Newton Method Revisited: Global Convergence Rates up to $\mathcal{O}\left(k^{-3}\right)$ for Stepsize Schedules and Linesearch Procedures

Slavomír Hanzely
MBZUAI*

Farshed Abdukhakimov
MBZUAI*

Martin Takáč
MBZUAI*

Abstract

This paper investigates the global convergence of stepsized Newton methods for convex functions with Hölder continuous Hessians or third derivatives. We propose several simple stepsize schedules with fast global convergence guarantees, up to $\mathcal{O}\left(k^{-3}\right)$. For cases with multiple plausible smoothness parameterizations or an unknown smoothness constant, we introduce a stepsize linesearch and a backtracking procedure with provable convergence as if the optimal smoothness parameters were known in advance. Additionally, we present strong convergence guarantees for the practically popular Newton method with exact linesearch.

1 Introduction

Second-order methods are fundamental to scientific computing. With its rich history that can be traced back to works Newton [1687], Raphson [1697], [Simpson, 1740], they have remained widely used up to the present day [Ypma, 1995, Conn et al., 2000]. The main advantage of second-order methods is their independence from the conditioning of the underlying problem, enabling an extremely fast local quadratic convergence rate, where precision doubles with each iteration. Additionally, they are inherently invariant to rescaling and coordinate transformations, which greatly simplifies parameter tuning. In contrast, the convergence of first-order methods is highly dependent on the problem's conditioning, resulting in a slower linear local convergence rate and a greater sensitivity to parameter tuning.

Despite their extremely fast local convergence, second-order methods often lack global convergence guarantees. Even the classical Newton method,

$$x^{k+1} = x^k - \left[\nabla^2 f(x^k)\right]^{-1} \nabla f(x^k), \tag{1}$$

can diverge when initialized far from the solution [Jarre and Toint, 2016, Mascarenhas, 2007]. Global convergence guarantees are typically achieved through various combinations of stepsize schedules [Nesterov and Nemirovski, 1994], line-search procedures [Kantorovich, 1948, Nocedal and Wright, 1999], trust-region methods [Conn et al., 2000], and Levenberg-Marquardt regularization [Levenberg, 1944, Marquardt, 1963].

The simplest globalization strategy is to employ stepsize schedules. These schedules can be based on implicit descent conditions, which often require an additional subroutine per iteration, such as exact linesearch [Cauchy, 1847, Shea and Schmidt, 2024], Armijo linesearch [Armijo, 1966], Wolfe condition [Wolfe, 1969], Goldstein condition [Nocedal and Wright, 1999]. However, those methods often lack global convergence guarantees achieved by simple stepsize schedules. Notably, Nesterov and Nemirovski [1994] introduced a simple stepsize schedule with global rate $\mathcal{O}\left(k^{-\frac{1}{2}}\right)$. Hanzely et al. [2022] improved upon this result by discovering duality between Newton stepsizes and Lavenberg-Marquardt regularization and proposing a stepsize with global rate $\mathcal{O}\left(k^{-2}\right)$ matching regularized Newton methods [Nesterov and Polyak, 2006, Mishchenko, 2023, Doikov and Nesterov, 2024].

Despite all recent advances, current guarantees still fall short of the optimal rate for functions with Hölder continuous Hessians, $\Omega\left(k^{-\frac{7}{2}}\right)$ [Gasnikov et al., 2019, Agarwal and Hazan, 2018, Arjevani et al., 2019]. It remains an open question whether the rate $\mathcal{O}\left(k^{-2}\right)$ achieved by Hanzely et al. [2022] is optimal for the Newton method or if more efficient stepsize schedules are yet to be discovered. In the context of first-order methods, several nontrivial stepsize schedules have been shown to improve convergence of Gradient Descent. Young [1953] introduced a stepsize schedule based on Chebyshev polynomials achieving the optimal rate for quadratic functions. Polyak [1987] proposed a stepsize schedule optimal for non-smooth convex functions, and Altschuler and Parrilo [2023], Grimmer et al. [2024] proposed stepsize schedules with guaranteed semi-accelerated rate for general convex, Lipschitz smooth functions. This motivates us to ask the question:

> *Is it possible to guarantee a global convergence rate better than $\mathcal{O}\left(k^{-2}\right)$ for a simple stepsize schedule of the Newton method?*

The answer is positive. We demonstrate that the stepsized Newton method can be analyzed under the

*Mohammed bin Zayed University of Artificial Inteligence.

assumption of Hölder continuity of third derivatives, achieving convergence guarantees resembling third-order tensor methods, up to $\mathcal{O}(k^{-3})$[1]. Analyzing the Newton method as the third-order method is a novel and unexpected approach, as the Newton method has traditionally been regarded as the most classical second-order method.

1.1 Benefits of basic methods

While it is possible to achieve optimal rates using acceleration techniques with a more complex structure [Gasnikov et al., 2019], basic methods are often preferred in practice for several reasons.

Firstly, basic methods are simple and easy to understand. They are also inherently robust, typically involving fewer hyperparameters, which minimizes the need for complex and costly hyperparameter tuning. In contrast, accelerated methods often require multiple sequences of iterates and additional hyperparameters, significantly increasing the complexity of tuning.

Moreover, basic methods can be seamlessly integrated with various techniques to enhance practical performance, such as parameter searches, data sampling strategies, momentum estimation, and gradient clipping. Combining these techniques with accelerated methods, however, introduces significant challenges. In the context of first-order methods, acceleration with parameter searches provides limited improvement over basic Gradient Descent with stepsize linesearch.

For second-order methods, the basic stepsized Newton method is particularly popular due to its affine invariance (i.e., invariance to changes in basis and data scaling), making it an efficient and convenient optimization tool.

1.2 Notation

For convex function $f : \mathbb{R}^d \to \mathbb{R}$, we consider the optimization objective

$$\min_{x\in\mathbb{R}^d} f(x), \tag{2}$$

where f is twice differentiable with nondegenerate Hessians and potentially ill-conditioned.

Our paper uses a nontrivial amount of notation; hence, we highlight definitions in gray and theorems in blue for easier reference. Denote any minimizer of the function $x^* \in \operatorname{argmin}_{x\in\mathbb{R}^d} f(x)$ and the optimal value $f_* \stackrel{\text{def}}{=} f(x^*)$. We define norms based on a symmetric positive definite matrix $\mathbf{H} \in \mathbb{R}^{d\times d}$. For all $x, g \in \mathbb{R}^d$,

$$\|x\|_{\mathbf{H}} \stackrel{\text{def}}{=} \langle \mathbf{H}x, x\rangle^{1/2}, \qquad \|g\|_{\mathbf{H}}^* \stackrel{\text{def}}{=} \left\langle g, \mathbf{H}^{-1}g\right\rangle^{1/2}.$$

As a special case $\mathbf{H} = \mathbf{I}$, we get l_2 norm $\|x\|_{\mathbf{I}} = \langle x, x\rangle^{1/2}$. We will utilize *local Hessian norm* $\mathbf{H} = \nabla^2 f(x)$, with shorthand notation for $h, g \in \mathbb{R}^d$

$$\|h\|_x \stackrel{\text{def}}{=} \left\langle \nabla^2 f(x)h, h\right\rangle^{1/2}, \|g\|_x^* \stackrel{\text{def}}{=} \left\langle g, \nabla^2 f(x)^{-1}g\right\rangle^{1/2}.$$

1.3 Stepsizes as a form of regularization

Hanzely et al. [2022] demonstrated that a stepsize schedule for the Newton method is equivalent to cubical regularization of the Newton method [Nesterov and Polyak, 2006] if the regularization is measured in the local Hessian norms. As the regularized Newton methods leverage the Taylor polynomial, we denote the second-order Taylor approximation of $f(y)$ by information at point x as

$$\Phi_x(y) \stackrel{\text{def}}{=} f(x) + \langle \nabla f(x), y-x\rangle + \frac{1}{2}\|y-x\|_x^2.$$

In particular, Hanzely et al. [2022] showed that

$$x^{k+1} = T(x^k),\ T(x) = \operatorname*{argmin}_{y\in\mathbb{R}^d} \left\{ \Phi_x(y) + \frac{\sigma}{3}\|y-x\|_x^3 \right\}$$

is equivalent to a Newton method with stepsize AICN[2]

$$x^{k+1} = x^k - \alpha_k [\nabla^2 f(x^k)]^{-1} \nabla f(x^k), \tag{3}$$

$$\text{for } \alpha_k = \frac{2}{1+\sqrt{1+2\sigma\|\nabla f(x^k)\|_{x^k}^*}}. \tag{4}$$

Note that stepsize schedule (4) preserves much larger stepsize when initialized far from the solution, $\left\|\nabla f(x^0)\right\|_{x^0}^* \gg 1$, compared to the stepsize of Damped Newton method [Nesterov and Nemirovski, 1994], which sets stepsize for L_{sc}-self-concordant functions as

$$\alpha_k = \frac{1}{1+L_{sc}\|\nabla f(x^k)\|_{x^k}^*}. \tag{5}$$

Aiming to extend this dependence beyond $L_{2,1}$-Hölder continuous functions (Definition 1), in Section 3 we present algorithm RN that under general $L_{p,\nu}$-Hölder continuity (Def 1) and $q = p+\nu \in [2,4]$ supports stepsize

$$\alpha_k = \frac{1}{1+(9L_{p,\nu})^{\frac{1}{q-1}}\|\nabla f(x^k)\|_{x^k}^{*\frac{q-2}{q-1}}}, \tag{6}$$

up to a constant recovering schedules of both AICN stepsize (4) (for $L_{2,1}$-Hölder continuous functions, $q=3$) and constant stepsizes of Karimireddy et al. [2018b], Gower et al. [2019a] (for $L_{2,0}$-Hölder continuous functions, $q=2$).

Remark. *Stepsized Newton methods often enjoy much simpler analysis compared to Newton methods regularized in l_2 norms, as it is possible to transition easily between gradients and model differences with an exact identity*

$$\left\|x^{k+1}-x^k\right\|_{x^k} \stackrel{(3)}{=} \alpha_k \left\|\nabla f(x^k)\right\|_{x^k}^*. \tag{7}$$

[1]For functions with Hölder continuous third derivatives, the achievable lower bound is $\Omega\left(k^{-5}\right)$ [Gasnikov et al., 2019].

[2]We present the stepsize in a simplified but equivalent form. Hanzely et al. [2022] expressed its stepsize as $\alpha_k = \frac{-1+\sqrt{1+2\sigma\left\|\nabla f(x^k)\right\|_{x^k}^*}}{\sigma\left\|\nabla f(x^k)\right\|_{x^k}^*}$.

1.4 Higher order of regularization

Extending cubic regularization [Nesterov and Polyak, 2006], tensor methods achieve better convergence guarantees by regularizing p-th order Taylor approximations by $(p+1)$-th order regularization (survey in Kamzolov et al. [2023]).

For third-order tensor methods, Nesterov [2021] showed that regularization can avoid computation of third-order derivatives, and Doikov et al. [2024] simplified regularization using technique of Mishchenko [2023] to

$$x^{k+1} = T(x^k), \text{ where for } \beta, \sigma \geq 0, \tag{8}$$

$$T(x) = \underset{y \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \Phi_x(y) + \frac{\sigma}{2} \|y - x\|_2^2 \|\nabla f(x)\|_2^{\beta} \right\}. \tag{9}$$

Combining insights about higher-order regularization with the regularization-stepsize duality of Hanzely et al. [2022] , we show that the higher-order regularization in local norms

$$x^{k+1} = T_{\sigma,\beta}\left(x^k\right), \text{ where for } \beta, \sigma \geq 0, \tag{10}$$

$$T_{\sigma,\beta}\left(x\right) = \underset{y \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \Phi_x(y) + \frac{\sigma}{2+\beta} \|y - x\|_x^{2+\beta} \right\}, \tag{11}$$

is equivalent to a Newton method with stepsize $\alpha_k \in (0, 1]$, where α_k is the *unique* positive root of the polynomial $P[\alpha] \overset{\text{def}}{=} 1 - \alpha - \alpha^{1+\beta}\sigma \left\|\nabla f(x^k)\right\|_{x^k}^{*\beta}$. Even though the polynomial P lacks an explicit formula for its roots, we derive algorithm RN (Algorithm 1) with a simple and exactly computed stepsize.

This method can be viewed as a third-order tensor method, as the model (11) bounds the third-order term of Taylor polynomial similarly to [Nesterov, 2021, Lemma 3].

Lemma 1. *Let function $f : \mathbb{R}^d \to \mathbb{R}$ be third-order $L_{3,\nu}$-Hölder continuous (Def. 1). Then $\forall x^k, x^{k+1} \in \mathbb{R}^d$,*

$$\left\|\nabla^3 f(x^k)[x^{k+1} - x^k]^2\right\|_{x^k}^{*} \leq 2 \left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{1}{1+\nu}} \left\|x^{k+1} - x^k\right\|_{x^k}^2.$$

Generality of higher-order regularization

Investigating generality of the regularization (11), w can observe that (11) also encapsulates all polynomial upper bounds of polynomials $P[\|x-y\|_x]$ with smaller exponents. Writing regularization as a polynomial,

$$f(y) \leq \Phi_x(y) + P[\|x - y\|_x], \tag{12}$$

this can be bounded as

$$f(y) \leq \Phi_x(y) + A_1 + A_2 \|x - y\|_x^p, \tag{13}$$

where constants $A_1, A_2 > 0$ and degree p are expressed in the lemma below. Notably, the next iterate x^+ set as the minimizer of the right-hand side of (13) is not affected by A_1, but the A_1 worsens guarantees on functional value decrease, $f(x^+) \leq f(x) + A_1$.

Lemma 2. *A polynomial P with d_P coefficients $a_k \geq 0$ and exponents $0 \leq b_1 \leq \cdots \leq b_{d_P}$,*

$$P[x] \overset{\text{def}}{=} \sum_{k=0}^{d_P} a_k x^{b_k},$$

satisfies following bound with any $p \geq \max_{k \in \{1,\ldots,d_P\}} b_k$,

$$P[x] \leq A_1 + A_2 x^p,$$

where $A_1 = \frac{1}{p}\sum_{k=0}^{d_P} a_k(p - b_k)$, $A_2 = \frac{1}{p}\sum_{k=0}^{d_P} a_k b_k$.

A surprising observation: Similarly, we can replace even the quadratic term from Taylor polynomial, $\frac{1}{2}\|y - x\|_x^2$, by an upper bound in the form $A_1 + A_2\|x - y\|_x^p$. This further simplifies the regularization and results in the Newton method with the **unbounded stepsize**

$$x^+ = x - \left(\frac{1}{(\sigma+1)\|\nabla f(x^k)\|_{x^k}^{*\beta}}\right)^{\frac{1}{1+\beta}} \left[\nabla^2 f(x)\right]^{-1} \nabla f(x).$$

As the gradient diminishes, the stepsize diverges to infinity. Yet, simultaneously, the functional value is guaranteed to not deteriorate by more than a constant factor. We refer the reader to the Appendix E for more details.

2 Contributions

Our contributions can be summarized as follows:

- **Newton method as a third-order tensor method:** We analyze the stepsized Newton method for functions with Hölder continuous third-derivatives (Definition 1). This reframes the classical second-order Newton method as a **third-order** method, bridging the gap between second-order methods and third-order tensor methods.
- **Simple stepsizes for fast global convergence:** We propose multiple stepsize schedules for the Newton method (RN, Alg 1), leveraging **various** Hölder continuity assumptions (Def 1). Although the stepsize is chosen to be a root of a non-quadratic polynomial, it is surprisingly **simple and directly computable**.

  Depending on the considered variant of the Hölder continuity assumption, they can achieve a global convergence rate up to $\mathcal{O}\left(k^{-3}\right)$ (Theorem 2). These are the first Newton method stepsizes improving upon the rate $\mathcal{O}\left(k^{-2}\right)$of Hanzely et al. [2022].

  Additionally, we establish the following guarantees:
  - a **local superlinear** convergence rate (Theorem 3),
  - a **global linear** convergence (Theorems 9, 10) under additional assumption of finite *s-relative size* (Definition 4) [Doikov et al., 2024],
  - and a **global superlinear** convergence (Theorem 7) under the additional assumption of uniform star-convexity (Definition 3) of degree $s \geq 2$.

- **Stepsize linesearches for unknown parameters:**
  In practice, smoothness constants are often unknown, requiring approximation or fine-tuning. To address this, we introduce a **linesearch** procedure GRLS (24) and a **stepsize backtracking** method UN (Algorithm 2), both of which provably converge as if the **optimal** parameterization was known in advance (Col 1, Th 5).
- **Guarantees for popular Newton linesearch:**
  As a byproduct of our analysis, we prove similar convergence guarantees for the popular Newton method with greedy linesearch (27) (Col 2, Th 7). This is, to our best knowledge, the first result of such kind.
- **Experimental comparison:**
  In Section 8, we experimentally compare the proposed algorithms (RN, UN, and GRLS) with existing methods and demonstrate that they outperform their counterparts in most of the considered scenarios.

  Also, we observe that the stepsizes of linesearch procedure GRLS closely resemble stepsizes of popular Greedy Newton linesearch.

2.1 Most relevant literature

Our theoretical framework leverages multiple insights of works Hanzely et al. [2022] and Doikov et al. [2024]. We will outline the key differences between those approaches.

Compared to our approach, the AICN method of Hanzely et al. [2022] is restricted to cubic regularization and achieves only an $\mathcal{O}\left(k^{-2}\right)$ convergence rate. In contrast, our schedules incorporate a range of smoothness notions, including the Hölder continuity of the third derivative, allowing Algorithm 1 to achieve rates up to $\mathcal{O}\left(k^{-3}\right)$. Additionally, while AICN requires prior knowledge of the smoothness constant, our backtracking linesearch Algorithm 2 provably converge as if the optimal parameterization was known in advance.

Furthermore, while Hanzely et al. [2022] relies on cubic regularization, resulting in a stepsize that is the root of a quadratic polynomial, higher-order regularizations yield a stepsize that is the root of a higher-order polynomial. Surprisingly, we show that even with higher-order regularization, there is a unique positive root in the interval $(0, 1]$, and we present algorithms (Algorithm 1 and Algorithm 2) that can operate without requiring any additional linesearch.

In comparison to Doikov et al. [2024], which utilizes standard l_2 norms for regularization, our approach leverages the local Hessian norms suggested by [Hanzely et al., 2022]. By utilizing local norms, the minimizers of various regularization models (11) align on the same line, naturally connecting different regularizations from a geometric perspective. Local norms also result in a simpler algorithm invariant to linear transformations (e.g., data scaling or choice of basis), which is a valuable property in practice, as it significantly reduces hyperparameter tuning.

We would like to highlight that our results explain the success of popular stepsize linesearches in the Newton direction. These insights have implications far beyond our newly proposed methods. In comparison, the results presented in Doikov et al. [2024] do not provide a novel theoretical explanation for any established method.

3 Simple stepsize schedule

Now we are ready to present our new stepsize schedule.

