convexity (q) interact. With the restriction to Lipschitz continuity the convergence rates seem to be improving in discrete steps as p increases, while the result above makes clear that actually these rates are part of a continuum, where the rates continuously improve as $p+\beta_p$ increases. Unfortunately, the $\frac{p+\beta_p-1}{q-1}$-order rate is only possibly with a priori knowledge of the β_p parameter as the value of β_p is used within the algorithm.[8]

From a second perspective, the Lipschitz continuous case is the most important case and the order of convergence does indeed improve in discrete steps as the algorithm gets access to more and more derivatives. When considering smooth functions

[8] An analysis of a variant of the adaptive regularization method when the regularization power r and the smoothness $p + \beta_p$ are not equal is provided by Cartis, Gould and Toint [3]. Their results show that the bound on the number of iterations does not degrade when $r \geq p + \beta_p$ and varies continuously with $p + \beta_p$. With this reasoning the authors suggest that $r = p + 1$ gives a universal algorithm that adapts to the parameter β_p of the function class at hand. Whether $(p + 1)$st-order regularization also leads to a universal algorithm in the context of local convergence, is an open question and outside the scope of this paper.

$f \in C^\infty$ and applying ARp for different p, then these functions have Lipschitz continuous pth derivatives on any bounded domain. With the additional assumption that the iterates stay bounded, we obtain rates that depend not on the smoothness itself, but on the number of derivatives the algorithm has access to, which can only be increased in discrete steps. This is why in this paper we focus on the Lipschitz continuity case and only mention the more general Hölder continuity case here.

4. Sharpness of local convergence results. Considering the $\frac{p}{q-1}$th-order convergence rates for successful iterations in Theorem 3.5 and the degraded rates in Theorem 3.6, it is natural to ask whether these rates are sharp and in which cases one might need to expect such rates. The example in subsection 4.1 shows that the order of convergence proved in Theorem 3.5 is indeed sharp and the main result in subsection 4.2 proves that in some cases oscillations in σ_k are unavoidable when $\boldsymbol{y}_k$ is the global model minimizer and σ_k is decreased during successful iterations.

4.1. $\frac{p}{q-1}$th-order convergence.

Example 4.1. Consider the one-dimensional function $f(x) = \frac{1}{q}x^q + \frac{1}{p+1}x^{p+1}$ for even $q \in \mathbb{N}$ and some $p \in \mathbb{N}$ with $p > q-1$. The function has a local minimizer at $x_* = 0$, its pth derivative is Lipschitz continuous with constant $L_p = p!$, and it is locally uniformly convex of order q. We choose $\theta = 3$ and the other parameters arbitrarily. Since $p \geq q$, the pth-order Taylor expansion of x^q is exact and the expansion of x^{p+1} is only missing the $(p+1)$st-order term. Therefore,

$$t_k(y) = f(y) - \frac{1}{p+1}(y-x_k)^{p+1} = \frac{1}{q}y^q + \frac{1}{p+1}y^{p+1} - \frac{1}{p+1}(y-x_k)^{p+1} \tag{4.1a}$$

$$t_k'(y) = y^{q-1} + y^p - (y-x_k)^p \tag{4.1b}$$

$$m_k'(y) = y^{q-1} + y^p - (y-x_k)^p + (p+1)\sigma_k|y-x_k|^{p-1}(y-x_k). \tag{4.1c}$$

At each iteration, the ARp algorithm computes an approximate minimizer y_k of m_k satisfying (2.1). We will show that

$$y_k = -|x_k|^{\frac{p}{q-1}}\frac{x_k}{|x_k|}$$

is an approximate minimizer as long as $|x_k|$ is small enough. Notice that the sign of y_k is always the opposite of the sign of x_k, which implies $|y_k - x_k| = |y_k| + |x_k|$. We can deduce that

$$|y_k^{q-1} + y_k^p| = \left|-|x_k|^p\frac{x_k}{|x_k|} + y_k^p\right| \leq |x_k|^p + |y_k|^p \leq (|x_k| + |y_k|)^p = |y_k - x_k|^p.$$

The other two terms in (4.1c) are bounded by $|y_k - x_k|^p$ and $(p+1)\sigma_k|y_k - x_k|^p$ respectively. Overall,

$$|m_k'(y_k)| \leq (2 + (p+1)\sigma_k)|y_k - x_k|^p \leq \theta|y_k - x_k|^p$$

for $\sigma_k \leq \frac{1}{p+1}$. This makes the tentative iterate y_k an approximate minimizer since $m_k(y_k) < m_k(x_k)$ as long as $|x_k|$ is small enough. Moreover, we can make $|x_k|$ small enough such that Lemma 3.10 implies that any approximate minimizer in a prescribed neighbourhood of $x_* = 0$ is successful and that y_k, satisfying $|y_k| = |x_k|^{\frac{p}{q-1}}$, lies in this neighbourhood. For $\sigma_0 < \frac{1}{p+1}$ and $|x_0|$ small enough, all iterations are thus successful and

$$|x_k| = |x_0|^{\left(\frac{p}{q-1}\right)^k}.$$

Therefore, x_k converges to x_* at a $\frac{p}{q-1}$th-order rate, which in turn implies that the order of convergence of $f'(x_k) \to 0$ and $f(x_k) \to f(x_*)$ is not higher by Lemma 3.2 (but also not lower by Theorem 3.5).

This example illustrates that the convergence rates derived in Theorems 3.5 and 3.12 are not an artefact of the analysis, but describe worst-case rates that can be attained.