Theorem 1. *For any constants $\sigma, \beta \geq 0$, the following modifications of the Newton method are equivalent:*

$$\textit{Regularize: } x^{k+1} = x^k + \operatorname*{argmin}_{y \in \mathbb{R}^d} T_{\sigma,\beta}\left(x^k\right), \tag{14}$$

$$\textit{Damping: } x^{k+1} = x^k - \alpha_k [\nabla^2 f(x^k)]^{-1} \nabla f(x^k), \tag{15}$$

where,

$$T_{\sigma,\beta}(x) = \operatorname*{argmin}_{y \in \mathbb{R}^d} \left\{ \Phi_x(y) + \frac{\sigma}{2+\beta} \|y - x\|_x^{2+\beta} \right\},$$

and $\alpha_k \in (0,1]$ is the only positive root of polynomial $P[\alpha] \overset{\text{def}}{=} 1 - \alpha - \alpha^{1+\beta}\sigma \left\|\nabla f(x^k)\right\|_{x^k}^{*\beta}$.

We call this algorithm Root Newton (RN), Algorithm 1.

To simplify calculations, we reparametrize the RN as $\theta \overset{\text{def}}{=} \alpha^\beta \sigma \|\nabla f(x)\|_x^{*\beta}$, and $\theta \geq 0$. Now, the polynomial P simplifies to $P[\alpha] = 1 - \alpha - \alpha\theta$ and for fixed θ, the positive root of P can be expressed as $\alpha = \frac{1}{1+\theta}$, with $\alpha\theta < 1$.

3.1 Hölder continuity

Our analysis is built upon the assumption that the function has Hölder continuous Hessian or third derivative.

Definition 1. *For $f: \mathbb{R}^d \to \mathbb{R}$, and $p \in \mathbb{N}$, we say that p-times differentiable convex function is* Hölder continuous *of p-th order, if for some $\nu \in [0,1]$ there exists a constant $L_{p,\nu} < \infty$, such thata $\forall x, y \in \mathbb{R}^d$,*

$$\|\nabla^p f(x) - \nabla^p f(y)\|_{op} \leq L_{p,\nu} \|x - y\|_x^\nu, \tag{16}$$

We say that the f has Hölder continuous Hessian if $L_{2,\nu} < \infty$ (for some $\nu \in [0,1]$) and Hölder continuous third derivative if $L_{3,\nu} < \infty$ (for some $\nu \in [0,1]$).

In particular, $L_{3,0} = \left\|\nabla^3 f(x) - \nabla^3 f(y)\right\|_{op}$ and $L_{2,1} = \sup_x \left\|\nabla^3 f(x)\right\|_{op}$ matches the definition of semi-strong

Table 1: Global convergence guarantees of stepsized Newton methods under various notions of Hölder continuity (Definition 1). For simplicity, we report dependence only on the number of iterations k.

Stepsize schedule	**Stepsize for** $g_x \overset{\text{def}}{=} \|\nabla f(x)\|_x^*$	**Smoothness assumption**	**Global rate**	**Reference**
Damped Newton B	$\frac{1}{1+L_{sc}g_x}$ (0)	L_{sc} (0)	$\mathcal{O}\left(k^{-\frac{1}{2}}\right)$ (1)	[Nesterov and Nemirovski, 1994] (1)
AICN	$\frac{2}{1+\sqrt{1+2L_{2,1}g_x}}$ (2)	$L_{2,1}$	$\mathcal{O}\left(k^{-2}\right)$	[Hanzely et al., 2022]
RN (Algorithm 1)	$\frac{1}{1+(9L_{p,\nu})^{\frac{1}{q-1}} g_x^{\frac{q-2}{q-1}}}$ (3)	$L_{p,\nu}$ (3)	$\mathcal{O}\left(k^{-(p+\nu-1)}\right)$ (3)	**This work** (Theorem 4)
GRLS (24)	Linesearched	$L_{p,\nu}$ (3) (unknown)	$\min_{p,\nu} \mathcal{O}\left(k^{-(p+\nu-1)}\right)$ (3)	**This work** (Corollary 1)
UN (Algorithm 2)	Backtracked	$L_{p,\nu}$ (3) (unknown)	$\min_{p,\nu} \mathcal{O}\left(k^{-(p+\nu-1)}\right)$ (3)	**This work** (Theorem 5)
Greedy Newton (27)	Linesearched	$L_{p,\nu}$ (3) (unknown)	$\min_{p,\nu} \mathcal{O}\left(k^{-(p+\nu-1)}\right)$ (3)	Folklore Rate: Corollary 2 (**new**)

(0) Constant L_{sc} represents self-concordance constant and is implied by $L_{2,1}$-Hölder continuity.
(1) Authors show global decrease $f(x^{k+1}) \leq f(x^k) - c$ for some $c > 0$. Rate $\mathcal{O}(k^{-\frac{1}{2}})$ is reported in Hanzely et al. [2022]. We were unable to prove or find the convergence guarantee for Damped Newton B of the form $\mathcal{O}(k^{-\alpha})$.
(2) We present a simplified form of the stepsize. Authors proposed AICN stepsize in equivalent form $\frac{-1+\sqrt{1+L_{2,1}g_x}}{L_{2,1}g_x}$.
(3) Parameters p, ν are fixed and satisfy $p \in \{2,3\}, \nu \in [0,1]$ and $p+\nu-1 \in [1,3]$.

self-concordance [Hanzely et al., 2022]. Function $L_{p,\nu}$ is log-convex in ν and hence for $0 \leq \nu_1 \leq \nu \leq \nu_2 \leq 1$, hold

$$L_{p,\nu} \leq [L_{p,\nu_1}]^{\frac{\nu_2-\nu}{\nu_2-\nu_1}} [L_{p,\nu_2}]^{\frac{\nu-\nu_1}{\nu_2-\nu_1}},$$
$$L_{p,\nu} \leq L_{p,0}^{1-\nu} L_{p,1}^{\nu}.$$

We will use the properties of the Hölder continuity summarized in the proposition below.

Proposition 1. *$L_{2,\nu}$-Hölder continuous functions satisfy*

$$\left\|\nabla f(y) - \nabla f(x) - \nabla^2 f(x)[y-x]\right\|_x^* \leq \frac{L_{2,\nu}}{1+\nu}\|y-x\|_x^{1+\nu}.$$

$L_{3,\nu}$-Hölder continuous functions satisfy

$$\Big\|\nabla f(y) - \nabla f(x) - \nabla^2 f(x)[y-x] - \frac{1}{2}\nabla^3 f(x)[y-x]^2\Big\|_x^* \leq \frac{L_{3,\nu}}{(1+\nu)(2+\nu)}\|y-x\|_x^{2+\nu}.$$

For further discussion of smoothness constants, we refer the reader to Appendix D.

3.2 One-step decrease Hölder continuity

We are going to show that from the Hölder continuity for sufficiently large θ_k follows bound

$$\left\langle \nabla f(x^{k+1}), \left[\nabla^2 f(x^k)\right]^{-1} \nabla f(x^k) \right\rangle \geq \frac{1}{2c_1(1-\alpha_k)} \left\|\nabla f(x^{k+1})\right\|_{x^k}^{*2},$$

for $c_1 \in \{1,2\}$, implying the one-step decrease as

$$\begin{aligned} f(x^k) - f(x^{k+1}) &\geq -\left\langle \nabla f(x^{k+1}), x^{k+1} - x^k \right\rangle \\ &= \left\langle \nabla f(x^{k+1}), \alpha_k \left[\nabla^2 f(x^k)\right]^{-1} \nabla f(x^k) \right\rangle \\ &\geq \frac{\alpha_k}{2c_1(1-\alpha_k)} \left\|\nabla f(x^{k+1})\right\|_{x^k}^{*2} \\ &= \frac{1}{2c_1\theta_k} \left\|\nabla f(x^{k+1})\right\|_{x^k}^{*2}. \end{aligned} \tag{17}$$

Lemma 3. *Let $\left\|\nabla f(x^k)\right\|_{x^k}^* > 0$, and $x^k \in \mathbb{R}^d, x^{k+1} = x^k - \alpha_k \left[\nabla^2 f(x^k)\right]^{-1} \nabla f(x^k)$, as* RN. *Hölder continuity of **Hessian** (Definition 1 with $p = 2$) implies that for θ_k larger than*

$$\theta_k \geq \frac{L_{2,\nu}}{1+\nu} \alpha_k^{\nu} \left\|\nabla f(x^k)\right\|_{x^k}^{*\nu}, \tag{18}$$

holds

$$\left\langle \nabla f(x^{k+1}), \left[\nabla^2 f(x^k)\right]^{-1} \nabla f(x^k) \right\rangle \geq \frac{1}{2(1-\alpha_k)} \left\|\nabla f(x^{k+1})\right\|_{x^k}^{*2}.$$

Lemma 4. *Let* $\left\|\nabla f(x^k)\right\|_{x^k}^* > 0$*, and* $x^k \in \mathbb{R}^d, x^{k+1} = x^k - \alpha_k \left[\nabla^2 f(x^k)\right]^{-1} \nabla f(x^k)$*, as* RN*. Hölder continuity of the* ***third derivative*** *(Definition 1 with* $p = 3$*) implies that for* θ_k *larger than*

$$\theta_k \geq \alpha_k \left\|\nabla f(x^k)\right\|_{x^k}^* \max\left\{6\left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{1}{1+\nu}}, \frac{\sqrt{3}L_{3,\nu}}{(1+\nu)(2+\nu)}\left(\alpha_k \left\|\nabla f(x^k)\right\|_{x^k}^*\right)^{\nu}\right\}, \quad (19)$$

holds

$$\left\langle \nabla f(x^{k+1}), \left[\nabla^2 f(x^k)\right]^{-1} \nabla f(x^k) \right\rangle \geq \frac{1}{4(1-\alpha_k)} \left\|\nabla f(x^{k+1})\right\|_{x^k}^{*2}.$$

3.3 Generalized one-step decrease

In Lemma 3 and Lemma 4, the requirement on θ_k is dependent on α_k. We can use the following observation to derive a bound dependent only on the norm of the gradient.

Lemma 5. *For* $c_3, \delta > 0$*, choice* $\theta_k \geq c_3^{\frac{1}{1+\delta}} \left\|\nabla f(x^k)\right\|_{x^k}^{*\frac{\delta}{1+\delta}}$ *ensures* $\theta_k \geq c_3 \left(\alpha_k \left\|\nabla f(x^k)\right\|_{x^k}^*\right)^{\delta}$.

With Lemma 5, we can unify the cases $p \in \{2, 3\}$ (see Corollary 3 for the additional explanation). Let us reparametrize as $q \stackrel{\text{def}}{=} p + \nu \in [2, 4]$, $M_q \stackrel{\text{def}}{=} L_{p,\nu}$.

Theorem 2. *Let* $\|\nabla f(x)\|_x^* > 0$*. Hölder continuity (Definition 1) with* $q = p + \nu \in [2, 4]$ *for points* $x^k, x^{k+1} = x^k - \alpha_k \left[\nabla^2 f(x^k)\right]^{-1} \nabla f(x^k)$ *from* RN *implies that for* θ_k *such that*

$$\theta_k \geq (9M_q)^{\frac{1}{q-1}} \left\|\nabla f(x^k)\right\|_{x^k}^{*\frac{q-2}{q-1}} \quad (20)$$

holds

$$\left\langle \nabla f(x^{k+1}), \left[\nabla^2 f(x^k)\right]^{-1} \nabla f(x^k) \right\rangle \geq \frac{1}{2\alpha_k \theta_k} \left\|\nabla f(x^{k+1})\right\|_{x^k}^{*2}. \quad (21)$$

In particular, in view of (17)*, we have that the choice* $\theta_k = (9M_q)^{\frac{1}{q-1}} \left\|\nabla f(x^k)\right\|_{x^k}^{*\frac{q-2}{q-1}}$ *guarantees decrease*

$$f(x^k) - f(x^{k+1}) \geq \frac{1}{2}\left(\frac{1}{9M_q}\right)^{\frac{1}{q-1}} \frac{\left\|\nabla f(x^{k+1})\right\|_{x^k}^{*2}}{\left\|\nabla f(x^k)\right\|_{x^k}^{*\frac{q-2}{q-1}}}. \quad (22)$$

This naturally leads to an optimization algorithm RN.

Algorithm 1 RN: Root Newton stepsize schedule

Requires: Initial point $x^0 \in \mathbb{R}^d$, Hölder continuity exponent $q \in [2, 4]$ and constant $M_q < \infty$.
for $k = 0, 1, 2 \dots$ **do**
  $n^k = \left[\nabla^2 f(x^k)\right]^{-1} \nabla f(x^k)$ ▷ Newton direction
  $g_k = \left\langle \nabla f(x^k), n^k \right\rangle^{\frac{1}{2}}$ ▷ $g_k = \left\|\nabla f(x^k)\right\|_{x^k}^*$
  $\theta_k = (9M_q)^{\frac{1}{q-1}} g_k^{\frac{q-2}{q-1}}$ ▷ Sufficient regularization
  $\alpha_k = \frac{1}{1+\theta_k}$ ▷ α_k is the root of $P[\alpha]$
  $x^{k+1} = x^k - \alpha_k n^k$ ▷ Step, $x^k = T_{\sigma_k,\beta}\left(x^k\right)$
end for

4 Convergence garantees of RN

Denote the functional suboptimality $f_k \stackrel{\text{def}}{=} f(x^k) - f_*$, the initial level set $\mathcal{Q}(x^0) \stackrel{\text{def}}{=} \left\{x \in \mathbb{R}^d : f(x) \leq f(x^0)\right\}$, and its diameter as $D \stackrel{\text{def}}{=} \sup_{x,y \in \mathcal{Q}(x^0)} \|x - y\|_x$. Note that convexity and bounded diameter of $\mathcal{Q}(x^0)$, $D < \infty$ together imply $D\left\|\nabla f(x^k)\right\|_{x^k}^* \geq f_k$. We need the Hessian not to change much between iterations to guarantee the global convergence rate.

Assumption 1. *There exists a* $\gamma > 0$ *bounding norms of the gradients in the consecutive iterates,*

$$\gamma \leq \frac{\left\|\nabla f(x^{k+1})\right\|_{x^k}^{*2}}{\left\|\nabla f(x^{k+1})\right\|_{x^{k+1}}^{*2}}.$$

Required γ exists in many cases. For L-smooth μ-strongly convex functions, $\gamma = \frac{\mu}{L}$. For functions with $\hat{c}$-stable Hessian [Karimireddy et al., 2018a], $\gamma = \hat{c}$. For L_{sc}-self-concordant functions, it holds when the points x, x^+ are close to each other [Nesterov and Nemirovski, 1994] or in the neighborhood of the solution (Proposition 2).

Proposition 2 (Hanzely et al. [2022], Lemma 4)**.** *For convex* L_{sc}*-self-concordant function* $f : \mathbb{R}^d \to \mathbb{R}$ *and for any* $0 < c_4 < 1$ *in the neighborhood of solution* $x^k \in \left\{x : \|\nabla f(x)\|_x^* \leq \frac{(2c_4+1)^2-1}{2L_{sc}}\right\}$ *holds*

$$\nabla^2 f(x^{k+1})^{-1} \preceq (1 - c_4)^{-2} \nabla^2 f(x^k)^{-1}.$$

First, we present the local convergence of the RN.

Theorem 3. *Let function $f : \mathbb{R}^d \to \mathbb{R}$ be convex, $L_{p,\nu}$-Hölder continuous ($q = p + \nu$) with γ-bounded Hessian change* (1)*. Algorithm* RN *has a superlinear local convergence rate,*

$$\left\|\nabla f(x^{k+1})\right\|^{*}_{x^{k+1}} \leq \frac{2}{\gamma}\left(9M_q\right)^{\frac{1}{q-1}} \left\|\nabla f(x^k)\right\|^{*\left(2-\frac{1}{q-1}\right)}_{x^k}.$$

Now we quantify the global convergence rate following from Theorem 2 and present the rate of RN.

Lemma 6. *Let function $f : \mathbb{R}^d \to \mathbb{R}$ be convex with γ-bounded Hessian change* (1) *and the bound level sets with diameter D. If an algorithm $\mathcal{A}$ generates the iterates $\left\{x^k\right\}_{k=1}^{n}$ with one-step decrease for $q \geq 2$ and $c_5 \geq 0$ as*

$$f(x^k) - f(x^{k+1}) \geq c_5 \frac{\left\|\nabla f(x^{k+1})\right\|^{*2}_{x^k}}{\left\|\nabla f(x^k)\right\|^{*\frac{q-2}{q-1}}_{x^k}}, \tag{23}$$

then $\mathcal{A}$ has the global convergence rate

$$f_n \leq \frac{D^q\left(2\gamma(q-1)\right)^{q-1}}{{c_5}^{q-1}n^{q-1}} + \left\|\nabla f(x^0)\right\|^{*}_{x^0} D \exp\left(-\frac{k}{4}\right).$$

Theorem 4. *Let function $f : \mathbb{R}^d \to \mathbb{R}$ be convex, $L_{p,\nu}$-Hölder continuous ($q = p + \nu$) with γ-bounded Hessian change* (1) *and the bound level sets with diameter D.* RN *(Algorithm 1) with known parameters q, M_q converges as*

$$f(x^k) - f_* \leq \frac{9M_qD^q\left(4\gamma(q-1)\right)^{q-1}}{k^{q-1}} + \left\|\nabla f(x^0)\right\|^{*}_{x^0} D \exp\left(-\frac{k}{4}\right),$$

which in $\mathcal{O}$ notation is simplifies to $\mathcal{O}\left(\frac{M_qD^q}{k^{q-1}}\right)$.

Note that the loss function can satisfy Hölder continuity (Definition 1) with multiple different $L_{p,\nu}$, and therefore different pairs (q, M_q) can be used. The best parametrization might not be known.

5 Unknown parametrization

To address unknown parameterization, we propose a stepsize linesearch Gradient-Regulated Line Search GRLS simultaneously minimizing loss and gradient norms as

$$x^{k+1} = \operatorname*{argmin}_{y \in \left\{x - \alpha n_{x^k} \mid \alpha \in [0,1]\right\}} \frac{f(y) - f(x^k)}{\left\|\nabla f(y)\right\|^{*2}_{x^k}}, \tag{24}$$

where $n_x \overset{\text{def}}{=} [\nabla^2 f(x)]^{-1}\nabla f(x)$ is a shorthand for Newton's direction at point x. Linesearch GRLS is directly minimizing bound (23) in Lemma 6, and therefore has the corresponding convergence rate.

Corollary 1. *Let function $f : \mathbb{R}^d \to \mathbb{R}$, be convex, Hölder continuous with some $M_q < \infty$, with γ-bounded Hessian change* (1)*, and the bound level sets with diameter $D < \infty$. Linesearch* GRLS *converges as $\min_{q\in[2,4]} \mathcal{O}\left(\frac{M_qD^q}{k^{q-1}}\right)$*

$$f(x^k) - f_* \leq \min_{q\in[2,4]} \frac{9M_qD^q\left(4\gamma(q-1)\right)^{q-1}}{k^{q-1}} + \left\|\nabla f(x^0)\right\|^{*}_{x^0} D \exp\left(-\frac{k}{4}\right). \tag{25}$$

Observe that for small stepsizes $\alpha_k \in [0, \overline{\alpha}]$, for some $\overline{\alpha} \ll 1$, model differences are small $x^{k+1} \approx x^k$ and $\nabla f(x^k) \approx \nabla f(x^{k+1})$. Therefore, expression (24) minimized by GRLS can be approximated as

$$\frac{f(y) - f(x^k)}{\left\|\nabla f(y)\right\|^{*2}_{x^k}} \approx \frac{f(y) - f(x^k)}{\left\|\nabla f(x^k)\right\|^{*2}_{x^k}}, \tag{26}$$

and the minimizer of the right-hand-side is equivalent to the practically popular Newton method with greedy linesearch

$$x^{k+1} = \operatorname*{argmin}_{y \in \left\{x^k - \alpha n_{x^k} \mid \alpha \in [0,1]\right\}} f(y), \tag{27}$$

which we will call *Greedy Newton* (GN). Leveraging this insight, we obtain the convergence rate for (27) in the corollary below. More details can be found in Appendix B.