4.2. Unsuccessful iterations. Example 2.3 already shows that there are cases where half of the iterations are unsuccessful even asymptotically, when $\boldsymbol{y}_k$ is the global model minimizer at each iteration and σ_k is decreased during successful iterations. Continuing with this global minimizer assumption, we want to make more precise in which cases we need to expect unsuccessful iterations, show that they occur for a wide variety of problems, and that the ratio of unsuccessful iterations to successful iterations is equal to $\alpha \geq 0$ where α describes the relationship between the increase and decrease factors for σ_k through $\gamma_1 = \gamma_2^{-\alpha}$. This then shows that the degradation from $\frac{p}{q-1}$th-order convergence rates for successful iterations to $\sqrt[1+\alpha]{\frac{p}{q-1}}$th-order rates in Theorem 3.6 are again sharp.

The key assumption in the following theorem is that $\boldsymbol{x}_*$ is a global minimizer of f but not of $t_{\boldsymbol{x}_*}$, the Taylor expansion at $\boldsymbol{x}_*$. The second part holds true, for example, when p is odd and the highest-order term is non-zero, because then the Taylor expansion is not lower bounded and has no global minimizer. When these two assumptions are satisfied, the smallest σ_k that guarantees a successful iteration and the largest σ_k that guarantees an unsuccessful iteration both converge to some value σ_* (Theorem 4.3). Therefore, the regularization parameters σ_k will eventually oscillate around σ_* (Theorem 4.4).

To prove this result we need some bound on the difference between the model functions when expanded at $\boldsymbol{x}_k$ and $\boldsymbol{x}_*$ to then exploit the fact that $\boldsymbol{x}_*$ is a local but not a global minimizer of the Taylor expansion at $\boldsymbol{x}_*$.

LEMMA 4.2. *Let $f \in C^p_{\text{Lip}}$. For any $r > 0$ and $\boldsymbol{x} \in \mathbb{R}^n$ there exists a constant $C_{\boldsymbol{x},r} > 0$ such that*

$$|m_{\boldsymbol{x},\sigma}(\boldsymbol{y}) - m_{\boldsymbol{x}',\sigma}(\boldsymbol{y})| \leq (C_{\boldsymbol{x},r} + 2^p(p+1)\sigma)\|\boldsymbol{x} - \boldsymbol{x}'\|\|\boldsymbol{y} - \boldsymbol{x}\|^p \text{ and} \tag{4.2a}$$

$$|m_{\boldsymbol{x},\sigma}(\boldsymbol{y}) - m_{\boldsymbol{x}',\sigma}(\boldsymbol{y})| \leq (C_{\boldsymbol{x},r} + 2^p(p+1)\sigma)\|\boldsymbol{x} - \boldsymbol{x}'\|\|\boldsymbol{y} - \boldsymbol{x}'\|^p \tag{4.2b}$$

hold for all $\boldsymbol{x}' \in B(\boldsymbol{x}, r)$ and $\boldsymbol{y} \notin B(\boldsymbol{x}, 2r)$.

Proof. To prove this bound on the difference between models with different expansion points, we need to first establish some simple bounds between differences of rank-one tensors and differences between $(p+1)$st-order regularization terms. For any $\boldsymbol{x}, \boldsymbol{x}', \boldsymbol{y} \in \mathbb{R}^n$ and $j \in \mathbb{N}_{\geq 0}$ we have

$$\|\otimes^j(\boldsymbol{y} - \boldsymbol{x})^T - \otimes^j(\boldsymbol{y} - \boldsymbol{x}')^T\| \leq j\|\boldsymbol{x} - \boldsymbol{x}'\| \max(\|\boldsymbol{y} - \boldsymbol{x}\|, \|\boldsymbol{y} - \boldsymbol{x}'\|)^{j-1}. \tag{4.3}$$

This follows from the following telescoping sum

$$\otimes^j (\boldsymbol{y} - \boldsymbol{x})^T - \otimes^j(\boldsymbol{y} - \boldsymbol{x}')^T = \sum_{i=0}^{j-1} \otimes^{j-i-1}(\boldsymbol{y} - \boldsymbol{x})^T \otimes (\boldsymbol{x}' - \boldsymbol{x})^T \otimes^i (\boldsymbol{y} - \boldsymbol{x}')^T$$

by taking norms and bounding each term individually. Note that when $j = 0$ the "tensors" involved are outer products of zero vectors, i.e. the scalar 1, so the difference

between them is zero and so is the right-hand side of (4.3). In the same way,

$$\big|\|\boldsymbol{y}-\boldsymbol{x}\|^{p+1} - \|\boldsymbol{y}-\boldsymbol{x}'\|^{p+1}\big| \leq (p+1)\|\boldsymbol{x}-\boldsymbol{x}'\| \max(\|\boldsymbol{y}-\boldsymbol{x}\|, \|\boldsymbol{y}-\boldsymbol{x}'\|)^p$$

can be derived from a telescoping sum using the reverse triangle inequality $|\|\boldsymbol{y}-\boldsymbol{x}\| - \|\boldsymbol{y}-\boldsymbol{x}'\|| \leq \|\boldsymbol{x}-\boldsymbol{x}'\|$.

Let us now turn to proving the original statement, by considering the difference between the Taylor expansions.

$$t_{\boldsymbol{x}}(\boldsymbol{y}) - t_{\boldsymbol{x}'}(\boldsymbol{y}) = \sum_{j=0}^{p} \frac{1}{j!}\big(\langle \nabla^j f(\boldsymbol{x}), \otimes^j(\boldsymbol{y}-\boldsymbol{x})^T\rangle_F - \langle \nabla^j f(\boldsymbol{x}'), \otimes^j(\boldsymbol{y}-\boldsymbol{x}')^T\rangle_F\big)$$

Ignoring the $\frac{1}{j!}$ coefficient, each summand on the right-hand side can be written as

$$\langle \nabla^j f(\boldsymbol{x}), \otimes^j(\boldsymbol{y}-\boldsymbol{x})^T - \otimes^j(\boldsymbol{y}-\boldsymbol{x}')^T\rangle_F + \langle \nabla^j f(\boldsymbol{x}) - \nabla^j f(\boldsymbol{x}'), \otimes^j(\boldsymbol{y}-\boldsymbol{x}')^T\rangle_F. \tag{4.4}$$

For $0 \leq j \leq p$ let $M_{\boldsymbol{x},j} = \|\nabla^j f(\boldsymbol{x})\|$ and let $L_{\boldsymbol{x},r,j}$ be the local Lipschitz constant of $\nabla^j f$ on $B(\boldsymbol{x}, r)$. The local Lipschitz constants exist and are finite because $\nabla^{j+1} f$ is continuous for $0 \leq j \leq p-1$ and because $\nabla^p f$ is Lipschitz continuous by assumption. Using these constants, the term in (4.4) is bounded by

$$M_{\boldsymbol{x},j} j \|\boldsymbol{x}-\boldsymbol{x}'\| \max(\|\boldsymbol{y}-\boldsymbol{x}\|, \|\boldsymbol{y}-\boldsymbol{x}'\|)^{j-1} + L_{\boldsymbol{x},r,j}\|\boldsymbol{x}-\boldsymbol{x}'\| \max(\|\boldsymbol{y}-\boldsymbol{x}\|, \|\boldsymbol{y}-\boldsymbol{x}'\|)^j.$$

Thus, overall we get

$$|m_{\boldsymbol{x},\sigma}(\boldsymbol{y}) - m_{\boldsymbol{x}',\sigma}(\boldsymbol{y})| \tag{4.5a}$$

$$\leq |t_{\boldsymbol{x}}(\boldsymbol{y}) - t_{\boldsymbol{x}'}(\boldsymbol{y})| + (p+1)\sigma\|\boldsymbol{x}-\boldsymbol{x}'\| \max(\|\boldsymbol{y}-\boldsymbol{x}\|, \|\boldsymbol{y}-\boldsymbol{x}'\|)^p \tag{4.5b}$$

$$\leq \sum_{j=0}^{p} \frac{L_{\boldsymbol{x},r,j} + M_{\boldsymbol{x},j+1}}{j!}\|\boldsymbol{x}-\boldsymbol{x}'\| \max(\|\boldsymbol{y}-\boldsymbol{x}\|, \|\boldsymbol{y}-\boldsymbol{x}'\|)^j \tag{4.5c}$$
$$+ (p+1)\sigma\|\boldsymbol{x}-\boldsymbol{x}'\| \max(\|\boldsymbol{y}-\boldsymbol{x}\|, \|\boldsymbol{y}-\boldsymbol{x}'\|)^p$$

where $M_{\boldsymbol{x},p+1}$ is defined to be zero. This result holds for all $\boldsymbol{y} \in \mathbb{R}^n$.

The simplified bounds in the statement of this lemma can be obtained using $\boldsymbol{y} \notin B(\boldsymbol{x}, 2r)$, since then

$$\|\boldsymbol{y}-\boldsymbol{x}'\| \leq \|\boldsymbol{y}-\boldsymbol{x}\| + \|\boldsymbol{x}-\boldsymbol{x}'\| \leq \|\boldsymbol{y}-\boldsymbol{x}\| + r < 2\|\boldsymbol{y}-\boldsymbol{x}\| \text{ and}$$
$$\|\boldsymbol{y}-\boldsymbol{x}\| < \|\boldsymbol{y}-\boldsymbol{x}\| + \|\boldsymbol{y}-\boldsymbol{x}\| - 2\|\boldsymbol{x}-\boldsymbol{x}'\| = 2(\|\boldsymbol{y}-\boldsymbol{x}\| - \|\boldsymbol{x}-\boldsymbol{x}'\|) \leq 2\|\boldsymbol{y}-\boldsymbol{x}'\|.$$

Using $\max(\|\boldsymbol{y}-\boldsymbol{x}\|, \|\boldsymbol{y}-\boldsymbol{x}'\|) > 2r$ and the first of these inequalities gives

$$\max(\|\boldsymbol{y}-\boldsymbol{x}\|, \|\boldsymbol{y}-\boldsymbol{x}'\|)^j \leq (2r)^{j-p} \max(\|\boldsymbol{y}-\boldsymbol{x}\|, \|\boldsymbol{y}-\boldsymbol{x}'\|)^p \leq 2^j r^{j-p}\|\boldsymbol{y}-\boldsymbol{x}\|^p \tag{4.6}$$

for all $0 \leq j \leq p$, and analogously the second inequality implies that the left-hand side is upper-bounded by $2^j r^{j-p}\|\boldsymbol{y}-\boldsymbol{x}'\|^p$ as well.

Lastly, combine (4.5) and (4.6) and choose

$$C_{\boldsymbol{x},r} = \sum_{j=0}^{p} \frac{2^j}{r^{p-j}} \frac{L_{\boldsymbol{x},r,j} + M_{\boldsymbol{x},j+1}}{j!}$$

to obtain (4.2). This proves the claim. ∎

THEOREM 4.3. *Let $f \in C^p_{\text{Lip}}$, $\boldsymbol{x}_* \in M^q(f)$ and $p > q-1$. Assume that $\boldsymbol{x}_*$ is a global minimizer of f but not a global minimizer of $t^p_{\boldsymbol{x}_*}$. There exists a value $\sigma_* > 0$ (independent of η, γ_1, and γ_2) such that for any $\varepsilon > 0$ there exists a radius $r_\varepsilon > 0$ with the property that for any $\boldsymbol{x}_k \in B(\boldsymbol{x}_*, r_\varepsilon)$ the iteration will be unsuccessful if $\sigma_k < \sigma_* - \varepsilon$ and successful if $\sigma_k > \sigma_* + \varepsilon$, given that $\boldsymbol{y}_k$ is the global minimizer of m_k.*