Corollary 2. *Let function $f : \mathbb{R}^d \to \mathbb{R}$, be convex, M_q-Hölder continuous for some $M_q < \infty$, with γ-bounded Hessian change* (1)*, and the bound level sets with diameter $D < \infty$. If the Greedy Newton linesearch* (27) *satisfies the inequality $\left\|\nabla f(x^{k+1})\right\|^{*}_{x^k} \leq \overline{c}\left\|\nabla f(x^k)\right\|^{*}_{x^k}$ with some constant $\overline{c} \geq 0$ for all iterates x^k, then it has convergence guarantee $\min_{q\in[2,4]} \mathcal{O}\left(\frac{M_qD^q\overline{c}^{2(q-1)}}{k^{q-1}}\right)$*

$$f(x^k) - f_* \leq \min_{q\in[2,4]} \frac{9M_qD^q\left(4\gamma\overline{c}^2(q-1)\right)^{q-1}}{k^{q-1}} + \left\|\nabla f(x^0)\right\|^{*}_{x^0} D \exp\left(-\frac{k}{4}\right).$$

Linesearches GRLS (24) and GN (27) have fast convergence guarantees without knowledge of smoothness parametrization (q, M_q), yet their implicit nature might not be suitable for all practical scenarios. To remedy that, in the next section, we present a stepsize backtracking procedure with matching convergence guarantees.

6 Universal stepsize backtracking

Our backtracking procedure is based on the observation that the knowledge of the parametrization (q, M_q) in RN (Algorithm 1) is required only for setting θ_k. We start with an estimate of θ_k smaller than the true value and increase it until it achieves the theoretically predicted decrease. We

Algorithm 2 UN: Universal stepsize backtracking procedure for the Newton method

Input: Initial point $x^0 \in \mathbb{R}^d$, any constants $\beta \in \left[\frac{2}{3}, 1\right], \sigma_0, \gamma > 1$ $\quad \triangleright\ \beta \geq \frac{q-2}{q-1}$ for $q \in [2,4]$
for $k = 0, 1, 2 \ldots$ **do**
  $n^k = \left[\nabla^2 f(x^k)\right]^{-1} \nabla f(x^k)$ $\quad \triangleright$ Newton direction
  $g_k = \langle \nabla f(x^k), n^k \rangle^{\frac{1}{2}}$ $\quad \triangleright = \left\|\nabla f(x^k)\right\|_{x^k}^*$
  for $j_k = 0, 1, 2 \ldots$ **do**
    $\theta_{k,j_k} = \gamma^{j_k} \sigma_k g_k^\beta$ $\quad \triangleright$ Increase regularization
    $\alpha_{k,j_k} = \frac{1}{1+\theta_{k,j_k}}$ $\quad \triangleright$ Update stepsize
    $x_{j_k}^k = x^k - \alpha_{k,j_k} n^k$ $\quad \triangleright = T_{\gamma^{j_k}\sigma_k,\beta_k}(x^k)$
    if $\langle \nabla f(x_{j_k}^k), n^k \rangle \geq \frac{1}{2\alpha_{k,j_k}\theta_{k,j_k}} \left\|\nabla f(x_{j_k}^k)\right\|_{x^k}^{*2}$ **then**
      $x^{k+1} = x_{j_k}^k$
      $\sigma_{k+1} = \gamma^{j_k - 1} \sigma_k$
      break
    end if
  end for
end for

claim that the resulting algorithm, UN, Algorithm 2, is well-defined with a bounded number of backtracking steps and a fast global convergence rate.

To formalize this claim, we first define a quantity to identify the smallest plausible true parameter θ_k to be estimated first, $\mathcal{H}(x) \stackrel{\text{def}}{=} \inf_{q\in[2,4]} (9M_q)^{\frac{1}{q-1}} \|\nabla f(x)\|_x^{*\left(\frac{q-2}{q-1}-\beta\right)}$, for $q \in [2,4]$ and $\beta \geq \frac{2}{3}$.

Lemma 7. *If $M_q < \infty$ for some $q \in [2,4]$, and the initial estimate σ_0 small enough, $\sigma_0 \leq \mathcal{H}(x^0)$, then all iterations $\{x^k\}_{k=0}^n$ of* UN*, such that $\left\|\nabla f(x^k)\right\|_{x^k}^* > 0$, satisfy $\sigma_{k+1} = \frac{\theta_{k,j_k-1}}{\|\nabla f(x^k)\|_{x^k}^{*\beta}} \leq \mathcal{H}(x^k)$. Moreover, the total number N_K of backtracking steps during the first k iterations is bounded,*

$$N_k \leq 2k + \log_c \frac{\mathcal{H}(x^{k-1})}{\sigma_0}.$$

Theorem 5. *Let function $f : \mathbb{R}^d \to \mathbb{R}$, be convex, Hölder continuous with some $M_q < \infty$, with γ-bounded Hessian change (Assumption 1), and the bound level sets with diameter $D < \infty$.* UN *(Algorithm 2) converges with the rate $\min_{q\in[2,4]} \mathcal{O}\left(\frac{M_q D^q}{k^{q-1}}\right)$,*

$$f(x^k) - f_* \leq \min_{q\in[2,4]} \frac{9M_q D^q \left(4\gamma^2(q-1)\right)^{q-1}}{k^{q-1}} + \left\|\nabla f(x^0)\right\|_{x^0}^* D \exp\left(-k/4\right).$$

7 Global (super)linear convergence rate

Stepsized Newton method is known to be able to achieve a global linear rate if the Hessian is bounded and stepsize is constant [Karimireddy et al., 2018b, Gower et al., 2019b], or when the function is $L_{2,1}$-Hölder continuous with stepsize following schedule AICN [Hanzely et al., 2022, proof in [Hanzely, 2023]].

In line with those results, we present global linear rates for algorithms RN, UN, GRLS on $L_{p,\nu}$-Hölder continuous functions with finite $(p+\nu)$-relative size characteristic [Doikov et al., 2024]. The proof is in Appendix G.

Definition 2 ([Doikov et al., 2024])**.** *For strictly convex function $f : \mathbb{R}^d \to \mathbb{R}$ we call s-*relative size *characteristic*

$$D_s \stackrel{\text{def}}{=} \sup_{x,y\in\mathcal{Q}(x^0)} \left\{ \|x-y\|_x \left(\frac{V_f}{\beta_f(x,y)}\right)^{\frac{1}{s}} \right\},$$

where $\beta_f(x,y) \stackrel{\text{def}}{=} \langle \nabla f(x) - \nabla f(y), x - y \rangle > 0$ and $V_f \stackrel{\text{def}}{=} \sup_{x,y\in\mathcal{Q}(x^0)} \beta_f(x,y)$.

Theorem 6. *Let function f be $L_{p,\nu}$-Hölder continuous, with finite relative size $D_q < \infty$ for $q = p + \nu$ (Definition 4) and γ-bounded Hessian change (Assumption 1). Algorithms* RN*,* UN *and* GRLS *find points in the ε-neighborhood, $f(x^k) - f(x^*) \leq \varepsilon$, in*

$$k \leq \mathcal{O}\left(\gamma \left(\frac{M_q D_q^q}{V_f}\right)^{\frac{1}{q-1}} \ln \frac{f_0}{\varepsilon} + \ln \frac{\left\|\nabla f(x^0)\right\|_{x^0}^* D}{\varepsilon}\right)$$

iterations, implying a global linear convergence rate.

Remark. *In view of* (26)*, analogous convergence guarantee (with a worse constant) can be proven for* GN.

Replacing relative size assumption with uniform star-convexity of degree s ($q > s \geq 2$), we can guarantee a global superlinear rate for RN and GN similarly to Kamzolov et al. [2024]. The proof is in Appendix F.

Definition 3. *For $s \geq 2$ and $\mu_s \geq 0$ we call function $f : \mathbb{R}^d \to \mathbb{R}$ μ_s-*uniformly star-convex *of degree s in local norms with respect to a minimizer x^* if $\forall x \in \mathbb{R}^d, \forall \eta \in [0,1]$ holds*

$$f(\eta x + (1-\eta)x^*) \leq \eta f(x) + (1-\eta) f_* - \frac{\eta(1-\eta)\mu_s}{s}\|x - x^*\|_x^s.$$

If this inequality holds for $\mu_s = 0$, we call function f star-convex in local norms *(w.r.t. minimizer x^*).*

Theorem 7. *Let the function $f : \mathbb{R}^d \to R$ be $L_{p,\nu}$-Hölder continuous (Definition 1) and μ_s-uniformly star-convex of degree s in local norms (Definition 3) and $q \overset{def}{=} p+\nu \geq s \geq 2$ then* RN *and* GN *have global decrease in functional value suboptimality,*

$$f(x^k) - f_* \leq \left(f(x^0) - f_*\right) \prod_{t=0}^{k-1} (1 - \hat{\eta}_t),$$

where $\hat{\eta}_k \in [0,1]$ is the only positive root of $E_k(\eta) \overset{def}{=} (1-\eta)\frac{\mu_s}{s} - \eta^{q-1}\left(\frac{M_q}{(p+1)!} + \frac{\sigma}{q}\right)\left\|x^k - x^*\right\|_{x^k}^{q-s}$.

*If $q = s$, then $\hat{\eta}_k$ is constant throughout all iterations and the rate is **globally linear**.*

*If $q > s$, then $\hat{\eta}_k$ is monotonically increasing as $\left\|x^k - x^*\right\|_{x^k}$ decreases, $1 - \hat{\eta}_k \to 0$, and therefore, the resutling rate is **globally superlinear**.*

8 Numerical experiments

Logistic regression

In Figure 2, we compare the performance of the proposed algorithms on binary classification on datasets from LIBSVM repository [Chang and Lin, 2011]. For datapoints $\{(a_i, b_i)\}_{i=1}^n$, where $a_i \in \mathbb{R}^d, b_i \in \{-1, +1\}$, and regularizer $\mu = 10^{-3}$, we aim to minimize

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) = \frac{1}{n}\sum_{i=1}^n \log\left(1 + e^{-b_i\langle a_i, x\rangle}\right) + \frac{\mu}{2}\|x\|_2^2 \right\}.$$

We initialize all methods at $x_0 = 10 \cdot [1, 1, \ldots, 1]^T \in \mathbb{R}^d$.

Polytope feasibility

In Figure 3, we compare proposed algorithms on *polytope feasibility* problem, aiming to find a point from a polytope $\mathcal{P} = \left\{ x \in \mathbb{R}^d : \langle a_i, x\rangle \leq b_i,\ 1 \leq i \leq n \right\}$, reformulated as

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) = \sum_{i=0}^n (\langle a_i, x\rangle - b_i)_+^p \right\}, \tag{28}$$

where $(t)_+ \overset{\text{def}}{=} \max\{t, 0\}$ and $p \geq 2$. We generate data points (a_i, b_i) and the solution x^* synthetically as $a_i, x^* \sim \mathcal{N}(0, 1)$ and set $b_i = \langle a_i, x^*\rangle$.

We initialize all methods at $x_0 = [1, 1, \ldots, 1]^T \in \mathbb{R}^d$.

Rosenbrock function

Linesearch procedures solve the abovementioned problems in just a few steps. For a more challenging task, Figure 1 presents the notorious d-dimensional *Rosenbrock* function,

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) = \sum_{i=0}^{d-1} [100(x_{i+1} - x_i^2)^2 + (1 - x_i)^2] \right\}. \tag{29}$$

Notably, the Rosenbrock function (29) is nonconvex, which breaks assumptions in our convergence theorems.

The function (29) has the global solution at $x^* = [1, \ldots, 1]^T$, and therefore we choose the initial point from a normal distribution, $x^0 \sim \mathcal{N}(0, I_d) \cdot 20$.

8.1 Experimental comparison

In Figures 2a, 3a, we compare higher-order methods *without* any linesearch procedures, namely RN (Algorithm 1), AICN [Hanzely et al., 2022] and Gradient Regularization of Newton Method (GRN) [Doikov et al., 2024, Alg. 1]. As additional baselines, we use the damped Newton method with a fixed fine-tuned stepsize and classical first-order Gradient Method (GM) [Nesterov, 2018]. RN and AICN show similar performance while GRN has a slight disadvantage. As expected, the first-order method GM that does not utilize Hessian has quicker iterations but slower per-iteration convergence.

In Figures 2b, 3b, we compare higher-order regularization methods *with* smoothness constant estimation procedures, UN and Super-universal Newton method [Doikov et al., 2024, Alg. 2]. As an additional baseline, we use the damped Newton method with a fixed but fine-tuned stepsize. We show that UN displays faster convergence than the Super-universal Newton method. Moreover, we show that the exponent of the regularization term β that appears in both UN and super-universal Newton method (8) does not have a significant impact on overall performance.

Figures 2c, 3c, 1 compare implicit linesearch procedures for Newton stepsizes, namely GRLS, Armijo stepsize, and Greedy Newton stepsize (GN) [Cauchy, 1847, Shea and Schmidt, 2024]. Our theory presents convergence guarantees for GRLS and GN with stepsizes limited to the interval $[0, 1]$. We go beyond this limitation and perform parameter linesearches over $\alpha \in \mathbb{R}_+$ instead.

Figures 2c, 3c demonstrate that on logistic regression and polytope feasibility problems, linesearch procedures GRLS and GN use almost indistinguishable stespsizes and converge faster than Armijo linesearch and fixed stepsize Newton. On the Rosenbrock function (Figure 1), GRLS significantly outperforms all other linesearches procedures.

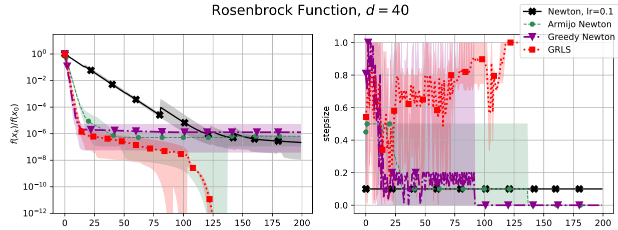
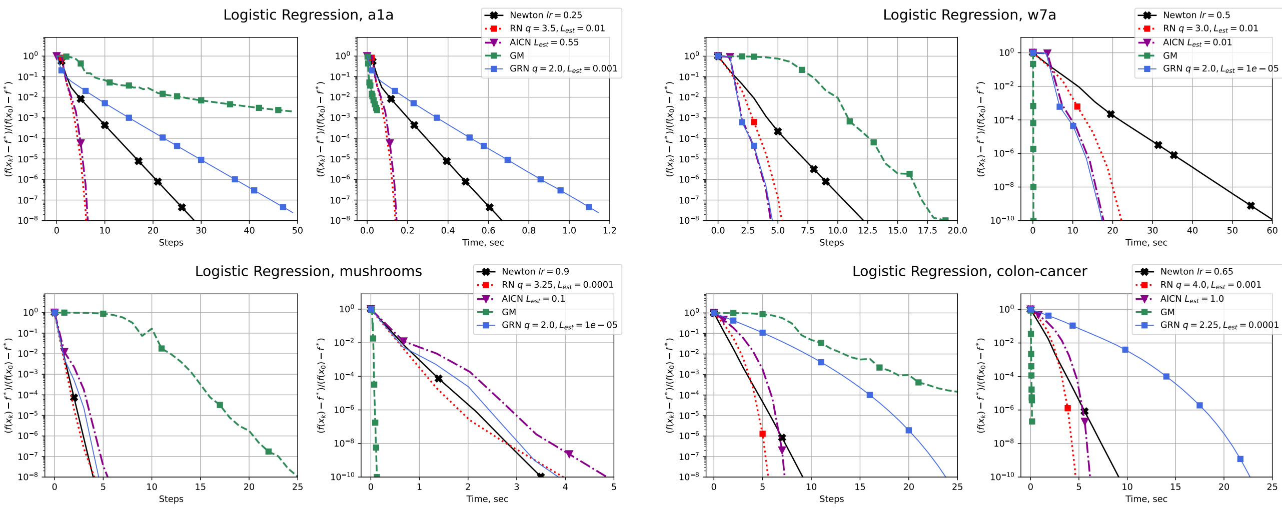
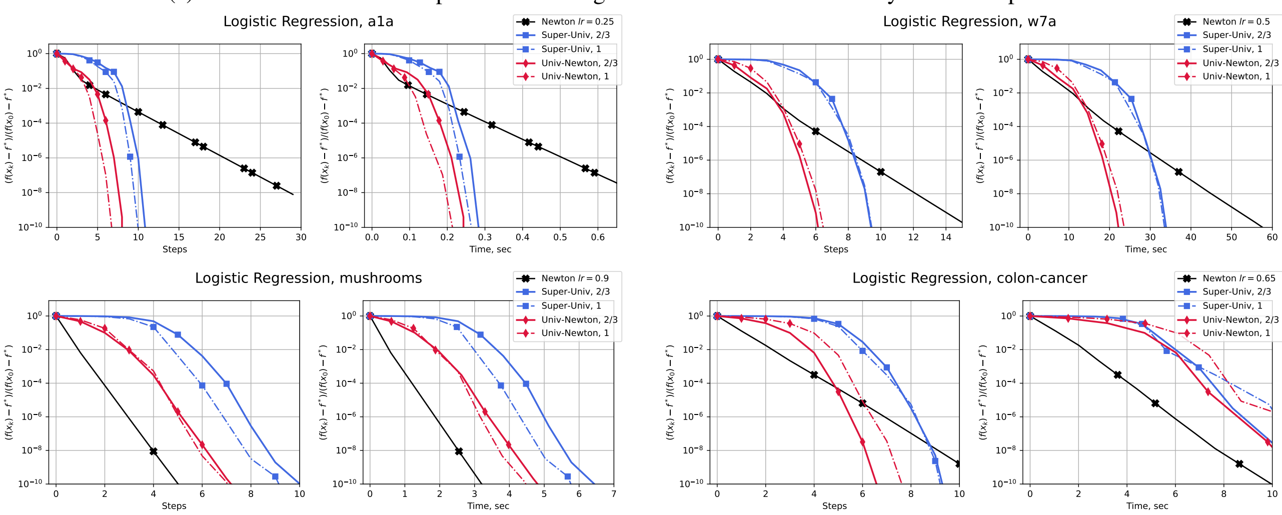


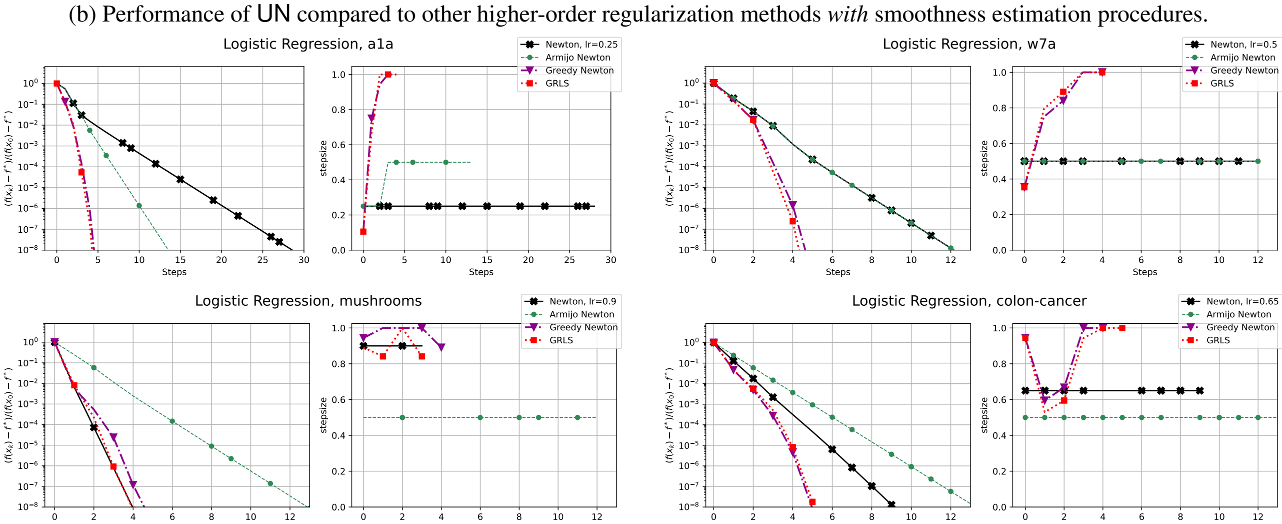
Figure 1: Performance of Newton method stepsize lineserch procedures on nonconvex **Rosenbrock function** (29). We plot mean ± standard deviation of 5 random initializations. We crop stepsize standard deviation at 0.