Proof. We will show that the claim holds for the following threshold value of the regularization parameter:

$$\sigma_* = \inf\left\{ \sigma > 0 \;\middle|\; \min_{\boldsymbol{y}\in\mathbb{R}^n} m_{\boldsymbol{x}_*,\sigma}(\boldsymbol{y}) = f(\boldsymbol{x}_*) \right\}.$$

Since $\boldsymbol{x}_*$ is not a global minimizer of $t^p_{\boldsymbol{x}_*}$ there exists a point $\boldsymbol{y}$ such that $t^p_{\boldsymbol{x}_*}(\boldsymbol{y}) < t^p_{\boldsymbol{x}_*}(\boldsymbol{x}_*) = f(\boldsymbol{x}_*)$. If σ is small enough, the regularization term $\sigma\|\boldsymbol{y}-\boldsymbol{x}_*\|^p$ will be smaller than the difference between the two values of the Taylor expansion, therefore $\sigma_* > 0$. Equation (2.3a) implies that with $\sigma \geq L_p/(p+1)!$ the model $m_{\boldsymbol{x}_*,\sigma}$ overestimates the objective function everywhere. Combining this fact with $m_{\boldsymbol{x}_*,\sigma}(\boldsymbol{x}_*) = f(\boldsymbol{x}_*)$ and the assumption that $\boldsymbol{x}_*$ is a global minimizer of f we get $\sigma_* < \infty$.

Let $\varepsilon > 0$ be given. We consider the case $\sigma_k < \sigma_* - \varepsilon$ first. By definition of σ_* we know that there exists a point $\boldsymbol{y}_\varepsilon$ with the property that $m_{\boldsymbol{x}_*,\sigma_*-\varepsilon}(\boldsymbol{y}_\varepsilon) < f(\boldsymbol{x}_*)$. Let the difference between these two values be c and assume $\|\boldsymbol{x}_k - \boldsymbol{x}_*\| \leq r < \|\boldsymbol{y}_\varepsilon - \boldsymbol{x}_*\|/2$, then the global minimizer $\boldsymbol{y}_k$ of m_k satisfies

$$\begin{aligned} m_k(\boldsymbol{y}_k) &\leq m_k(\boldsymbol{y}_\varepsilon) = m_{\boldsymbol{x}_k,\sigma_k}(\boldsymbol{y}_\varepsilon) < m_{\boldsymbol{x}_k,\sigma_*-\varepsilon}(\boldsymbol{y}_\varepsilon) \\ &\leq m_{\boldsymbol{x}_*,\sigma_*-\varepsilon}(\boldsymbol{y}_\varepsilon) + (C_{\boldsymbol{x}_*,r} + 2^p(p+1)(\sigma_* - \varepsilon))\|\boldsymbol{x}_k - \boldsymbol{x}_*\|\|\boldsymbol{y}_\varepsilon - \boldsymbol{x}_*\|^p \\ &\leq m_{\boldsymbol{x}_*,\sigma_*-\varepsilon}(\boldsymbol{y}_\varepsilon) + \tfrac{c}{2} \leq f(\boldsymbol{x}_*) - \tfrac{c}{2} \end{aligned}$$

if $\|\boldsymbol{x}_k - \boldsymbol{x}_*\|$ is small enough using Lemma 4.2. This implies

$$\rho_k = \frac{f(\boldsymbol{x}_k) - f(\boldsymbol{y}_k)}{t_k(\boldsymbol{x}_k) - t_k(\boldsymbol{y}_k)} \leq \frac{f(\boldsymbol{x}_k) - f(\boldsymbol{x}_*)}{f(\boldsymbol{x}_k) - m_k(\boldsymbol{y}_k)} \leq \frac{f(\boldsymbol{x}_k) - f(\boldsymbol{x}_*)}{f(\boldsymbol{x}_k) - f(\boldsymbol{x}_*) + \frac{c}{2}} < \eta$$

whenever $f(\boldsymbol{x}_k) - f(\boldsymbol{x}_*) < \frac{\eta}{1-\eta}\frac{c}{2}$. Overall, as long as $\|\boldsymbol{x}_k - \boldsymbol{x}_*\|$ is small enough (which also means that $f(\boldsymbol{x}_k) - f(\boldsymbol{x}_*)$ becomes arbitrarily small) any iteration with $\sigma_k < \sigma_* - \varepsilon$ will be unsuccessful.

Next, consider the case $\sigma_k > \sigma_* + \varepsilon$. Choose r_x and r_y such that they satisfy the properties in Lemma 3.10 and let $r_x < r_y/2$. Apply Lemma 4.2 to $\boldsymbol{x}_k \in B(\boldsymbol{x}_*, r_x)$ and $\boldsymbol{y} \notin B(\boldsymbol{x}_*, r_y)$ to see that

$$\begin{aligned} &m_k(\boldsymbol{y}) = m_{\boldsymbol{x}_k,\sigma_k}(\boldsymbol{y}) > m_{\boldsymbol{x}_k,\sigma_*}(\boldsymbol{y}) + \varepsilon\|\boldsymbol{y} - \boldsymbol{x}_k\|^{p+1} \\ &\geq m_{\boldsymbol{x}_*,\sigma_*}(\boldsymbol{y}) + (\varepsilon\|\boldsymbol{y} - \boldsymbol{x}_k\| - (C_{\boldsymbol{x}_*,r_x} + 2^p(p+1)\sigma_*)\|\boldsymbol{x}_k - \boldsymbol{x}_*\|)\|\boldsymbol{y} - \boldsymbol{x}_k\|^p. \end{aligned}$$