(a) Performance of RN compared to other higher-order methods *without* any linesearch procedure.

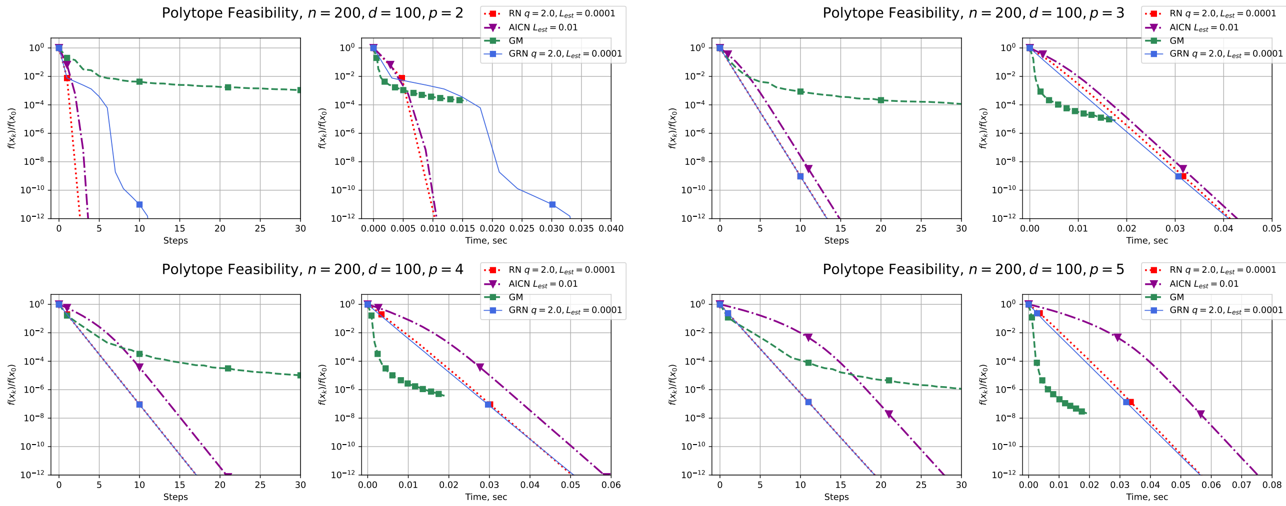


(b) Performance of UN compared to other higher-order regularization methods *with* smoothness estimation procedures.

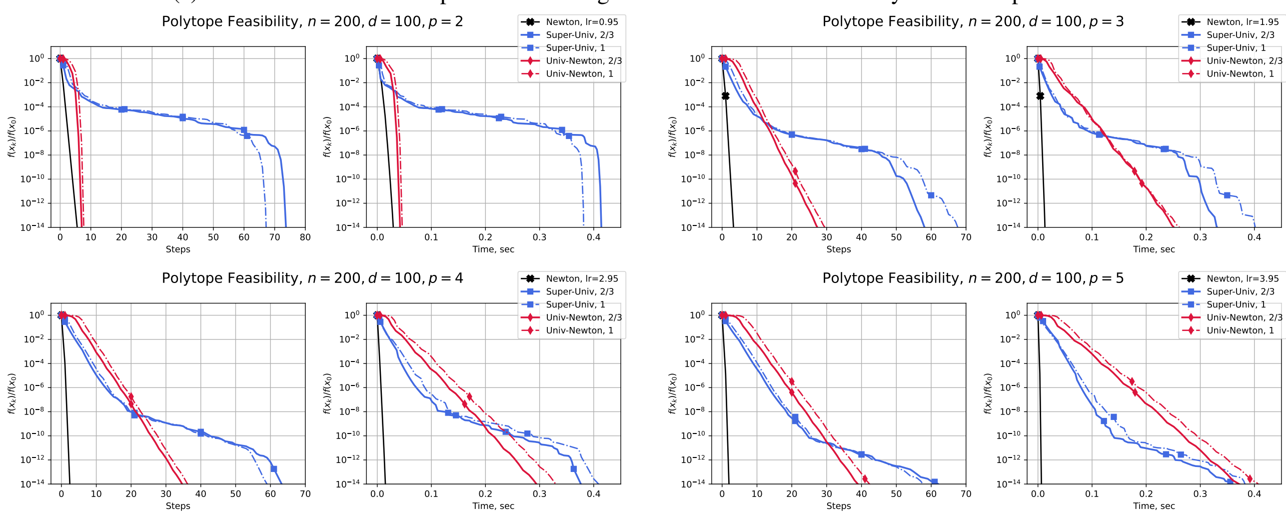


(c) Performance of Linesearch GRLS (24) compared to other linesearch procedures.

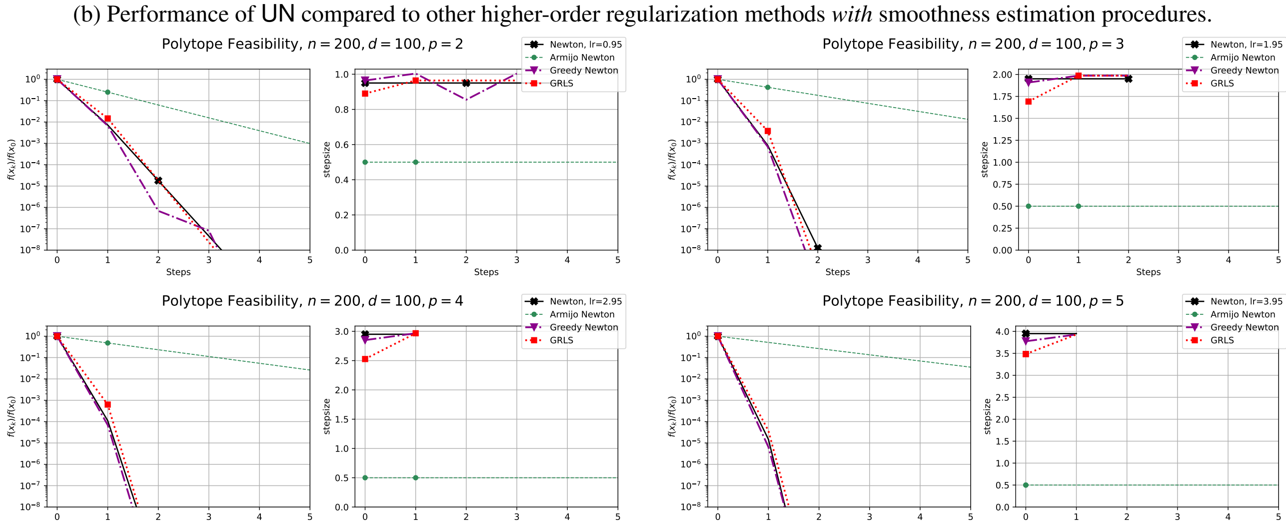
Figure 2: Binary classification **logistic regression** problem on LIBSVM datasets.



(a) Performance of RN compared to other higher-order methods *without* any linesearch procedure.



(b) Performance of UN compared to other higher-order regularization methods *with* smoothness estimation procedures.



(c) Performance of Linesearch GRLS (24) compared to other linesearch procedures.

Figure 3: **Polytope feasibility** problem (28) on a synthetic datasets.

References

Naman Agarwal and Elad Hazan. Lower bounds for higher-order convex optimization. In *Conference On Learning Theory*, pages 774–792. PMLR, 2018.

Jason Altschuler and Pablo Parrilo. Acceleration by stepsize hedging I: multi-step descent and the silver stepsize schedule, 2023.

Yossi Arjevani, Ohad Shamir, and Ron Shiff. Oracle complexity of second-order methods for smooth convex optimization. *Mathematical Programming*, 178(1):327–360, 2019.

Larry Armijo. Minimization of functions having lipschitz continuous first partial derivatives. *Pacific Journal of mathematics*, 16(1):1–3, 1966.

Augustin Cauchy. Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris*, 25 (1847):536–538, 1847.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):1–27, 2011. URL `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`.

Andrew Conn, Nicholas Gould, and Philippe Toint. *Trust Region Methods*. SIAM, 2000.

Nikita Doikov and Yurii Nesterov. Gradient regularization of Newton method with bregman distances. *Mathematical programming*, 204(1):1–25, 2024.

Nikita Doikov, Konstantin Mishchenko, and Yurii Nesterov. Super-universal regularized Newton method. *SIAM Journal on Optimization*, 34(1):27–56, 2024.

Alexander Gasnikov, Pavel Dvurechensky, Eduard Gorbunov, Evgeniya Vorontsova, Daniil Selikhanovych, and César Uribe. Optimal tensor methods in smooth convex and uniformly convexoptimization. In *Conference on Learning Theory*, pages 1374–1391. PMLR, 2019.

Robert Gower, Dmitry Kovalev, Felix Lieder, and Peter Richtárik. RSN: randomized subspace Newton. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 616–625. Curran Associates, Inc., 2019a. URL `http://papers.nips.cc/paper/8351-rsn-randomized-subspace-newton.pdf`.

Robert Gower, Dmitry Kovalev, Felix Lieder, and Peter Richtárik. RSN: randomized subspace Newton. *Advances in Neural Information Processing Systems*, 32, 2019b.

Benjamin Grimmer, Kevin Shu, and Alex Wang. Accelerated objective gap and gradient norm convergence for gradient descent via long steps. *arXiv preprint arXiv:2403.14045*, 2024.

Slavomír Hanzely. Sketch-and-project meets Newton method: Global $\mathcal{O}(1/k^2)$ convergence with low-rank updates. *arXiv preprint arXiv:2305.13082*, 2023.

Slavomír Hanzely, Dmitry Kamzolov, Dmitry Pasechnyuk, Alexander Gasnikov, Peter Richtárik, and Martin Takáč. A damped Newton method achieves global $\mathcal{O}(k^{-2})$ and local quadratic convergence rate. *Advances in Neural Information Processing Systems*, 35:25320–25334, 2022.

Florian Jarre and Philippe Toint. Simple examples for the failure of Newton's method with line search for strictly convex minimization. *Mathematical Programming*, 158(1):23–34, 2016.

Dmitry Kamzolov, Alexander Gasnikov, Pavel Dvurechensky, Artem Agafonov, and Martin Takáč. Exploiting higher order derivatives in convex optimization methods. In *Encyclopedia of Optimization*, pages 1–13. Springer, 2023.

Dmitry Kamzolov, Dmitry Pasechnyuk, Artem Agafonov, Alexander Gasnikov, and Martin Takáč. OPTAMI: Global superlinear convergence of high-order methods. *arXiv preprint arXiv:2410.04083*, 2024.

Leonid Kantorovich. Functional analysis and applied mathematics. *Uspekhi Matematicheskikh Nauk*, 3(6):89–185, 1948.

Sai Karimireddy, Sebastian Stich, and Martin Jaggi. Global linear convergence of Newton's method without strong-convexity or Lipschitz gradients. *arXiv:1806:0041*, 2018a.

Sai Karimireddy, Sebastian Stich, and Martin Jaggi. Global linear convergence of Newton's method without strong-convexity or Lipschitz gradients. *arXiv preprint arXiv:1806.00413*, 2018b.

Kenneth Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics*, 2(2):164–168, 1944.

Donald Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.

Walter Mascarenhas. On the divergence of line search methods. *Computational & Applied Mathematics*, 26(1):129–169, 2007.

Konstantin Mishchenko. Regularized Newton method with global convergence. *SIAM Journal on Optimization*, 33(3): 1440–1462, 2023.

Yurii Nesterov. *Lectures on Convex Optimization*. Springer Publishing Company, Incorporated, 2nd edition, 2018. ISBN 3319915770.

Yurii Nesterov. Implementable tensor methods in unconstrained convex optimization. *Mathematical Programming*, 186: 157–183, 2021.

Yurii Nesterov and Arkadi Nemirovski. *Interior-Point Polynomial Algorithms in Convex Programming*. SIAM, 1994.

Yurii Nesterov and Boris Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

Isaac Newton. *Philosophiae Naturalis Principia Mathematica*. Jussu Societatis Regiae ac Typis Josephi Streater, 1687.

Jorge Nocedal and Stephen Wright. *Numerical Optimization*. Springer, 1999.

Boris Polyak. *Introduction to Optimization.*, volume 1. Inc., Publications Division, New York, 1987.

Joseph Raphson. *Analysis Aequationum Universalis Seu Ad Aequationes Algebraicas Resolvendas Methodus Generalis & Expedita, Ex Nova Infinitarum Serierum Methodo, Deducta Ac Demonstrata*. Th. Braddyll, 1697.

Betty Shea and Mark Schmidt. Greedy newton: Newton's method with exact line search. *arXiv preprint arXiv:2401.06809*, 2024.

Thomas Simpson. *Essays on Several Curious and Useful Subjects, in Speculative and Mix'd Mathematicks. Illustrated by a Variety of Examples.* Printed by H. Woodfall, jun. for J. Nourse, at the Lamb without Temple-Bar, 1740.

Philip Wolfe. Convergence conditions for ascent methods. *SIAM review*, 11(2):226–235, 1969.

David Young. On richardson's method for solving linear systems with positive definite matrices. *Journal of Mathematics and Physics*, 32(1-4):243–255, 1953.

Tjalling Ypma. Historical development of the Newton–Raphson method. *SIAM Review*, 37(4):531–551, 1995.

Appendix

A Technical details of experiments

All hyperparameters were fine-tuned to achieve the best possible performance for both objectives and every dataset. All experiments were conducted on a workstation with specifications: AMD EPYC 7742 64-Core Processor with 32Gb of RAM. Source code is available at `https://anonymous.4open.science/r/root-newton-8D65`.

Extended comparison on Rosenbrock function

Here we present an extended comparison of linesearch procedures on Rosenbrock function (29) (similar to Figure 1), with 10 random initializations and the limit of 1000 steps. We observe that none of the considered algorithms consistently converge to the exact solution for all of the random seeds, and that GRLS performs better than the other linesearch methods.

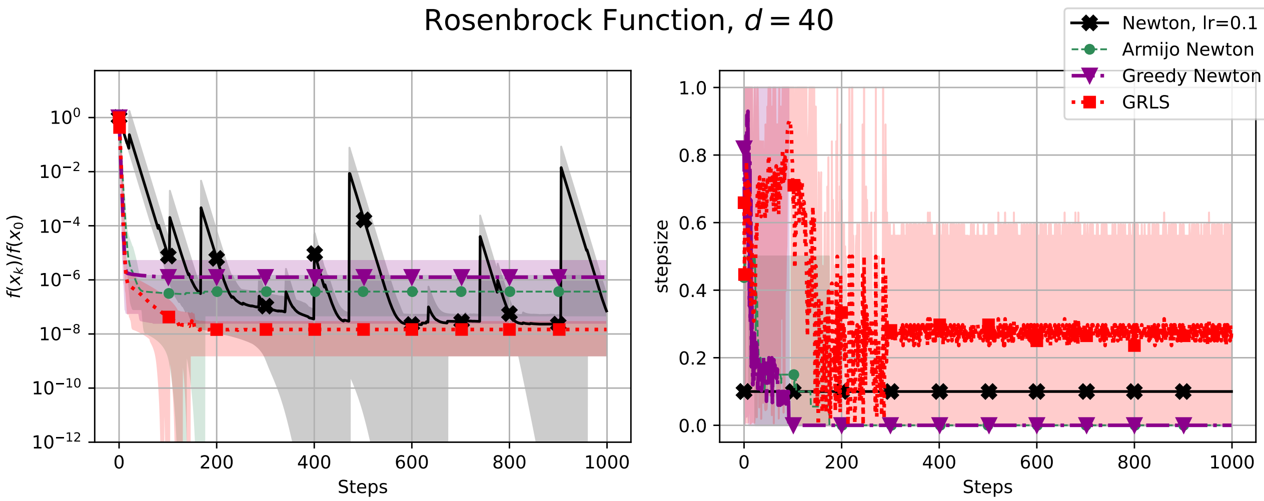


Figure 4: Performance of Newton method stepsize lineserch procedures on nonconvex **Rosenbrock function** (29). We plot mean $\pm$ standard deviation of 10 random initializations. We crop stepsize standard deviation at 0.

B Fast convergence guarantees for Greedy Newton linesearch

If the inequality $\|\nabla f(y)\|_{x^k}^* \leq \bar{c}\left\|\nabla f(x^k)\right\|_{x^k}^*$ holds for constant $\bar{c} \geq 0$, we have that for stepsizes in a range $[\underline{\alpha}, \overline{\alpha}]$ holds

$$\min_{\substack{\alpha \in [\underline{\alpha}, \overline{\alpha}] \\ y = x - \alpha n_{x^k}}} \frac{f(y) - f(x^k)}{\|\nabla f(x^k)\|_{x^k}^{*2}} \leq \bar{c}^2 \cdot \min_{\substack{\alpha \in [\underline{\alpha}, \overline{\alpha}] \\ y = x - \alpha n_{x^k}}} \frac{f(y) - f(x^k)}{\|\nabla f(y)\|_{x^k}^{*2}}, \tag{30}$$

proving that Greedy Newton minimizes the target metric of GRLS up to a constant $\times \bar{c}^2$. If we denote $\hat{c_5}$ constant with which GRLS satisfies Lemma 6, then Greedy Newton satisfies Lemma 6 with constant $\hat{c_5}\bar{c}^2$ and guarantee convergence similar to Corollary 1.

Now we are going to discuss how constant $\bar{c}$ can be found in different scenarios.