Note that $\|\boldsymbol{y} - \boldsymbol{x}_k\| \geq \|\boldsymbol{y} - \boldsymbol{x}_*\| - \|\boldsymbol{x}_k - \boldsymbol{x}_*\| > r_y/2$. Assuming

$$\|\boldsymbol{x}_k - \boldsymbol{x}_*\| \leq \frac{\varepsilon r_y}{4(C_{\boldsymbol{x}_*,r_x} + 2^p(p+1)\sigma_*)} \text{ and } f(\boldsymbol{x}_k) \leq f(\boldsymbol{x}_*) + \frac{\varepsilon}{2}\Big(\frac{r_y}{2}\Big)^{p+1}$$

we can deduce that

$$m_k(\boldsymbol{y}) > m_{\boldsymbol{x}_*,\sigma_*}(\boldsymbol{y}) + \frac{\varepsilon}{2}\frac{r_y}{2}\|\boldsymbol{y} - \boldsymbol{x}_k\|^p \geq f(\boldsymbol{x}_*) + \frac{\varepsilon}{2}\Big(\frac{r_y}{2}\Big)^{p+1} \geq f(\boldsymbol{x}_k) = m_k(\boldsymbol{x}_k).$$

Therefore, if $\boldsymbol{x}_k$ is close enough to $\boldsymbol{x}_*$ then any point outside $B(\boldsymbol{x}_*, r_y)$ is not a global minimizer of the local model m_k. Since we assumed $\boldsymbol{y}_k$ is the global minimizer of m_k it must be contained in $B(\boldsymbol{x}_*, r_y)$ and the iteration is successful by Lemma 3.10. ∎

THEOREM 4.4. *Let $f \in C^p_{\text{Lip}}$, $\boldsymbol{x}_* \in M^q(f)$ and $p > q-1$. Assume that $\boldsymbol{x}_*$ is a global minimizer of f but not a global minimizer of $t^p_{\boldsymbol{x}_*}$, that $\gamma_1 = \gamma_2^{-\alpha}$ for some $\alpha \in \mathbb{Q}$ and that $\sigma_0 \neq \sigma_*$ for the value of σ_* from Theorem 4.3. For $\boldsymbol{x}_0$ close enough to $\boldsymbol{x}_*$ the values of σ_k will eventually cycle through finitely many values and the ratio of unsuccessful to successful iterations is α in each cycle, if $\boldsymbol{y}_k$ is the global minimizer of m_k at each iteration.*

Proof. By Theorem 4.3 there exists the value σ_* for the given f and global minimizer $\boldsymbol{x}_*$. Let $\alpha = \frac{a}{b}$ for some $a, b \in \mathbb{N}_{\geq 0}$ coprime if $\alpha > 0$ and $a = 0$ and $b = 1$ if $\alpha = 0$. Select

$$\varepsilon < \min\left\{ |\gamma_2^{c/b}\sigma_0 - \sigma_*| \;\middle|\; c \in \mathbb{Z} \right\}$$

and let r_ε be chosen such that the conclusions of Theorem 4.3 holds. By assuming $\boldsymbol{x}_*$ to be a global minimizer, a small adaptation of Theorem 3.7 implies that all iterates satisfy $\boldsymbol{x}_k \in B(\boldsymbol{x}_*, r_\varepsilon)$ if $\boldsymbol{x}_0$ is close enough to $\boldsymbol{x}_*$. Clearly, for any integer c we have $\gamma_2^{c/b}\sigma_0 \notin [\sigma_* - \varepsilon, \sigma_* + \varepsilon]$. By the mechanism by which σ_k is updated in Algorithm 2.1, we have that for all iterations $\sigma_k = \gamma_2^{c_k/b}\sigma_0$ for some integer $c_k \in \mathbb{Z}$. Let $c_* \in \mathbb{Z}$ be such that $\gamma_2^{c_*/b}\sigma_0$ is the smallest value larger than σ_*. Whenever $c_k \geq c_*$ the iteration is successful by Theorem 4.3 and $c_{k+1} = c_k - a$, and by the same logic whenever $c_k < c_*$ the iteration is unsuccessful and $c_{k+1} = c_k + b$.

If $a = 0$ then the regularization parameter will increase until $c_k \geq c_*$ and then stay constant, since from that point onwards all iterations are successful. The claim holds as σ_k eventually cycles through one value and the ratio of unsuccessful to successful iterations is $0 = \frac{a}{b}$.

Assume $a \neq 0$ from now on. Clearly, after each $c_k \geq c_*$ there will eventually be a $c_k < c_*$ and vice versa. Moreover, c_k will at some point enter the range $[c_* - a, c_* + b)$ and will never leave it again, so it suffices to show that once it does, the values cycle through finitely many values as claimed.

Assume for the moment that $c_0 = c_*$. Since there are only $a + b$ distinct values that c_k can take, the values must cycle. Consider the remainder of $c_k - c_*$ modulo b, then it does not change in unsuccessful iterations $c_k < c_*$, but it does change by $-a$ whenever $c_k \geq c_*$. As a and b are coprime, this implies every c_k between c_* and $c_* + b - 1$ is attained in the cycle. Similarly, modulo a the value of $c_k - c_*$ does not change in successful iterations, but does change by $+b$ whenever $c_k < c_*$ and so c_k will attain each value between $c_* - a$ and $c_* - 1$. Overall, the cycle starting at $c_0 = c_*$ includes $c_* - a, \ldots, c_* - 1$ and $c_*, \ldots, c_* + b - 1$ and so it has length $a + b$. As the cycle includes all values in $[c_* - a, c_* + b)$, even if $c_0 \neq c_*$ there will eventually be an iteration with c_k in the given range, which will then cycle through all $a + b$ values. During that cycle a iterations are unsuccessful and b iterations are successful, meaning that the ratio of unsuccessful to successful iterations is exactly $\alpha = \frac{a}{b}$ as claimed. ∎

5. Numerical Illustration. Since the impact of the abstract convergence orders derived in the previous sections may not become immediately clear, we provide illustrations for the example functions from Examples 2.3 and 4.1. The experiments are run using arbitrary precision arithmetic and are terminated when the distance to the true minimizer is less than 10^{-100}. The parameter values of Algorithm 2.1 are chosen as $\eta = \frac{1}{2}$, $\gamma_1 = \frac{1}{2}$, $\gamma_2 = 2$, $\theta = 0$, $\sigma_0 = \frac{1}{2}$, $x_0 = x_* + \frac{1}{10}$. To showcase convergence orders larger than one, the graphs in Figure 5.1 show $\log_{10}(\|x_k - x_*\|^{-1})$, an estimate of the number of correct digits, on a logarithmic scale.