Remark (General M_q-Hölder continuous functions)**.** *To find $\bar{c}$ we note that Theorem 2 shows that stepsize $\theta_k \overset{def}{=} \frac{1-\alpha_k}{\alpha_k} \geq (9M_q)^{\frac{1}{q-1}} \left\|\nabla f(x^k)\right\|_{x^k}^{*\frac{q-2}{q-1}}$ for M_q-Hölder continuous function implies*

$$\frac{1}{2(1-\alpha_k)}\|\nabla f(y)\|_{x^k}^{*2}, \leq \left\langle \nabla f(y), \left[\nabla^2 f(x^k)\right]^{-1} \nabla f(x^k) \right\rangle \leq \|\nabla f(y)\|_{x^k}^* \left\|\nabla f(x^k)\right\|_{x^k}^*,$$

which after rearranging yields $\|\nabla f(y)\|_{x^k}^* \leq 2(1-\alpha_k)\left\|\nabla f(x^k)\right\|_{x^k}^*$. Therefore if

$$\alpha \leq \frac{1}{1 + (9M_q)^{\frac{1}{q-1}} \left\|\nabla f(x^k)\right\|_{x^k}^{*\frac{q-2}{q-1}}} \tag{31}$$

or equivalently

$$\overline{\alpha} \leq \left(1 + (9M_q)^{\frac{1}{q-1}} \left\|\nabla f(x^k)\right\|_{x^k}^{*\frac{q-2}{q-1}}\right)^{-1} \leq \left(1 + \sup_{q \in [2,4]} (9M_q)^{\frac{1}{q-1}} \left\|\nabla f(x^0)\right\|_{x^0}^{*\frac{q-2}{q-1}}\right)^{-1}. \tag{32}$$

In such case, $\overline{c}$ can be set as $\overline{c} = 2(1-\underline{\alpha})$.

Note that (32) is satisfied by smaller stepsizes, which damped Newton methods use globally until they converge to the neighborhood of the solution.

Remark (Hölder continuity of Hessians)**.** *For $L_{2,\nu}$-Holder, Lemma 8 yields*

$$\|\nabla f(y)\|_{x^k}^* \leq \left(|1-\alpha| + \frac{L_{2,\nu}}{1+\nu} \alpha^{1+\nu} \left\|\nabla f(x^k)\right\|_{x^k}^{*\nu} \right) \left\|\nabla f(x^k)\right\|_{x^k}^*, \tag{33}$$

ensuring that without any limitation on $\overline{\alpha}$

$$\overline{c}_x \overset{def}{=} \sup_{\alpha \in [\underline{\alpha}, \overline{\alpha}]} |1-\alpha| + \frac{L_{2,\nu}}{1+\nu} \alpha^{1+\nu} \left\|\nabla f(x^k)\right\|_{x^k}^{*\nu} \tag{34}$$

$$= \max_{\alpha \in \{\underline{\alpha}, \overline{\alpha}, 1\}} |1-\alpha| + \frac{L_{2,\nu}}{1+\nu} \alpha^{1+\nu} \left\|\nabla f(x^k)\right\|_{x^k}^{*\nu}. \tag{35}$$

For $\underline{\alpha} \leftarrow 0, \overline{\alpha} \leftarrow 1$, we can set

$$\overline{c} = \max\left\{1, \frac{L_{2,\nu}}{1+\nu} \left\|\nabla f(x^k)\right\|_{x^k}^{*\nu}\right\} \leq \max\left\{1, \frac{L_{2,\nu}}{1+\nu} \left\|\nabla f(x^0)\right\|_{x^0}^{*\nu}\right\}. \tag{36}$$

Remark ($L_{2,0}$-Hölder continuity)**.** *For $L_{2,0}$-Hölder functions with $L_{2,0} \geq 1$, constant $\overline{c}$ simplifies to $\overline{c} \overset{def}{=} \overline{\alpha} \frac{L_{2,0}}{2} + |1-\overline{\alpha}|$, because*

$$\begin{cases} \overline{\alpha}\left(\frac{L_{2,0}}{2} - 1\right) + 1 \geq \alpha\left(\frac{L_{2,0}}{2} - 1\right) + 1 \geq \frac{1}{2}, & \text{if } \alpha \leq 1, \\ \overline{\alpha}\left(\frac{L_{2,0}}{2} + 1\right) - 1 \geq \alpha\left(\frac{L_{2,0}}{2} + 1\right) - 1 \geq \frac{L_{2,0}}{2}, & \text{if } \alpha \geq 1. \end{cases} \tag{37}$$

C Connection between stepsizes and regularization

We show connections of particular stepsizes to regularized Newton methods. For fixed $\sigma > 0, \beta \geq 0$ define regularized model as

$$T_{\sigma,\beta}(x) \overset{\text{def}}{=} \operatorname*{argmin}_{y \in \mathbb{R}^d} \left\{ f(x) + \langle \nabla f(x), y-x \rangle + \frac{1}{2}\|y-x\|_x^2 + \frac{\sigma}{2+\beta}\|y-x\|_x^{2+\beta} \right\}. \tag{38}$$

We can define optimization algorithm RN as

$$x^{k+1} \overset{\text{def}}{=} T_{\sigma,\beta}\left(x^k\right) \tag{39}$$

By first-order optimality condition, solution of model $h^* \overset{\text{def}}{=} T_{\sigma,\beta}(x) - x$ satisfy

$$\left(1 + \sigma\|h^*\|_x^\beta\right) \left[\nabla^2 f(x)\right] h^* = -\nabla f(x), \tag{40}$$

$$h^* = -\underbrace{\left(1 + \sigma\|h^*\|_x^\beta\right)^{-1}}_{\overset{\text{def}}{=}\alpha > 0} \left[\nabla^2 f(x)\right]^{-1} \nabla f(x). \tag{41}$$

Now iterates of RN are in the direction of Newton method (for any σ and β) and we can write

$$h^* = -\alpha \left[\nabla^2 f(x)\right]^{-1} \nabla f(x), \tag{42}$$

$$\left[\nabla^2 f(x)\right] h^* = -\alpha \nabla f(x), \tag{43}$$

$$\|h^*\|_x = \alpha \|\nabla f(x)\|_x^*. \tag{44}$$

Substituting $\left[\nabla^2 f(x)\right] h^*$ back to the first-order optimality conditions we get

$$0 = \nabla f(x) \left(1 - \alpha - \alpha^{1+\beta} \sigma \|\nabla f(x)\|_x^{*\beta}\right). \tag{45}$$

Thus, α defined as a root of the polynomial

$$P[\alpha] \stackrel{\text{def}}{=} 1 - \alpha - \alpha^{1+\beta}\sigma \|\nabla f(x)\|_x^{*\beta} \tag{46}$$

satisfies first-order optimality condition. Note that $P[0] > 0$ and $P[1] \leq 0$, hence P has root on interval $(0, 1]$. This will be the stepsize of our algorithm. Also note that P is monotone on $\mathbb{R}_+$,

$$P'[\alpha] = -1 - (1+\beta)\alpha^{\beta}\sigma \|\nabla f(x)\|_x^{*\beta} < 0, \tag{47}$$

and consequently, the positive root of P is unique.

D Extra smoothness relations

Let $\gamma \in [0, 1]$. From Hölders continuity, triangle inequality and definition of $L_{p,\nu}$,

$$\begin{aligned}
\left\|\nabla^3 f(x)[y-x]\right\|_{op} &\leq \left\|\nabla^2 f(x) - \nabla^2 f(y)\right\|_{op} + \frac{L_{3,\nu}}{1+\nu}\|y-x\|_x^{1+\nu} && (48)\\
&\leq L_{2,\gamma}\|x-y\|_x^{\gamma} + \frac{L_{3,\nu}}{1+\nu}\|y-x\|_x^{1+\nu} && (49)
\end{aligned}$$

For $y \leftarrow x + \tau h$, where $\|h\|_x = 1, \tau > 0$, we can continue

$$\begin{aligned}
\left\|\nabla^3 f(x)\right\|_{op} &\leq \frac{L_{2,\gamma}}{\tau^{1-\gamma}} + \frac{L_{3,\nu}}{1+\nu}\tau^{\nu}, && (50)\\
&\leq \frac{2+\nu}{1+\nu}[L_{2,\gamma}]^{\frac{\nu}{1+\nu-\gamma}}\tau^{1-\gamma}[L_{3,\nu}]^{\frac{1}{1+\nu-\gamma}}, \qquad // \text{ by } \tau \leftarrow \left[\frac{L_{2,\gamma}}{L_{3,\nu}}\right]^{\frac{1}{1+\nu-\gamma}} && (51)\\
&\leq \frac{3}{2}\sqrt{L_{2,0}L_{3,1}}, \qquad // \text{ by } \gamma \leftarrow 0, \nu \leftarrow 1 && (52)
\end{aligned}$$

and we can summarize

$$\begin{aligned}
L_{3,0} = \sup_{x \neq y}\left\|\nabla^3 f(x) - \nabla^3 f(y)\right\|_{op} &\leq \sup_{x\neq y}\left(\left\|\nabla^3 f(x)\right\|_{op} + \left\|\nabla^3 f(y)\right\|_{op}\right) = 2\sup_x \left\|\nabla^3 f(x)\right\|_{op} && (53)\\
&\leq \begin{cases} 2L_{2,1} \\ 3\sqrt{L_{2,0}L_{3,1}} \end{cases}. && (54)
\end{aligned}$$

Lemma 8. *If $L_{2,\nu}$ exists, for points $x^k, x^{k+1} = x^k - \alpha_k \left[\nabla^2 f(x^k)\right]^{-1}\nabla f(x^k)$ holds decrease*

$$\left\|\nabla f(x^{k+1})\right\|_{x^k}^{*} \leq \left(\theta_k + \frac{L_{2,\nu}}{1+\nu}\alpha_k^{\nu}\left\|\nabla f(x^k)\right\|_{x^k}^{*\nu}\right)\alpha_k\left\|\nabla f(x^k)\right\|_{x^k}^{*},$$

and hence, if $\nu > 0$ and $\theta_k \geq \left\|\nabla f(x^k)\right\|_{x^k}^{*\varepsilon}$ for $\varepsilon > 0$, and if the bound (131) *exists (meaning that the Hessian does not change much), we have guaranteed superlinear local rate.*

Remark. *Hanzely et al. [2022] shows that $L_{2,1}$-Hölder continuity implies self-concordance, and [Nesterov, 2018, Theorem 4.1.3] proves that self-concordance implies positive definiteness of Hessian $\nabla^2 f$ the domain of function f contains no straight line.*

E Simplified regularization

In the view of Section 1.4 and Lemma 2, we can bound the majorization as

$$\begin{aligned}
T_{\sigma,\beta}(x) &= \operatorname*{argmin}_{y\in\mathbb{R}^d}\left\{f(x) + \langle\nabla f(x), y-x\rangle + \frac{1}{2}\|y-x\|_x^2 + \frac{\sigma}{2+\beta}\|y-x\|_x^{2+\beta}\right\} && (55)\\
&\leq \operatorname*{argmin}_{y\in\mathbb{R}^d}\left\{f(x) + \langle\nabla f(x), y-x\rangle + \frac{\beta}{2(\beta+2)} + \frac{\sigma+1}{2+\beta}\|y-x\|_x^{2+\beta}\right\} && (56)\\
&= x - \left(\frac{1}{(\sigma+1)\|\nabla f(x^k)\|_{x^k}^{*\beta}}\right)^{\frac{1}{1+\beta}}\left[\nabla^2 f(x)\right]^{-1}\nabla f(x), && (57)
\end{aligned}$$

where stepsize was obtained as the positive root of polynomial $P[\alpha] \stackrel{\text{def}}{=} 1 - \alpha^{1+\beta}(\sigma+1)\left\|\nabla f(x^k)\right\|_{x^k}^{*\beta}$.

Surprisingly, stepsize is unbounded, and when $\|\nabla f(x)\|_x^* \to 0$, then $\alpha \to \infty$. This puzzling result has a simple explanation – such stepsize converges only to a neighborhood of the solution.

In practice, we could not observe stepsize larger than 5 on any considered dataset. When close to the solution and the stepsize becomes larger than one, algorithm (57) stops converging closer to the solution, and functional values oscillate.

F Analysis under uniform star-convexity assumption in local norms

Proof of Theorem 7. We have that updates of RN with $q = p + \nu = 2 + \beta$ and any $\sigma \geq M_q$ can be written as

$$f(x^{k+1}) \leq \Phi_{x^k}(x^{k+1}) + \frac{\sigma}{q}\left\|x^{k+1} - x^k\right\|_{x^k}^q \tag{58}$$

$$= \min_{y \in \mathbb{R}^d} \left\{ \Phi_{x^k}(y) + \frac{\sigma}{q}\|y - x\|_{x^k}^q \right\}, \tag{59}$$

using standard integration arguments from M_q-Hölder continuity

$$\leq \min_{y \in \mathbb{R}^d} \left\{ f(y) + \frac{M_q}{(p+1)!}\left\|y - x^k\right\|_{x^k}^q + \frac{\sigma}{q}\left\|y - x^k\right\|_{x^k}^q \right\} \tag{60}$$

$$= \min_{y \in \mathbb{R}^d} \left\{ f(y) + \left(\frac{M_q}{(p+1)!} + \frac{\sigma}{q}\right)\left\|y - x^k\right\|_{x^k}^q \right\}, \tag{61}$$

setting $y \leftarrow x + \eta_k(x^* - x^k)$ for arbitrary $\eta_k \in [0,1]$,

$$\leq f\left(x^k + \eta_k(x^* - x^k)\right) + \eta_k^q \left(\frac{M_q}{(p+1)!} + \frac{\sigma}{q}\right)\left\|x^k - x^*\right\|_{x^k}^q, \tag{62}$$

assuming μ_s-strong star-convexity for $q \geq s \geq 2$,

$$\leq (1-\eta_k) f(x^k) + \eta_k f_* - \frac{\eta_k(1-\eta_k)\mu_s}{s}\left\|x^k - x^*\right\|_{x^k}^s + \eta_k^q \left(\frac{M_q}{(p+1)!} + \frac{\sigma}{q}\right)\left\|x^k - x^*\right\|_{x^k}^q, \tag{63}$$

denoting functional suboptimality $\delta_k \stackrel{\text{def}}{=} f(x^k) - f_*$,

$$\delta_{k+1} \leq (1-\eta_k)\delta_k - \eta_k\left\|x^k - x^*\right\|_{x^k}^s \left((1-\eta_k)\frac{\mu_s}{s} - \eta_k^{q-1}\left(\frac{M_q}{(p+1)!} + \frac{\sigma}{q}\right)\left\|x^k - x^*\right\|_{x^k}^{q-s}\right). \tag{64}$$

Denote expression $E(\eta) \stackrel{\text{def}}{=} (1-\eta)\frac{\mu_s}{s} - \eta^{q-1}\left(\frac{M_q}{(p+1)!} + \frac{\sigma}{q}\right)\|x - x^*\|_x^{q-s}$ for $\eta \in [0,1]$. Observe that $E'(\eta) < 0$ and therefore E is monotonically decreasing on $\mathbb{R}^+$; with $E(0) \geq 0 \leq E(1)$ we can conclude that it has a unique root $\hat{\eta}$ on $[0,1]$. With choice $\eta \leftarrow \hat{\eta}$ in the last inequality we can conclude global convergence rate

$$\delta_{k+1} \leq (1 - \hat{\eta}_k)\delta_k. \tag{65}$$

Note that the root of the expression E is inversely proportional to the distance from the solution $\|x - x^*\|_x$, and therefore as the method converges, $x^k \to x^*$, then the size of its root increases $\hat{\eta}_k \to 1$. Therefore, the global convergence rate (65) is superlinear.

Unrolling the recurrence (65) yields the inequality from the Theorem 7.

Note that the decrease is based solely on the decrease in functional values, which allows us to prove the identical guarantee for Greedy Newton linesearch GN. In particular, GN implies $f(x_{\mathsf{GN}}^+) \leq f(x_{\mathsf{RN}}^+)$, and we can analogically conclude

$$f(x_{\mathsf{GN}}^{k+1}) - f_* \leq \left(f(x_{\mathsf{GN}}^k) - f_*\right)(1 - \hat{\eta}_k). \tag{66}$$

□

G Analysis under s-relative size assumption

In this section, we present global convergence guarantees under a novel characteristic called s-relative size recently proposed by Doikov et al. [2024].

Definition 4 ([Doikov et al., 2024])**.** *For strictly convex function $f : \mathbb{R}^d \to \mathbb{R}$ we call* s-relative size *characteristic*

$$D_s \overset{def}{=} \sup_{x,y\in\mathcal{Q}(x^0)} \left\{ \|x-y\|_x \left(\frac{V_f}{\beta_f(x,y)} \right)^{\frac{1}{s}} \right\},$$

where $\beta_f(x,y) \overset{def}{=} \langle \nabla f(x) - \nabla f(y), x-y \rangle > 0$ and $V_f \overset{def}{=} \sup_{x,y\in\mathcal{Q}(x^0)} \beta_f(x,y)$.

Theorem 8. *Let function f be $L_{p,\nu}$-Hölder continuous, with finite relative size $D_q < \infty$ for $q = p+\nu$ (Definition 4) and γ-bounded Hessian change (Assumption 1). Algorithms* RN, UN *and* GRLS *find points in the ε-neighborhood, $f(x^k) - f(x^*) \le \varepsilon$, in*

$$k \le \mathcal{O}\left(\gamma \left(\frac{M_q D_q^q}{V_f} \right)^{\frac{1}{q-1}} \ln \frac{f_0}{\varepsilon} + \ln \frac{\left\|\nabla f(x^0)\right\|_{x^0}^* D}{\varepsilon} \right)$$

iterations, enjoying a global linear convergence rate.

Strict convexity implies $\beta_f(x,y) > 0$, we also have $\lim_{s\to\infty} D_s = D$, also $\frac{\beta_f(x,y)}{V_f} \le 1$, and

$$\langle \nabla f(x) - \nabla f(y), x-y \rangle \ge V_f \left(\frac{\|x-y\|_x}{D_s} \right)^s \tag{67}$$

Characteristic D_s is log-convex function in s, and if $D_{s_1}, D_{s_2} < \infty$, then for $2 \le s_1 \le s \le s_2$ holds

$$D_s \le [D_{s_1}]^{\frac{s_2-s}{s_2-s_1}} [D_{s_2}]^{\frac{s-s_1}{s_2-s_1}}, \tag{68}$$

and D_s is continuous on this segment.

Remark. *For self-concordant functions, it holds $\beta_f(x,y) \ge \|y-x\|_x^2$, and $D_s \le D^{1-\frac{2}{s}} V_f^{\frac{1}{s}}$.*

Remark. *For functions such that $\beta_f(x,y) \ge \mu_s \|x-y\|_x^s$ it holds $D_s \le \left(\frac{V_f}{\mu_s}\right)^{\frac{1}{s}}$. In particular, for self-concordant functions holds $\beta_f(x,y) \ge \|y-x\|_x^2$, and therefore $D_2 \le \sqrt{V_f}$.*

Assumption 2. *For some $s \ge 2$, value of D_s is finite, $D_s < \infty$.*

Lemma 9. *For any $2 \le s \le q$, we have*

$$\left(\frac{D_q}{D}\right)^q \le \left(\frac{D_s}{D}\right)^s \tag{69}$$

Proof of Lemma 9. Analogical to Doikov et al. [2024]. □

Now for any $x, y \in \mathcal{Q}(x^0)$,

$$f(y) = f(x) + \langle \nabla f(x), y-x \rangle + \int_0^1 \frac{1}{\tau} \langle \nabla f(x+\tau(y-x)) - \nabla f(x), \tau(y-x) \rangle \, d\tau \tag{70}$$

$$\ge f(x) + \langle \nabla f(x), y-x \rangle + \frac{1}{s} V_f \left(\frac{\|x-y\|_x}{D_s} \right)^s, \tag{71}$$

and minimizing both sides w.r.t. y independently, we get

$$\frac{s-1}{s} \left(\frac{D_s \|\nabla f(x)\|_x^*}{V_f} \right)^{\frac{s}{s-1}} \ge \frac{f(x) - f_*}{V_f} \tag{72}$$

Let us denote some constants that will appear in proofs.

$$\hat{\gamma} \stackrel{\text{def}}{=} \frac{q(s-1)}{(q-1)s} \in \left[\frac{2}{3}, 2\right], \qquad \text{and} \qquad 1 - \hat{\gamma} = \frac{q-s}{(q-1)s} \tag{73}$$

$$\omega_{q,s} \stackrel{\text{def}}{=} \frac{1}{2}\left(\frac{s}{s-1}\right)^{\hat{\gamma}} \left(\frac{V_f^{\frac{q}{s}}}{9M_q D_s^q}\right)^{\frac{1}{q-1}} = \frac{1}{2}\left(\frac{s}{s-1}\right)^{\frac{q(s-1)}{(q-1)s}} \left(\frac{V_f^{\frac{q}{s}}}{9M_q D_s^q}\right)^{\frac{1}{q-1}} \tag{74}$$

$$C_q \stackrel{\text{def}}{=} 2\gamma(q-1)(9M_q)^{\frac{1}{q-1}} D^{\frac{q}{q-1}} \tag{75}$$

Note that $\frac{\omega_{q,s} C_q}{\gamma(q-1)} = \left(\left(\frac{s}{s-1}\right)^{\frac{s-1}{s}} \frac{V_f^{\frac{1}{s}} D}{D_s}\right)^{\frac{q}{q-1}}$.

Lemma 10. *For $q \in [2,4]$ and $s \in [2, \infty)$, we have*

$$\frac{1}{(\hat{\gamma}-1) f_{k+1}^{\hat{\gamma}-1}} - \frac{1}{(\hat{\gamma}-1) f_k^{\hat{\gamma}-1}} \geq \omega_{q,s} \frac{\|\nabla f(x_{k+1})\|_{x_{k+1}}^{*2}}{\|\nabla f(x_k)\|_{x_k}^{*2}}. \tag{76}$$

Proof. Analogically to Doikov et al. [2024].

$$f_k - f_{k+1} \stackrel{(22)}{\geq} \frac{1}{2}\left(\frac{1}{9M_q}\right)^{\frac{1}{q-1}} \frac{\left\|\nabla f(x^k)\right\|_{x^k}^{*2}}{\|\nabla f(x^k)\|_{x^k}^{*2}} \left\|\nabla f(x^k)\right\|_{x^k}^{*\frac{q}{q-1}} \tag{77}$$

$$\stackrel{(72)}{\geq} \frac{1}{2}\left(\frac{1}{9M_q}\right)^{\frac{1}{q-1}} \frac{\left\|\nabla f(x^{k+1})\right\|_{x^k}^{*2}}{\|\nabla f(x^k)\|_{x^k}^{*2}} \left(\frac{V_f^{\frac{1}{s}}}{D_s}\right)^{\frac{q}{q-1}} \left(\frac{s}{s-1}\right)^{\hat{\gamma}} f_k^{\hat{\gamma}} \tag{78}$$

$$= \frac{1}{2}\left(\frac{s}{s-1}\right)^{\hat{\gamma}} \left(\frac{V_f^{\frac{q}{s}}}{9M_q D_s^q}\right)^{\frac{1}{q-1}} \frac{\left\|\nabla f(x^{k+1})\right\|_{x^k}^{*2}}{\|\nabla f(x^k)\|_{x^k}^{*2}} f_k^{\hat{\gamma}} \tag{79}$$

$$= \omega_{q,s} \frac{\left\|\nabla f(x^{k+1})\right\|_{x^k}^{*2}}{\|\nabla f(x^k)\|_{x^k}^{*2}} f_k^{\hat{\gamma}}. \tag{80}$$

If $s \geq q$, then $\hat{\gamma} \in [1,2]$ and the function $y(x) \stackrel{\text{def}}{=} x^{\hat{\gamma}-1}$ is concave. With monotonicity of $\{f_k\}_{k\geq 0}$, we have

$$\frac{1}{(\hat{\gamma}-1) f_{k+1}^{\hat{\gamma}-1}} - \frac{1}{(\hat{\gamma}-1) f_k^{\hat{\gamma}-1}} = \frac{f_k^{\hat{\gamma}-1} - f_{k+1}^{\hat{\gamma}-1}}{(\hat{\gamma}-1) f_{k+1}^{\hat{\gamma}-1} f_k^{\hat{\gamma}-1}} \geq \frac{f_k - f_{k+1}}{f_{k+1}^{\hat{\gamma}-1} f_k} \geq \omega_{q,s} \frac{\|\nabla f(x_{k+1})\|_{x_k}^{*2}}{\|\nabla f(x_k)\|_{x_k}^{*2}}. \tag{81}$$

If $2 \leq s < q$, then $\hat{\gamma} < 1$ and the function $y(x) \stackrel{\text{def}}{=} x^{\hat{\gamma}-1}$ is concave. We have

$$\frac{1}{(\hat{\gamma}-1) f_{k+1}^{\hat{\gamma}-1}} - \frac{1}{(\hat{\gamma}-1) f_k^{\hat{\gamma}-1}} = \frac{f_k^{1-\hat{\gamma}} - f_{k+1}^{1-\hat{\gamma}}}{1-\hat{\gamma}} \geq \frac{f_k - f_{k+1}}{f_k^{\hat{\gamma}}} \geq \omega_{q,s} \frac{\|\nabla f(x_{k+1})\|_{x_k}^{*2}}{\|\nabla f(x_k)\|_{x_k}^{*2}}. \tag{82}$$

□

Theorem 9. *Let function f be $L_{p,\nu}$-Hölder continuous with finite s-relative size and γ-bounded Hessian change, $M_q, D_s < \infty$ for some $q \in [2,4]$ and $s \geq q$ and sequence of iterates $x^0, \dots, x^k$ by generated by one of the algorithms* RN, UN, GRLS. *If all iterates had function suboptimality worse than $\varepsilon > 0$, $f_t \geq \varepsilon$ for $t \in \{0, \dots k\}$, then the* *algorithm did at most*

$$k \leq \frac{\gamma}{\omega_{q,s}(\hat{\gamma}-1)}\left[\frac{1}{f_k^{\hat{\gamma}-1}} - \frac{1}{f_0^{\hat{\gamma}-1}}\right] + 2\ln\frac{\left\|\nabla f(x^0)\right\|_{x^0}^* D}{f_k} \tag{83}$$

$$\leq 2\gamma\frac{s(q-1)}{s-q}\left(\frac{s-1}{s}\right)^{\frac{q(s-1)}{(q-1)s}}\left(\frac{9M_q D_s^q}{V_f^{\frac{q}{s}}}\right)^{\frac{1}{q-1}}\left[\varepsilon^{-\frac{s-q}{s(q-1)}} - f_0^{-\frac{s-q}{s(q-1)}}\right] + 2\ln\frac{\left\|\nabla f(x^0)\right\|_{x^0}^* D}{\varepsilon} \tag{84}$$

steps. If $s = q$*, treating RHS as limit together with* $\lim_{a\to 0}\frac{b^{-a}-c^{-a}}{a} = \ln\left(\frac{c}{b}\right)$ *guarantees the linear convergence rate*

$$k \leq 2\gamma\frac{q-1}{q}\left(\frac{9M_q D_q^q}{V_f}\right)^{\frac{1}{q-1}}\ln\frac{f_0}{\varepsilon} + 2\ln\frac{\left\|\nabla f(x^0)\right\|_{x^0}^* D}{\varepsilon}. \tag{85}$$

Remark. *We can analogically guarantee the global linear convergence of Greedy Newton linesearch* GN (27)*, but with a slightly different constant.*

Proof. Telescoping Lemma 10,

$$\frac{1}{(\hat{\gamma}-1)f_k^{\hat{\gamma}-1}} - \frac{1}{(\hat{\gamma}-1)f_0^{\hat{\gamma}-1}} \geq \omega_{q,s}\sum_{t=0}^{k-1}\frac{\left\|\nabla f(x^{t+1})\right\|_{x^t}^{*2}}{\left\|\nabla f(x^t)\right\|_{x^t}^{*2}} \tag{86}$$

$$\geq k\omega_{q,s}\left(\prod_{t=0}^{k-1}\frac{\left\|\nabla f(x^{t+1})\right\|_{x^t}^{*2}}{\left\|\nabla f(x^t)\right\|_{x^t}^{*2}}\right)^{\frac{1}{k}} \tag{87}$$

$$\geq \frac{k\omega_{q,s}}{\gamma}\left(\frac{f_k}{\left\|\nabla f(x^0)\right\|_{x^0}^* D}\right)^{\frac{k}{2}} \tag{88}$$

$$\geq \frac{k\omega_{q,s}}{\gamma}\exp\left(-\frac{2}{k}\ln\frac{\left\|\nabla f(x^0)\right\|_{x^0}^* D}{f_k}\right) \tag{89}$$

$$\geq \frac{k\omega_{q,s}}{\gamma}\left(1 - \frac{2}{k}\ln\frac{\left\|\nabla f(x^0)\right\|_{x^0}^* D}{f_k}\right) \tag{90}$$

$$= \frac{k\omega_{q,s}}{\gamma} - \frac{2\omega_{q,s}}{\gamma}\ln\frac{\left\|\nabla f(x^0)\right\|_{x^0}^* D}{f_k}, \tag{91}$$

hence

$$k \leq \frac{\gamma}{\omega_{q,s}(\hat{\gamma}-1)}\left[\frac{1}{f_k^{\hat{\gamma}-1}} - \frac{1}{f_0^{\hat{\gamma}-1}}\right] + 2\ln\frac{\left\|\nabla f(x^0)\right\|_{x^0}^* D}{f_k} \tag{92}$$

$$\leq \frac{\gamma}{\omega_{q,s}(\hat{\gamma}-1)}\left[\frac{1}{f_k^{\hat{\gamma}-1}} - \frac{1}{f_0^{\hat{\gamma}-1}}\right] + 2\ln\frac{\left\|\nabla f(x^0)\right\|_{x^0}^* D}{\varepsilon}. \tag{93}$$

□

Theorem 10. *Let funciton* f *be* $L_{p,\nu}$*-Hölder continuous with finite* s*-relative size and* γ*-bounded Hessian change,* $M_q, D_s < \infty$ *for some* $q \in [2,4]$ *and* $2 \leq s \leq q$ *and sequence of iterates* $x^0, \ldots, x^k$ *by generated by one of the algorithms* RN, UN, GRLS*. If all iterates were far from solution,* $f_t \geq \varepsilon > 0$ *and* $g_t \overset{def}{=} \left\|\nabla f(x^t)\right\|_{x^t}^* \geq \delta > 0$ *for* $t \in \{0, \ldots k\}$*, then the algorithm did at most*

$$k \leq 2\gamma\frac{q}{s}\left(\frac{s-1}{s}\right)^{\frac{s-1}{q-1}}\left(\frac{9M_q D_s^s D^{q-s}}{V_f}\right)^{\frac{1}{q-1}}\frac{s(q-1)}{q-s}\left[1 - \frac{s}{q}\left(\left(\frac{s}{s-1}\right)^{s-1}\frac{D_s^s}{V_f D^s}\varepsilon\right)^{\frac{q-s}{s(q-1)}}\right] + 2\ln\frac{g_0}{\delta} \tag{94}$$

steps. If $s = q$, treating RHS as a limit guarantees linear convergence rate

$$k \leq 2\gamma \frac{q-1}{q} \left(\frac{9 M_q D_q^q}{V_f} \right)^{\frac{1}{q-1}} \ln \left(\left(\frac{q}{q-1} \right)^{q-1} \frac{V_f D^q}{D_q^q \varepsilon} \right) + 2 \ln \frac{g_0}{\delta}. \tag{95}$$

Proof. Note $1 - \hat{\gamma} = \frac{q-s}{s(q-1)} > 0$. Let's split the analysis of the method into two stages, $k = m + n$. With $C_q = 2\gamma(q-1)(9M_q)^{\frac{1}{q-1}} D^{\frac{q}{q-1}}$, we bound the first stage,

$$C_q \frac{1}{f_m^{\frac{1}{q-1}}} \geq C_q \left[\frac{1}{f_m^{\frac{1}{q-1}}} - \frac{1}{f_0^{\frac{1}{q-1}}} \right] \overset{(126)}{\geq} m \left(\frac{g_m}{g_0} \right)^{\frac{2}{m}} = m \exp\left(\frac{2}{m} \ln \frac{g_m}{g_0} \right) \tag{96}$$

$$\geq m + 2 \ln \frac{g_m}{g_0} = m + 2 \ln \frac{g_m}{\delta} - 2 \ln \frac{g_0}{\delta}. \tag{97}$$

For the second stage, telescoping inequalities for $t = m, \ldots, k-1$

$$\frac{1}{\omega_{q,s}(1-\hat{\gamma})} \left[f_{t+1}^{1-\hat{\gamma}} - f_t^{1-\hat{\gamma}} \right] \geq \frac{\|\nabla f(x_{t+1})\|_{x_{t+1}}^{*2}}{\|\nabla f(x_t)\|_{x_t}^{*2}}, \tag{98}$$

we get

$$\frac{\gamma}{\omega_{q,s}(1-\hat{\gamma})} \left[f_m^{1-\hat{\gamma}} - \varepsilon^{1-\hat{\gamma}} \right] \geq \gamma \sum_{t=m}^{k-1} \frac{\|\nabla f(x_{t+1})\|_{x_{t+1}}^{*2}}{\|\nabla f(x_t)\|_{x_t}^{*2}} \geq n \left(\frac{g_k}{g_m} \right)^{\frac{2}{n}} \geq n \left(\frac{\delta}{g_m} \right)^{\frac{2}{n}} \tag{99}$$

$$\geq n - 2 \ln \frac{g_m}{\delta}. \tag{100}$$

Expressing n, m from the inequalities above and adding them together yields

$$k \leq C_q \frac{1}{f_m^{\frac{1}{q-1}}} + \frac{\gamma}{\omega_{q,s}(1-\hat{\gamma})} \left[f_m^{1-\hat{\gamma}} - \varepsilon^{1-\hat{\gamma}} \right] + 2 \ln \frac{g_0}{\delta}. \tag{101}$$

Note that $1 - \hat{\gamma} = \frac{q-s}{s(q-1)}$. Minimizer of RHS in f_m is achieved at

$$f_m^* \overset{\text{def}}{=} \left(\frac{C_q \omega_{q,s}}{\gamma(q-1)} \right)^{\frac{s(q-1)}{q}} = \left(\frac{s}{s-1} \right)^{\frac{s}{-}1} \frac{V_f D^s}{D_s^s}. \tag{102}$$

Substituting definitions of $f_m^*, \omega_{q,s}, C_q, \hat{\gamma}$ into the terms we get

$$C_q \frac{1}{f_m^{*\,\frac{1}{q-1}}} = 2\gamma(q-1) \left(\frac{s-1}{s} \right)^{\frac{s-1}{q-1}} \left(\frac{9 M_q D_s^s D^{q-s}}{V_f} \right)^{\frac{1}{q-1}},$$

$$\frac{\gamma}{\omega_{q,s}(1-\hat{\gamma})} f_m^{*\,(1-\hat{\gamma})} = \gamma \frac{s(q-1)}{q-s} \frac{1}{\omega_{q,s}} f_m^{*\,\frac{q-s}{s(q-1)}}$$

$$= 2\gamma \frac{s(q-1)}{q-s} \left(\frac{s-1}{s} \right)^{\frac{s-1}{q-1}} \left(\frac{9 M_q D_s^s D^{q-s}}{V_f} \right)^{\frac{1}{q-1}},$$

$$\frac{\gamma}{\omega_{q,s}(1-\hat{\gamma})} \varepsilon^{1-\hat{\gamma}} = 2\gamma \frac{s(q-1)}{q-s} \left(\frac{s-1}{s} \right)^{\frac{q(s-1)}{(q-1)s}} \left(\frac{9 M_q D_s^q}{V_f^{\frac{q}{s}}} \right)^{\frac{1}{q-1}} \varepsilon^{\frac{q-s}{s(q-1)}},$$

and plugging them back in, we conclude

$$
\begin{aligned}
k &\le C_q \frac{1}{f_m^{*\frac{1}{q-1}}} + \frac{\gamma}{\omega_{q,s}(1-\hat{\gamma})}\left[f_m^{*(1-\hat{\gamma})} - \varepsilon^{1-\hat{\gamma}}\right] + 2\ln\frac{g_0}{\delta} \\
&= 2\gamma(q-1)\frac{q}{q-s}\left(\frac{s-1}{s}\right)^{\frac{s-1}{q-1}}\left(\frac{9M_q D_s^s D^{q-s}}{V_f}\right)^{\frac{1}{q-1}} - \frac{\gamma}{\omega_{q,s}(1-\hat{\gamma})}\varepsilon^{1-\hat{\gamma}} + 2\ln\frac{g_0}{\delta} \\
&= 2\gamma\frac{q}{s}\left(\frac{s-1}{s}\right)^{\frac{s-1}{q-1}}\left(\frac{9M_q D_s^s D^{q-s}}{V_f}\right)^{\frac{1}{q-1}}\frac{s(q-1)}{q-s}\times \\
&\qquad \times\left[1 - \frac{s}{q}\left(\left(\frac{s}{s-1}\right)^{s-1}\frac{V_f D^s}{D_s^s}\right)^{\frac{q-s}{s(q-1)}}\varepsilon^{\frac{q-s}{s(q-1)}}\right] + 2\ln\frac{g_0}{\delta}.
\end{aligned}
$$