For the non-degenerate case in Example 2.3, Newton's method converges quadratically. This means the number of correct digits roughly doubles at each iteration and

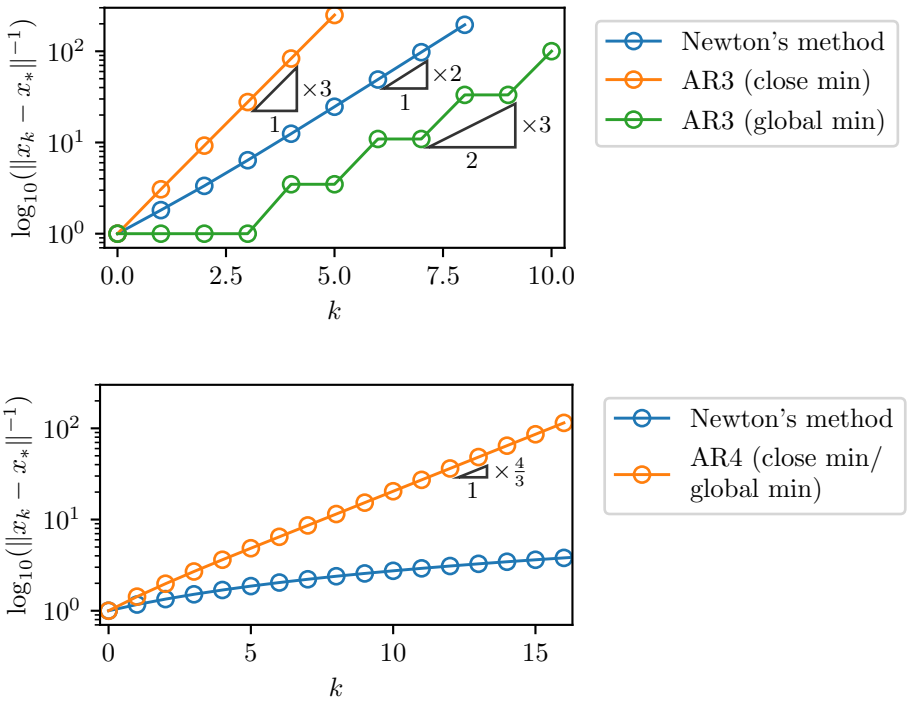

FIG. 5.1. *Convergence of Newton's method and ARp for Example 2.3 (top) and Example 4.1 (bottom, $p = 4$, $q = 4$). The slope of the triangles correspond to the theoretical order of convergence. Variants using the global model minimizer and the model minimizer close to the expansion point (see Lemma 3.10) are shown.*

thus, on a logarithmic scale, the corresponding line has a slope of $\log(2)$. The variant of AR3 using the successful model minimizer that is close to the expansion point, achieves cubic convergence, which corresponds to a line with slope $\log(3)$. When the global model minimizer is chosen at each step instead, the fourth iteration is the first successful one, since we have $\sigma_k > 2$ for the first time. Afterwards, the value of σ_k oscillates, and we can see that in effect only every second iteration is successful as predicted. With cubic progress during successful steps and no progress during unsuccessful steps, the overall slope is $\log(3)/2 = \log(\sqrt{3})$.

The second graph covers Example 4.1 with $p = 4$ and $q = 4$. On this degenerate example, Newton's method only achieves linear convergence, which corresponds to a line with decreasing slope. The fourth-order method, on the other hand, achieves superlinear convergence of order $\frac{p}{q-1}$, i.e., with slope $\log(\frac{4}{3})$.

6. Discussion and Conclusion. To round up the results in this paper, we compare them to those of Doikov and Nesterov [11] and Cartis, Gould and Toint [8, Section 5.3] and discuss connections to a characterization of local minimizers introduced by Cartis et al. [10].

As pointed out in the introduction, our paper can be seen as an extension to the analysis in [11] for objective functions without a non-smooth component. We replace a constant regularization parameter that requires knowledge of the Lipschitz constant with an adaptive one, relax global uniform convexity to local uniform convexity, and avoid the need for the exact global model minimizer by requiring only an approximate

local one. This relaxation of assumptions comes at the cost of slightly weaker results. Compare Lemmas 3.3 and 3.4 showing that the function value gap $f(\boldsymbol{x}_k) - f(\boldsymbol{x}_*)$ and gradient norm $\|\nabla f(\boldsymbol{x}_k)\|$ decrease at a $\frac{p}{q-1}$th-order rate during successful iterations to Theorems 1 and 2 in [11]. The conclusions are the same up to small differences in the constants and the fact that Lemma 3.4 only applies when the norm of the gradient is small enough, which Doikov and Nesterov can avoid by exploiting convexity of the local model. Similarly, they do not need to consider the possibility of multiple local minimizers or unsuccessful iterations as in Example 2.3.

Applying our results to their case, we take $\sigma_0 \geq \frac{pL_p}{(p+1)!}$, $\gamma_1 = 1$, $\theta = 0$ and $\eta = \frac{1}{2}$. By Lemma 2.2 every iteration is successful and so the regularization parameter stays constant and large enough. Moreover, the convexity of the model implies the sublevel set $\mathcal{L}_k$ from Lemma 3.11 is convex with only one connected component, implying that the global minimizer of the model lies in it and thus by Theorem 3.12 the same asymptotic Q-superlinear convergence of function values and gradients as in [11] is proved.

Section 5.3 in the book by Cartis, Gould and Toint [8] considers a wider class of functions, namely ones that are measure-dominated. The relevant class to compare with is the class of gradient-dominated functions of level $\mu = \frac{q}{q-1}$, which are a generalization of uniformly convex functions of order q that satisfy only (3.5), i.e.

$$f(\boldsymbol{x}) - f(\boldsymbol{x}_*) \leq \kappa \|\nabla f(\boldsymbol{x})\|^{\mu}$$

for some $\kappa > 0$. This inequality is also sometimes referred to as the Polyak–Łojasiewicz property. The algorithm analysed in [8][9] requires an approximate minimizer of the model in the sense of (2.1) at each iteration and updates the regularization parameter σ_k in a very similar way to Algorithm 2.1. There is some additional flexibility in [8] by introducing unsuccessful, successful and very successful iterations depending on the value of ρ_k and by allowing the increase and decrease factors to vary between iterations within given bounds. Simplifying these updates as done in Algorithm 2.1 does not fundamentally change the behaviour of the method, but was adopted here so that Theorems 3.6 and 4.4 are easier to state. A key change, however, is that the algorithm in [8] enforces a minimum regularization value $\sigma_{\min} > 0$. By using this restriction, it is possible to show a lower bound on the function decrease in terms of the gradient norm during successful iterations with an exponent of $\nu = \frac{p+1}{p}$:

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \geq \eta \sigma_{\min} R_{\max}^{-\frac{p+1}{p}} \|\nabla f(\boldsymbol{x}_{k+1})\|^{\frac{p+1}{p}}$$

Combining the gradient-domination of level μ and the function decrease result with exponent ν, Cartis, Gould and Toint are able to derive sublinear rates when $\nu > \mu$, linear rates when $\nu = \mu$ and superlinear rates for $\nu < \mu$. Specifically, for the $p > q - 1$ case considered in this paper we have $\nu < \mu$, and so they show that only $O(\log(\log(\varepsilon^{-1})))$ iterations are required to achieve a tolerance of $\varepsilon > 0$ for the function value gap $f(\boldsymbol{x}_k) - f(\boldsymbol{x}_*)$. The order of convergence for the function values is not stated explicitly in [8, Theorem 5.3.3], however from the proof we can see that a scaled function value gap ω_k satisfies $\omega_{k+1} \leq \omega_k^{\mu/\nu}$ for all k large enough if iteration k is successful [8, p. 465]. This corresponds to Q-convergence of order

$$\frac{\mu}{\nu} = \frac{pq}{(p+1)(q-1)} < \frac{p}{q-1}$$

[9] This is Algorithm 4.1.1 in [8] with $\varepsilon_2 = \varepsilon_3 = \infty$.

for the function values during successful iterations. The proof shows R-convergence of the same order for the gradient norms during successful iterations as well. The overall order of convergence is again not considered, because it does not change the log-log dependency on the tolerance ε. With the same setup as in Theorem 3.6 one would get R-convergence of order $\sqrt[1+\alpha]{\frac{pq}{(p+1)(q-1)}}$ for function values and gradient norms.

With their weaker assumptions on the function class and the addition of $\sigma_{\min}$, the authors are able to derive impressive results for the $p < q-1$ and $p = q-1$ cases, which go beyond the scope of this paper. For uniformly convex functions of order q and $p > q-1$, however, their rates are suboptimal and as shown in Theorem 3.5 the R-convergence of gradient norms can be strengthened to Q-convergence. Note that adding $\sigma_{\min}$ to Algorithm 2.1 does not change any of the results in section 3, and only requires $\sigma_{\min} < \gamma_1 \sigma_*$ in Theorem 4.4.

There are also connections of the notion of "right" and "wrong" minimizers in this paper to persistent and transient model minimizers introduced by Cartis et al. [10]. The authors of that study observe that the minimizers of regularized third-order Taylor expansions in general do not vary continuously with the regularization parameter σ, unlike the global minimizer of regularized first- and second-order polynomials. A local minimizer $\boldsymbol{y}_k$ of m_k is persistent if there exists a continuous curve $\boldsymbol{\psi}\colon [\sigma_k, \infty) \to \mathbb{R}^n$ with $\boldsymbol{\psi}(\sigma_k) = \boldsymbol{y}_k$ mapping each regularization parameter value to a local minimizer of the correspondingly regularized model, i.e., if one can continuously trace the local minimizer as σ grows to infinity while the Taylor polynomial stays unchanged. Any local minimizer that is not persistent is called transient. Cartis et al. [10] prove that any global minimizer of the AR1 and AR2 subproblem is persistent and that this property breaks for $p \geq 3$. Moreover, they develop an approximate way of detecting persistence and observe empirically that the performance of third-order methods can be improved by rejecting points that are deemed approximately transient. While the exact relationship between persistent and asymptotically successful minimizers remains an open question, we conjecture that for $\boldsymbol{x}_k$ close enough to $\boldsymbol{x}_*$ there always exists a persistent minimizer and that any persistent minimizer lies in the same connected component of the sublevel set of m_k as $\boldsymbol{x}_k$ and thus leads to a successful iteration according to Lemma 3.11. If correct, the persistent minimizer definition distinguishes useful from spurious model minimizers both in the local and non-local regime. This would encourage research into subproblem solvers aiming for (approximations of) persistent instead of global model minimizers.