□

H Proofs

H.1 Proof of Lemma 2

Proof of Lemma 2. Using weighed AG inequality, for $0 \le b \le p$, we have

$$
x^b \le \frac{(p-b) + bx^p}{p}. \tag{103}
$$

We use this inequality for each term of the polynomial. □

H.2 Proof of Proposition 1

Proof of Proposition 1. We can derive all of the inequalities straightforwardly

$$
\begin{aligned}
\nabla f(y) - \nabla f(x) - \nabla^2 f(x)[y-x] &= \int_0^1 \left(\nabla^2 f(x - \tau(y-x)) - \nabla^2 f(x)\right)[y-x]d\tau \\
\left\|\nabla f(y) - \nabla f(x) - \nabla^2 f(x)[y-x]\right\|_x^* &\le \int_0^1 \left\|\nabla^2 f(x - \tau(y-x)) - \nabla^2 f(x)\right\|_{op}\|y-x\|_x d\tau \\
&\le L_{2,\nu}\|y-x\|_x^{1+\nu}\int_0^1 \tau^\nu d\tau \\
&= \frac{L_{2,\nu}}{1+\nu}\|y-x\|_x^{1+\nu},
\end{aligned}
$$

$$
\begin{aligned}
\nabla^2 f(y) - \nabla^2 f(x) - \nabla^3 f(x)[y-x] &= \int_0^1 \left(\nabla^3 f(x - \tau(y-x)) - \nabla^3 f(x)\right)[y-x]d\tau \\
\left\|\nabla^2 f(y) - \nabla^2 f(x) - \nabla^3 f(x)[y-x]\right\|_{op} &\le \int_0^1 \left\|\nabla^3 f(x - \tau(y-x)) - \nabla^3 f(x)\right\|_{op}\|y-x\|_x d\tau \\
&\le L_{3,\nu}\|y-x\|_x^{1+\nu}\int_0^1 \tau^\nu d\tau \\
&= \frac{L_{3,\nu}}{1+\nu}\|y-x\|_x^{1+\nu},
\end{aligned}
$$

$$\nabla f(y)-\nabla f(x)-\nabla^2 f(x)\,[y-x]-\frac{1}{2}\nabla^3 f(x)[y-x]^2=\int_0^1\int_0^\tau\left(\nabla^3 f(x+\sigma(y-x))-\nabla^3 f(x)\right)[y-x]^2 d\sigma d\tau$$

$$\begin{aligned}\left\|\nabla f(y)-\nabla f(x)-\nabla^2 f(x)\,[y-x]-\frac{1}{2}\nabla^3 f(x)[y-x]^2\right\|_x^* &\le \int_0^1\int_0^\tau \|\nabla^3 f(x+\sigma(y-x))-\nabla^3 f(x)\|_x^*\|y-x\|_x^2 d\sigma d\tau\\ &\le L_{3,\nu}\|y-x\|_x^{2+\nu}\int_0^1\int_0^\tau \sigma^\nu d\sigma d\tau\\ &= \frac{L_{3,\nu}}{(1+\nu)(2+\nu)}\|y-x\|_x^{2+\nu}.\end{aligned}$$

□

H.3 Proof of Lemma 1

Proof of Lemma 1. For any $x,h,y\in\mathbb{E}$ and taking $y=x+\tau u$ for $\tau>0$, $\|u\|_x=1$

$$\begin{aligned}0\le\|h\|_y^2 &\le \|h\|_x^2+\left\langle\nabla^3 f(x)[h]^2,y-x\right\rangle+\frac{L_{3,\nu}}{1+\nu}\|y-x\|_x^{1+\nu}\|h\|_x^2\\ 0&\le\frac{1}{\tau}\|h\|_x^2+\left\langle\nabla^3 f(x)[h]^2,u\right\rangle+\frac{L_{3,\nu}\tau^\nu}{1+\nu}\|h\|_x^2\\ \|\nabla^3 f(x)[h]^2\|_x^*&\le\left(\frac{1}{\tau}+\frac{L_{3,\nu}\tau^\nu}{1+\nu}\right)\|h\|_x^2\end{aligned}$$

Setting

$$\tau=\left(\frac{1+\nu}{L_{3,\nu}}\right)^{\frac{1}{1+\nu}},$$

we get

$$\|\nabla^3 f(x)[h]^2\|_x^*\le 2\left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{1}{1+\nu}}\|h\|_x^2.$$

Setting $x^k=x, h=x^{k+1}-x^k$ we get

$$\|\nabla^3 f(x^k)[x^{k+1}-x^k]^2\|_{x^k}^*\le 2\left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{1}{1+\nu}}\|x^{k+1}-x^k\|_{x^k}^2=2\left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{1}{1+\nu}}\alpha_k^2\|\nabla f(x^k)\|_{x^k}^{*2}$$

□

H.4 Proof of Lemma 8

Proof. Proof of Lemma 8.

$$\begin{aligned}\|\nabla f(x^{k+1})\|_{x^k}^* &= \|\nabla f(x^{k+1})-\nabla^2 f(x^k)\left[x^{k+1}-x^k\right]-\alpha_k\nabla f(x^k)\|_{x^k}^*\\ &=\|\nabla f(x^{k+1})-\nabla f(x^k)-\nabla^2 f(x^k)\left[x^{k+1}-x^k\right]+(1-\alpha_k)\nabla f(x^k)\|_{x^k}^*\\ &\le\|\nabla f(x^{k+1})-\nabla f(x^k)-\nabla^2 f(x^k)\left[x^{k+1}-x^k\right]\|_{x^k}^*+(1-\alpha_k)\|\nabla f(x^k)\|_{x^k}^*\\ &\le\frac{L_{2,\nu}}{1+\nu}\|x^{k+1}-x^k\|_{x^k}^{1+\nu}+(1-\alpha_k)\|\nabla f(x^k)\|_{x^k}^* \qquad \text{(if } L_{2,\nu} \text{ exists)}\\ &=\frac{L_{2,\nu}}{1+\nu}\alpha_k^{1+\nu}\|\nabla f(x^k)\|_{x^k}^{*(1+\nu)}+(1-\alpha_k)\|\nabla f(x^k)\|_{x^k}^*\\ &=\left(1-\alpha_k+\frac{L_{2,\nu}}{1+\nu}\alpha_k^{1+\nu}\|\nabla f(x^k)\|_{x^k}^{*\nu}\right)\|\nabla f(x^k)\|_{x^k}^*\\ &=\left(\theta_k+\frac{L_{2,\nu}}{1+\nu}\alpha_k^{\nu}\|\nabla f(x^k)\|_{x^k}^{*\nu}\right)\alpha_k\|\nabla f(x^k)\|_{x^k}^*.\end{aligned}$$

Hence

$$\left\|\nabla f(x^{k+1})\right\|_{x^k}^* \leq \begin{cases} 2\frac{L_{2,\nu}}{1+\nu}\alpha_k^{1+\nu}\left\|\nabla f(x^k)\right\|_{x^k}^{*(1+\nu)} & \text{if } \theta_k \leq \frac{L_{2,\nu}}{1+\nu}\alpha_k^{\nu}\left\|\nabla f(x^k)\right\|_{x^k}^{*\nu} \\ 2\theta_k\alpha_k\left\|\nabla f(x^k)\right\|_{x^k}^* & \text{if } \theta_k \geq \frac{L_{2,\nu}}{1+\nu}\alpha_k^{\nu}\left\|\nabla f(x^k)\right\|_{x^k}^{*\nu} \end{cases}$$

□

H.5 Proof of Lemma 3

Proof of Lemma 3. We can rewrite the Hölder continuity for points x^k, x^{k+1} s.t. $x^{k+1} = x^k - \alpha_k \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k)$

$$\begin{aligned}
&\left(\frac{L_{2,\nu}}{1+\nu}\left(\alpha_k\left\|\nabla f(x^k)\right\|_{x^k}^*\right)^{1+\nu}\right)^2 \\
&= \left(\frac{L_{2,\nu}}{1+\nu}\left\|x^{k+1} - x^k\right\|_{x^k}^{1+\nu}\right)^2 \\
&\geq \left\|\nabla f(x^{k+1}) - \nabla f(x^k) - \nabla^2 f(x^k)\left[x^{k+1} - x^k\right]\right\|_{x^k}^{*2} \\
&= \left\|\nabla f(x^{k+1}) - \nabla f(x^k) + \alpha_k \nabla f(x^k)\right\|_{x^k}^{*2} \\
&= \left\|\nabla f(x^{k+1}) - (1-\alpha_k)\nabla f(x^k)\right\|_{x^k}^{*2} \\
&= \left\|\nabla f(x^{k+1})\right\|_{x^k}^{*2} + (1-\alpha_k)^2\left\|\nabla f(x^k)\right\|_{x^k}^{*2} - 2(1-\alpha_k)\left\langle \nabla f(x^{k+1}), \left[\nabla^2 f(x^k)\right]^{-1}\nabla f(x^k)\right\rangle.
\end{aligned}$$

We are going to set σ so that

$$\frac{1-\alpha_k}{2}\left\|\nabla f(x^k)\right\|_{x^k}^{*2} \geq \frac{1}{2(1-\alpha_k)}\left(\frac{L_{2,\nu}}{1+\nu}\left(\alpha_k\left\|\nabla f(x^k)\right\|_{x^k}^*\right)^{1+\nu}\right)^2, \tag{104}$$

and hence, we can conclude the proof by rearranging,

$$\begin{aligned}
&\left\langle \nabla f(x^{k+1}), \left[\nabla^2 f(x^k)\right]^{-1}\nabla f(x^k)\right\rangle \\
&\geq \frac{1}{2(1-\alpha_k)}\left\|\nabla f(x^{k+1})\right\|_{x^k}^{*2} + \frac{1-\alpha_k}{2}\left\|\nabla f(x^k)\right\|_{x^k}^{*2} - \frac{1}{2(1-\alpha_k)}\left(\frac{L_{2,\nu}}{1+\nu}\left(\alpha_k\left\|\nabla f(x^k)\right\|_{x^k}^*\right)^{1+\nu}\right)^2 \\
&\geq \frac{1}{2(1-\alpha_k)}\left\|\nabla f(x^{k+1})\right\|_{x^k}^{*2}.
\end{aligned}$$

Now we are going to choose σ to satisfy (104). Because α_k is a root of a polynomial P, we have

$$1 - \alpha_k - \alpha_k^{1+\beta}\lambda_k = 0,$$

so the equation (104) is equivalent to

$$\begin{aligned}
1 - \alpha_k = \alpha_k^{1+\beta}\lambda_k &\geq \frac{L_{2,\nu}}{1+\nu}\alpha_k^{1+\nu}\left\|\nabla f(x^k)\right\|_{x^k}^{*\nu}, \\
\theta_k &\geq \frac{L_{2,\nu}}{1+\nu}\alpha_k^{\nu}\left\|\nabla f(x^k)\right\|_{x^k}^{*\nu}.
\end{aligned}$$

□

H.6 Proof of Lemma 4

Proof of Lemma 4. We can rewrite the Hölder continuity for points x^k, x^{k+1} s.t. $x^{k+1} = x^k - \alpha_k \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k)$

$$\frac{L_{3,\nu}}{(1+\nu)(2+\nu)} \left(\alpha_k \left\|\nabla f(x^k)\right\|_{x^k}^*\right)^{2+\nu} \tag{105}$$
$$= \frac{L_{3,\nu}}{(1+\nu)(2+\nu)} \left\|x^{k+1} - x^k\right\|_{x^k}^{2+\nu} \tag{106}$$
$$\geq \left\|\nabla f(x^{k+1}) - \nabla f(x^k) - \nabla^2 f(x^k)[x^{k+1} - x^k] - \frac{1}{2}\nabla^3 f(x^k)[x^{k+1} - x^k]^2\right\|_{x^k}^* \tag{107}$$
$$= \left\|\nabla f(x^{k+1}) - (1-\alpha_k)\nabla f(x^k) - \frac{1}{2}\nabla^3 f(x^k)[x^{k+1} - x^k]^2\right\|_{x^k}^*. \tag{108}$$

Squaring

$$\begin{aligned}
&\left(\frac{L_{3,\nu}}{(1+\nu)(1+\nu)} \left(\alpha_k \left\|\nabla f(x^k)\right\|_{x^k}^*\right)^{2+\nu}\right)^2 \\
&\geq \left\|\nabla f(x^{k+1}) - (1-\alpha_k)\nabla f(x^k) - \frac{1}{2}\nabla^3 f(x^k)[x^{k+1} - x^k]^2\right\|_{x^k}^{*2} \\
&= \left\|\nabla f(x^{k+1})\right\|_{x^k}^{*2} + (1-\alpha_k)^2 \left\|\nabla f(x^k)\right\|_{x^k}^{*2} + \frac{1}{4}\left\|\nabla^3 f(x^k)[x^{k+1} - x^k]^2\right\|_{x^k}^{*2} \\
&\quad -2(1-\alpha_k)\left\langle \nabla f(x^{k+1}), \left[\nabla^2 f(x^k)\right]^{-1} \nabla f(x^k)\right\rangle \\
&\quad +(1-\alpha_k)\left\langle \left[\nabla^2 f(x^k)\right]^{-\frac{1}{2}} \nabla f(x^k), \left[\nabla^2 f(x^k)\right]^{-\frac{1}{2}} \nabla^3 f(x^k)[x^{k+1} - x^k]^2\right\rangle \\
&\quad -\left\langle \left[\nabla^2 f(x^k)\right]^{-\frac{1}{2}} \nabla f(x^{k+1}), \left[\nabla^2 f(x^k)\right]^{-\frac{1}{2}} \nabla^3 f(x^k)[x^{k+1} - x^k]^2\right\rangle \\
&\geq \frac{1}{2}\left\|\nabla f(x^{k+1})\right\|_{x^k}^{*2} + (1-\alpha_k)^2 \left\|\nabla f(x^k)\right\|_{x^k}^{*2} - \frac{1}{4}\left\|\nabla^3 f(x^k)[x^{k+1} - x^k]^2\right\|_{x^k}^{*2} \\
&\quad -2(1-\alpha_k)\left\langle \nabla f(x^{k+1}), \left[\nabla^2 f(x^k)\right]^{-1} \nabla f(x^k)\right\rangle \\
&\quad -(1-\alpha_k)\left\|\nabla f(x^k)\right\|_{x^k}^{*}\left\|\nabla^3 f(x^k)[x^{k+1} - x^k]^2\right\|_{x^k} \\
&\geq \frac{1}{2}\left\|\nabla f(x^{k+1})\right\|_{x^k}^{*2} + (1-\alpha_k)^2 \left\|\nabla f(x^k)\right\|_{x^k}^{*2} - \left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{2}{1+\nu}} \alpha_k^4 \left\|\nabla f(x^k)\right\|_{x^k}^{4} \\
&\quad -2(1-\alpha_k)\left\langle \nabla f(x^{k+1}), \left[\nabla^2 f(x^k)\right]^{-1} \nabla f(x^k)\right\rangle \\
&\quad -2\left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{1}{1+\nu}} \alpha_k^2 (1-\alpha_k)\left\|\nabla f(x^k)\right\|_{x^k}^{*3}.
\end{aligned}$$

Rearranging yields

$$\begin{aligned}
&\left\langle \nabla f(x^{k+1}), \left[\nabla^2 f(x^k)\right]^{-1} \nabla f(x^k)\right\rangle \\
&\geq \frac{1}{4(1-\alpha_k)}\left\|\nabla f(x^{k+1})\right\|_{x^k}^{*2} + \frac{1-\alpha_k}{2}\left\|\nabla f(x^k)\right\|_{x^k}^{*2} - \frac{1}{2}\left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{2}{1+\nu}} \frac{\alpha_k^4}{1-\alpha_k}\left\|\nabla f(x^k)\right\|_{x^k}^{*4} \\
&\quad -\left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{1}{1+\nu}} \alpha_k^2 \left\|\nabla f(x^k)\right\|_{x^k}^{*3} - \frac{1}{2(1-\alpha_k)}\left(\frac{L_{3,\nu}}{(1+\nu)(2+\nu)}\right)^2 \left(\alpha_k \left\|\nabla f(x^k)\right\|_{x^k}^*\right)^{2(2+\nu)}.
\end{aligned}$$

Finally, we are going to set θ_k so that

$$\frac{1-\alpha_k}{6}\left\|\nabla f(x^k)\right\|_{x^k}^{*2} \geq \frac{1}{2}\left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{2}{1+\nu}} \frac{\alpha_k^4}{1-\alpha_k}\left\|\nabla f(x^k)\right\|_{x^k}^{*4} \tag{109}$$

$$\frac{1-\alpha_k}{6}\left\|\nabla f(x^k)\right\|_{x^k}^{*2} \geq \left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{1}{1+\nu}} \alpha_k^2\left\|\nabla f(x^k)\right\|_{x^k}^{*3} \tag{110}$$

$$\frac{1-\alpha_k}{6}\left\|\nabla f(x^k)\right\|_{x^k}^{*2} \geq \frac{1}{2(1-\alpha_k)}\left(\frac{L_{3,\nu}}{(1+\nu)(2+\nu)}\right)^2 \left(\alpha_k\left\|\nabla f(x^k)\right\|_{x^k}^{*}\right)^{2(2+\nu)} \tag{111}$$

and then we can conclude

$$\left\langle \nabla f(x^{k+1}), \left[\nabla^2 f(x^k)\right]^{-1}\nabla f(x^k)\right\rangle \geq \frac{1}{4(1-\alpha_k)}\left\|\nabla f(x^{k+1})\right\|_{x^k}^{*2}.$$

Note that the choice of stepsize implies

$$1-\alpha_k = \alpha_k^{1+\beta}\lambda_k$$

and (109), (110), (111) are satisfied as

$$1-\alpha_k = \alpha_k^{1+\beta}\lambda_k \geq \begin{cases} \sqrt{3}\left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{1}{1+\nu}} \alpha_k^2\left\|\nabla f(x^k)\right\|_{x^k}^{*} & \text{if } \theta_k \geq \sqrt{3}\left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{1}{1+\nu}}\alpha_k\left\|\nabla f(x^k)\right\|_{x^k}^{*} \\ 6\left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{1}{1+\nu}} \alpha_k^2\left\|\nabla f(x^k)\right\|_{x^k}^{*} & \text{if } \theta_k \geq 6\left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{1}{1+\nu}}\alpha_k\left\|\nabla f(x^k)\right\|_{x^k}^{*} \\ \frac{\sqrt{3}L_{3,\nu}}{(1+\nu)(1+\nu)}\alpha_k^{2+\nu}\left\|\nabla f(x^k)\right\|_{x^k}^{*(1+\nu)} & \text{if } \theta_k \geq \frac{\sqrt{3}L_{3,\nu}}{(1+\nu)(2+\nu)}\alpha_k^{1+\nu}\left\|\nabla f(x^k)\right\|_{x^k}^{*(1+\nu)}. \end{cases}$$

We can ensure (109), (110), (111) by

$$\theta_k \geq \alpha_k\left\|\nabla f(x^k)\right\|_{x^k}^{*} \max\left\{6\left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{1}{1+\nu}}, \frac{\sqrt{3}L_{3,\nu}}{(1+\nu)(2+\nu)}\alpha_k^{\nu}\left\|\nabla f(x^k)\right\|_{x^k}^{*\nu}\right\}.$$

□

H.7 Towards the proof of Theorem 2

We unify cases $p = 2, 3$ with the Lemma 5.