Overall, in this paper we show that the results of Doikov and Nesterov [11] for convex functions hold more generally and that those of Cartis, Gould and Toint [8] can be strengthened for locally uniformly convex functions to $p/(q-1)$th-order convergence during successful iterations (Theorem 3.5), but not further (Example 4.1). Moreover, we clarify the importance of choosing the right local model minimizer for the asymptotic rates as shown in Example 2.3 as well as Theorems 3.6 and 3.12. Two different characterizations of "right" are provided (Lemmas 3.10 and 3.11) and an analysis that shows that oscillations of σ_k around some σ_* are fundamentally unavoidable for fully adaptive regularization methods and odd p if the "wrong" minimizer is chosen (Theorem 4.4). We hereby elucidate both the advantages and potential pitfalls of higher-order methods from the perspective of local convergence.

REFERENCES

[1] S. Bellavia and B. Morini, *Strong local convergence properties of adaptive regularized meth-*

ods for nonlinear least squares, IMA Journal of Numerical Analysis, 35 (2015), pp. 947–968, https://doi.org/10.1093/imanum/dru021.

[2] E. G. Birgin, J. L. Gardenghi, J. M. Martínez, S. A. Santos, and Ph. L. Toint, *Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models*, Mathematical Programming, 163 (2017), pp. 359–368, https://doi.org/10.1007/s10107-016-1065-8.

[3] C. Cartis, N. I. Gould, and P. L. Toint, *Universal Regularization Methods: Varying the Power, the Smoothness and the Accuracy*, SIAM Journal on Optimization, 29 (2019), pp. 595–615, https://doi.org/10.1137/16M1106316.

[4] C. Cartis, N. I. M. Gould, and P. L. Toint, *Adaptive cubic regularisation methods for unconstrained optimization. Part I: Motivation, convergence and numerical results*, Mathematical Programming, 127 (2011), pp. 245–295, https://doi.org/10.1007/s10107-009-0286-5.

[5] C. Cartis, N. I. M. Gould, and P. L. Toint, *Adaptive cubic regularisation methods for unconstrained optimization. Part II: Worst-case function- and derivative-evaluation complexity*, Mathematical Programming, 130 (2011), pp. 295–319, https://doi.org/10.1007/s10107-009-0337-y.

[6] C. Cartis, N. I. M. Gould, and P. L. Toint, *Second-Order Optimality and Beyond: Characterization and Evaluation Complexity in Convexly Constrained Nonlinear Optimization*, Foundations of Computational Mathematics, 18 (2018), pp. 1073–1107, https://doi.org/10.1007/s10208-017-9363-y.

[7] C. Cartis, N. I. M. Gould, and P. L. Toint, *Sharp Worst-Case Evaluation Complexity Bounds for Arbitrary-Order Nonconvex Optimization with Inexpensive Constraints*, SIAM Journal on Optimization, 30 (2020), pp. 513–541, https://doi.org/10.1137/17M1144854.

[8] C. Cartis, N. I. M. Gould, and P. L. Toint, *Evaluation Complexity of Algorithms for Nonconvex Optimization: Theory, Computation and Perspectives*, Society for Industrial and Applied Mathematics, Philadelphia, PA, Jan. 2022, https://doi.org/10.1137/1.9781611976991.

[9] C. Cartis, N. I. M. Gould, and Ph. L. Toint, *A concise second-order complexity analysis for unconstrained optimization using high-order regularized models*, Optimization Methods and Software, 35 (2020), pp. 243–256, https://doi.org/10.1080/10556788.2019.1678033.

[10] C. Cartis, R. Hauser, Y. Liu, K. Welzel, and W. Zhu, *Efficient Implementation of Third-order Tensor Methods with Adaptive Regularization for Unconstrained Optimization*, Feb. 2025, https://doi.org/10.48550/arXiv.2501.00404, https://arxiv.org/abs/2501.00404v2.

[11] N. Doikov and Y. Nesterov, *Local convergence of tensor methods*, Mathematical Programming, 193 (2022), pp. 315–336, https://doi.org/10.1007/s10107-020-01606-x.

[12] J. Fan and Y. Yuan, *A regularized Newton method for monotone nonlinear equations and its application*, Optimization Methods and Software, 29 (2014), pp. 102–119, https://doi.org/10.1080/10556788.2012.746344.

[13] A. Griewank, *The modification of Newton's method for unconstrained optimization by bounding cubic terms*, Technical Report NA/12, DAMTP, Cambridge University, 1981.

[14] B. Mendelson, *Introduction to Topology*, Allyn and Bacon, Inc., Boston, second edition ed., 1968, https://lccn.loc.gov/68021467.

[15] Y. Nesterov, *Lectures on Convex Optimization*, vol. 137 of Springer Optimization and Its Applications, Springer International Publishing, Cham, 2018, https://doi.org/10.1007/978-3-319-91578-4.

[16] Y. Nesterov and B. Polyak, *Cubic regularization of Newton method and its global performance*, Mathematical Programming, 108 (2006), pp. 177–205, https://doi.org/10.1007/s10107-006-0706-8.

[17] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer Series in Operations Research and Financial Engineering, Springer, New York, 2nd ed ed., 2006, https://doi.org/10.1007/978-0-387-40065-5.

[18] Q. Rebjock and N. Boumal, *Fast convergence to non-isolated minima: four equivalent conditions for C^2 functions*, Sept. 2024, https://doi.org/10.48550/arXiv.2303.00096, https://arxiv.org/abs/2303.00096.

[19] M.-C. Yue, Z. Zhou, and A. Man-Cho So, *On the Quadratic Convergence of the Cubic Regularization Method under a Local Error Bound Condition*, SIAM Journal on Optimization, 29 (2019), pp. 904–932, https://doi.org/10.1137/18M1167498.