Corollary 3. *Lemma 5 with $\gamma = \nu$ implies that choice $\theta_k = \left(\frac{L_{2,\nu}}{1+\nu}\right)^{\frac{1}{1+\nu}}\left\|\nabla f(x^k)\right\|_{x^k}^{*\frac{\nu}{1+\nu}}$ satisfies (18), and therefore Lemma 3 implies decrease as Doikov et al. [2024],*

$$f(x^k) - f(x^{k+1}) \geq \frac{1}{\theta_k}\left\|\nabla f(x^{k+1})\right\|_{x^k}^{*2} \geq \left(\frac{1+\nu}{L_{2,\nu}}\right)^{\frac{1}{1+\nu}} \frac{\left\|\nabla f(x^{k+1})\right\|_{x^k}^{*2}}{\left\|\nabla f(x^k)\right\|_{x^k}^{*\frac{\nu}{1+\nu}}}. \tag{112}$$

Lemma 5 with $\gamma \in \{1, 1+\nu\}$ implies that the choice

$$\theta_k \geq \left\|\nabla f(x^k)\right\|_{x^k}^{*\frac{1}{2}} \max\left\{\left(\frac{6^{1+\nu}L_{3,\nu}}{1+\nu}\right)^{\frac{1}{2(1+\nu)}}, \left(\frac{\sqrt{3}L_{3,\nu}}{(1+\nu)(2+\nu)}\right)^{\frac{1}{2+\nu}}\left\|\nabla f(x^k)\right\|_{x^k}^{*\frac{\nu}{2(2+\nu)}}\right\}, \tag{113}$$

satisfies (19), and therefore Lemma 4 implies decrease

$$f(x^k) - f(x^{k+1}) \geq \frac{1}{2\theta_k}\left\|\nabla f(x^{k+1})\right\|_{x^k}^{*2} \tag{114}$$

$$\geq \frac{1}{\max\left\{\left(\frac{6^{1+\nu}L_{3,\nu}}{1+\nu}\right)^{\frac{1}{2(1+\nu)}}, \left(\frac{\sqrt{3}L_{3,\nu}}{(1+\nu)(2+\nu)}\right)^{\frac{1}{2+\nu}}\left\|\nabla f(x^k)\right\|_{x^k}^{*\frac{\nu}{2(2+\nu)}}\right\}} \frac{\left\|\nabla f(x^{k+1})\right\|_{x^k}^{*2}}{\left\|\nabla f(x^k)\right\|_{x^k}^{*\frac{1}{2}}}. \tag{115}$$

On the other hand, choice of $\theta_k = \left(\frac{6^{1+\nu} L_{3,\nu}}{1+\nu}\right)^{\frac{1}{2+\nu}} \left\|\nabla f(x^k)\right\|_{x^k}^{*\frac{1+\nu}{2+\nu}}$ *in Lemma 4 implies decrease as Doikov et al. [2024],*

$$f(x^k) - f(x^{k+1}) \geq \frac{1}{2\theta_k}\left\|\nabla f(x^{k+1})\right\|_{x^k}^{*2} \geq \frac{1}{2}\left(\frac{1+\nu}{6^{1+\nu} L_{3,\nu}}\right)^{\frac{1}{2+\nu}} \frac{\left\|\nabla f(x^{k+1})\right\|_{x^k}^{*2}}{\left\|\nabla f(x^k)\right\|_{x^k}^{*\frac{1+\nu}{2+\nu}}}. \tag{116}$$

H.7.1 Proof of Theorem 2

We can combine previous corollaries.

Proof of Theorem 2. For $p = 2$, choice $\theta_k = \left(\frac{L_{p,\nu}}{p-1+\nu}\right)^{\frac{1}{p-1+\nu}} \left\|\nabla f(x^k)\right\|_{x^k}^{*\frac{p-2+\nu}{p-1+\nu}}$ implies

$$f(x^k) - f(x^{k+1}) \geq \left(\frac{p-1+\nu}{L_{p,\nu}}\right)^{\frac{1}{p-1+\nu}} \frac{\left\|\nabla f(x^{k+1})\right\|_{x^k}^{*2}}{\left\|\nabla f(x^k)\right\|_{x^k}^{*\frac{p-2+\nu}{p-1+\nu}}}. \tag{117}$$

For $p = 3$, choice $\theta_k = 6\left(\frac{L_{p,\nu}}{3(p-1+\nu)}\right)^{\frac{1}{p-1+\nu}} \left\|\nabla f(x^k)\right\|_{x^k}^{*\frac{p-2+\nu}{p-1+\nu}}$ implies

$$f(x^k) - f(x^{k+1}) \geq \frac{1}{12}\left(\frac{3(p-1+\nu)}{L_{p,\nu}}\right)^{\frac{1}{p-1+\nu}} \frac{\left\|\nabla f(x^{k+1})\right\|_{x^k}^{*2}}{\left\|\nabla f(x^k)\right\|_{x^k}^{*\frac{p-2+\nu}{p-1+\nu}}}. \tag{118}$$

And for any $p \in \{2, 3\}$ we have that $\theta_k = 6\left(\frac{L_{p,\nu}}{3(p-1+\nu)}\right)^{\frac{1}{p-1+\nu}} \left\|\nabla f(x^k)\right\|_{x^k}^{*\frac{p-2+\nu}{p-1+\nu}}$ implies

$$f(x^k) - f(x^{k+1}) \geq \frac{1}{12}\left(\frac{3(p-1+\nu)}{L_{p,\nu}}\right)^{\frac{1}{p-1+\nu}} \frac{\left\|\nabla f(x^{k+1})\right\|_{x^k}^{*2}}{\left\|\nabla f(x^k)\right\|_{x^k}^{*\frac{p-2+\nu}{p-1+\nu}}}. \tag{119}$$

□

H.8 Proof of Lemma 5

Proof of Lemma 5. Consider any $c_2, \delta > 0$. Inequality $\theta_k \geq c_2^{\frac{1}{1+\delta}}$ implies

$$\frac{1}{\theta_k{}^{\delta}} c_2 \geq c_2 \alpha_k^{\delta},$$

which is ensured by

$$\theta_k \geq \frac{1}{\theta_k{}^{\delta}} c_2,$$

or equivalently

$$\theta_k \geq c_2^{\frac{1}{1+\delta}}.$$

Now, choice $c_2 = c_3\left\|\nabla f(x^k)\right\|_{x^k}^{*\delta}$ guarantees that $\theta_k \geq c_3^{\frac{1}{1+\delta}} \left\|\nabla f(x^k)\right\|_{x^k}^{*\frac{\delta}{1+\delta}}$ ensures $\theta_k \geq c_3\left(\alpha_k\left\|\nabla f(x^k)\right\|_{x^k}^{*}\right)^{\delta}$. □

H.9 Proof of Corollary 3

Proof of Corollary 3. For the first part of (19), we use $\alpha_k, \nu \in [0, 1]$ to bound $\frac{1}{\theta_k{}^{\frac{1}{1+\nu}}} \geq \alpha_k^{\frac{1}{1+\nu}} \geq \alpha_k$ and

$$\frac{1}{\theta_k{}^{\frac{1}{1+\nu}}} 6\left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{1}{1+\nu}} \left\|\nabla f(x^k)\right\|_{x^k}^{*} \geq 6\left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{1}{1+\nu}} \alpha_k\left\|\nabla f(x^k)\right\|_{x^k}^{*}.$$

Now, the first part of (19) is ensured by θ_k so that

$$\theta_k \geq \frac{1}{\theta_k^{\frac{1}{1+\nu}}} 6 \left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{1}{1+\nu}} \left\|\nabla f(x^k)\right\|_{x^k}^*,$$

or equivalently

$$\theta_k \geq \left(\frac{6^{1+\nu} L_{3,\nu}}{1+\nu}\right)^{\frac{1}{2+\nu}} \left\|\nabla f(x^k)\right\|_{x^k}^{*\frac{1+\nu}{2+\nu}}.$$

We ensure the second part of (19) directly using Lemma 5 and together with first part we have

$$\begin{aligned}
\theta_k &\geq \max\left\{\left(\frac{6^{1+\nu} L_{3,\nu}}{1+\nu}\right)^{\frac{1}{2+\nu}} \left\|\nabla f(x^k)\right\|_{x^k}^{*\frac{1+\nu}{2+\nu}}, \left(\frac{\sqrt{3} L_{3,\nu}}{(1+\nu)(2+\nu)}\right)^{\frac{1}{2+\nu}} \left\|\nabla f(x^k)\right\|_{x^k}^{*\frac{1+\nu}{2+\nu}}\right\} \\
&= \left(\frac{L_{3,\nu}}{1+\nu}\right)^{\frac{1}{2+\nu}} \left\|\nabla f(x^k)\right\|_{x^k}^{*\frac{1+\nu}{2+\nu}} \max\left\{6^{\frac{1+\nu}{2+\nu}}, \left(\frac{\sqrt{3}}{2+\nu}\right)^{\frac{1}{2+\nu}}\right\} \\
&= \left(\frac{6^{1+\nu} L_{3,\nu}}{1+\nu}\right)^{\frac{1}{2+\nu}} \left\|\nabla f(x^k)\right\|_{x^k}^{*\frac{1+\nu}{2+\nu}}.
\end{aligned}$$

□

H.10 Proof of Lemma 6

Proof of Lemma 6. For $0 \leq \beta \leq 1$, function $y(x) = x^\beta, x \geq 0$ is concave, which implies

$$a^\beta - b^\beta \geq \frac{\beta}{a^{1-\beta}}(a-b), \quad \forall a > b \geq 0, \tag{120}$$

which we will be using for $\beta \overset{\text{def}}{=} \frac{1}{q-1} = (0,1]$. We rewrite functional value decrease as

$$\frac{1}{f_{k+1}^\beta} - \frac{1}{f_k^\beta} = \frac{f_k^\beta - f_{k+1}^\beta}{f_k^\beta f_{k+1}^\beta} \overset{(120)}{\geq} \frac{\beta(f_k - f_{k+1})}{f_k f_{k+1}^\beta} \overset{(22)}{\geq} \beta c_5 \frac{\left\|\nabla f(x^{k+1})\right\|_{x^k}^{*2}}{\left\|\nabla f(x^k)\right\|_{x^k}^{*\frac{q-2}{q-1}}} \frac{1}{f_k f_{k+1}^{\frac{1}{q-1}}} \tag{121}$$

$$\geq \beta c_5 \frac{\left\|\nabla f(x^{k+1})\right\|_{x^k}^{*2}}{\left\|\nabla f(x^k)\right\|_{x^k}^{*\left(2-\frac{q}{q-1}\right)}} \frac{1}{f_k^{\frac{q}{q-1}}} \geq \frac{\beta c_5}{D^{1+\beta}} \frac{\left\|\nabla f(x^{k+1})\right\|_{x^k}^{*2}}{\left\|\nabla f(x^k)\right\|_{x^k}^{*2}}, \tag{122}$$

where in the last step we used the convexity of f in the form $f_k \leq D\left\|\nabla f(x^k)\right\|_{x^k}^*$. We can continue by summing it for $k = 0, \ldots, n-1$,

$$
\begin{aligned}
\frac{1}{f_n^{\beta}} - \frac{1}{f_0^{\beta}} &\geq \frac{\beta c_5}{D^{1+\beta}} \sum_{k=0}^{n-1} \frac{\left\|\nabla f(x^{k+1})\right\|_{x^k}^{*2}}{\|\nabla f(x^k)\|_{x^k}^{*2}} && (123)\\
&\overset{AG}{\geq} \frac{\beta c_5 n}{D^{1+\beta}} \left(\prod_{k=0}^{n-1} \frac{\left\|\nabla f(x^{k+1})\right\|_{x^k}^{*2}}{\|\nabla f(x^k)\|_{x^k}^{*2}} \right)^{\frac{1}{n}} && (124)\\
&= \frac{\beta c_5 n}{D^{1+\beta}} \left(\prod_{k=1}^{n-1} \frac{\|\nabla f(x^k)\|_{x^{k-1}}^{*2}}{\|\nabla f(x^k)\|_{x^k}^{*2}} \right)^{\frac{1}{n}} \left(\frac{\|\nabla f(x^n)\|_{x^{n-1}}^{*}}{\|\nabla f(x^0)\|_{x^0}^{*}} \right)^{\frac{2}{n}} && (125)\\
&\geq \frac{\beta c_5 n}{\gamma D^{1+\beta}} \left(\frac{f_n}{\|\nabla f(x^0)\|_{x^0}^{*} D} \right)^{\frac{2}{n}} && (126)\\
&= \frac{\beta c_5 n}{\gamma D^{1+\beta}} \exp\left(-\frac{2}{n} \ln\left(\frac{\left\|\nabla f(x^0)\right\|_{x^0}^{*} D}{f_n} \right) \right) && (127)\\
&\geq \frac{\beta c_5 n}{\gamma D^{1+\beta}} \left(1 - \frac{2}{n} \ln\left(\frac{\left\|\nabla f(x^0)\right\|_{x^0}^{*} D}{f_n} \right) \right) && (128)
\end{aligned}
$$

We can bound f_n based on the size of $\frac{2}{n} \frac{\left\|\nabla f(x^0)\right\|_{x^0}^{*} D}{f_n}$.

1. If $\frac{2}{n} \ln\left(\frac{\left\|\nabla f(x^0)\right\|_{x^0}^{*} D}{f_n} \right) \geq \frac{1}{2}$, then $f_n \leq \left\|\nabla f(x^0)\right\|_{x^0}^{*} D \exp\left(-\frac{k}{4}\right)$.

2. If $\frac{2}{n} \ln\left(\frac{\left\|\nabla f(x^0)\right\|_{x^0}^{*} D}{f_n} \right) < \frac{1}{2}$, then

$$
\frac{1}{f_n^{\beta}} > \frac{1}{f_n^{\beta}} - \frac{1}{f_0^{\beta}} \geq \frac{\beta c_5 n}{2\gamma D^{1+\beta}} \Leftrightarrow f_n < \left(\frac{2\gamma D^{1+\beta}}{\beta c_5 n} \right)^{\frac{1}{\beta}} = \frac{D^q \left(2\gamma(q-1)\right)^{q-1}}{c_5{}^{q-1} n^{q-1}} \tag{129}
$$

Hence

$$
f_n \leq \frac{D^q \left(2\gamma(q-1)\right)^{q-1}}{c_5{}^{q-1} n^{q-1}} + \left\|\nabla f(x^0)\right\|_{x^0}^{*} D \exp\left(-\frac{k}{4}\right). \tag{130}
$$

□

H.11 Proof of Theorem 3

Proof of Theorem 3. Bounded Hessian change together with condition (21) in Theorem 2 imply inequalities

$$
\begin{aligned}
&\left\|\nabla f(x^{k+1})\right\|_{x^k}^{*} \left\|\nabla f(x^k)\right\|_{x^k}^{*} \geq \left\langle \nabla f(x^{k+1}), \left[\nabla^2 f(x^k)\right]^{-1} \nabla f(x^k) \right\rangle \geq \frac{1}{2\alpha_k \theta_k} \left\|\nabla f(x^{k+1})\right\|_{x^k}^{*2},\\
&\left\|\nabla f(x^k)\right\|_{x^k}^{*} \geq \frac{1}{2\alpha_k \theta_k} \left\|\nabla f(x^{k+1})\right\|_{x^k}^{*} \geq \frac{\gamma}{2\alpha_k \theta_k} \left\|\nabla f(x^{k+1})\right\|_{x^{k+1}}^{*} \qquad \left(\geq \frac{\gamma}{2} \left\|\nabla f(x^{k+1})\right\|_{x^{k+1}}^{*} \right),
\end{aligned} \tag{131}
$$

which for θ_k from (20) guarantees local superlinear rate for $q > 2$. □

H.12 Proof of Theorem 4

Proof of Theorem 4. Theorem 2 implies that Algorithm 1 satisfies requirements of Lemma 6 with correspondent q and $c_5 = \frac{1}{2} \left(\frac{1}{9M_q} \right)^{\frac{1}{q-1}}$. The convergence rate follows. □

H.13 Proof of Lemma 7

Proof of Lemma 7. We will prove the statement by induction. The base for σ_0 holds. For k-th iteration, consider 2 cases based on the number of iterations of the inner loop.

1. Algorithm continues after $j_k > 0$ inner iterations. Note that if θ_{k,j_k-1} satisfied (20), Theorem 2 guarantees the continuation condition to be satisfied for $j_k - 1$. Consequently, θ_{k,j_k-1} does not satisfy (20) for any $q \in [2,4]$, and hence

$$\sigma_{k+1} = \frac{\theta_{k,j_k-1}}{\|\nabla f(x^k)\|_{x^k}^{*\beta}} < \inf_{q\in[2,4]} (9M_q)^{\frac{1}{q-1}} \left\|\nabla f(x^k)\right\|_{x^k}^{*\frac{q-2}{q-1}-\beta} = \mathcal{H}\left(x^k\right). \tag{132}$$

2. Algorithm continues after $j = 0$ iterates, then from (131) we have

$$\sigma_{k+1} = \frac{\sigma_k}{\gamma} \le \frac{1}{\gamma}\mathcal{H}\left(x^{k-1}\right) \le \gamma^{\frac{q-2}{q-1}-1}\mathcal{H}\left(x^k\right) \le \mathcal{H}\left(x^k\right). \tag{133}$$

For the total number of oracle calls N_K,

$$N_K = \sum_{k=0}^{K-1}(1+j_k) = K + \sum_{k=0}^{K-1}\log_c\frac{c\sigma_{k+1}}{\sigma_k} = 2K + \log_c\frac{\sigma_K}{\sigma_0} \tag{134}$$

$$\le 2K + \log_c\frac{\mathcal{H}\left(\left\|x^{k-1}\right\|_{x^{k-1}}^*\right)}{\sigma_0}. \tag{135}$$

□

H.14 Proof of Theorem 5

Proof of Theorem 5. Algorithm 2 sets $x^{k+1} = x^k_{j_k}$ so that

$$\left\langle \nabla f(x^k_{j_{k-1}}), n^k\right\rangle < \frac{1}{2\alpha_{k,j_{k-1}}\theta_{k,j_{k-1}}}\left\|\nabla f(x^k_{j_{k-1}})\right\|_{x^k}^{*2}, \tag{136}$$

$$\left\langle \nabla f(x^k_{j_k}), n^k\right\rangle \ge \frac{1}{2\alpha_{k,j_k}\theta_{k,j_k}}\left\|\nabla f(x^k_{j_k})\right\|_{x^k}^{*2}. \tag{137}$$

From Theorem 2 we can see that while $\theta_{k,j_{k-1}} = \theta_{k,j_k}/\gamma$ does not satisfy (21) for any $q \in [2,4]$ and θ_{k,j_k} satisfies (20) for some q, therefore

$$\theta_{k,j_k} \ge (9M_q)^{\frac{1}{q-1}}\left\|\nabla f(x^k)\right\|_{x^k}^{*\frac{q-2}{q-1}} \qquad \exists q\in[2,4] \tag{138}$$

$$\theta_{k,j_k} < \gamma\,(9M_q)^{\frac{1}{q-1}}\left\|\nabla f(x^k)\right\|_{x^k}^{*\frac{q-2}{q-1}} \qquad \forall q\in[2,4] \tag{139}$$

$$\theta_{k,j_k} < \gamma \inf_{q\in[2,4]}(9M_q)^{\frac{1}{q-1}}\left\|\nabla f(x^k)\right\|_{x^k}^{*\frac{q-2}{q-1}}, \tag{140}$$

hence estimate θ_{k,j_k} is at most constant γ times worse than any plausible parametrization of (q, M_q), and therefore, even the best plausible parametrization. In particular, for

$$q^* \overset{\text{def}}{=} \operatorname*{argmin}_{q\in[2,4]} \frac{9M_qD^q\left(4\gamma^2(q-1)\right)^{q-1}}{k^{q-1}} + \left\|\nabla f(x^0)\right\|_{x^0}^* D\exp\left(-\frac{k}{4}\right), \tag{141}$$

we have that from Theorem 2

$$f(x^k) - f(x^{k+1}) \ge \frac{1}{2\gamma}\left(\frac{1}{9M_{q^*}}\right)^{\frac{1}{q^*-1}} \frac{\left\|\nabla f(x^{k+1})\right\|_{x^k}^{*2}}{\left\|\nabla f(x^k)\right\|_{x^k}^{*\frac{q^*-2}{q^*-1}}}. \tag{142}$$

The rest of the proof is analogous to the proof of Theorem 4. □