

Adaptive Newton-CG methods with global and local analysis for unconstrained optimization with Hölder continuous Hessian

Ziyang Zeng*

Junyu Zhang*

Chuan He[†]

April 1, 2026

Abstract

In this paper, we study Newton-conjugate gradient (Newton-CG) methods for minimizing a nonconvex function f whose Hessian is (H_f, ν) -Hölder continuous with modulus $H_f > 0$ and exponent $\nu \in (0, 1]$. Recently proposed Newton-CG methods for this problem [13] adopt (i) non-adaptive regularization and (ii) a nested line-search procedure, where (i) often leads to inefficient early progress and the loss of local superlinear convergence, and (ii) may incur high computational cost due to multiple solves of the Newton system per iteration. To address these limitations, we propose two novel Newton-CG algorithms, depending on the availability of ν , that adaptively regularize the Newton system by leveraging the auto-conditioning technique to eliminate the nested line search. The proposed algorithms achieve the best-known iteration complexity $\mathcal{O}(H_f^{1/(1+\nu)} \epsilon^{-(2+\nu)/(1+\nu)})$ for finding an ϵ -stationary point and, simultaneously, enjoy local superlinear convergence near nondegenerate local minimizers. Numerical experiments further demonstrate the practical advantages of our algorithms over existing approaches.

1 Introduction

We consider the nonconvex unconstrained optimization problem

$$\min_{x \in \mathbb{R}^n} f(x), \quad (1)$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is twice continuously differentiable. Our focus is on second-order algorithms that compute an ϵ -stationary point \bar{x} satisfying $\|\nabla f(\bar{x})\| \leq \epsilon$, under the assumption that the Hessian is (H_f, ν) -Hölder continuous on a suitable compact set \mathcal{X} that shall be specified later. That is, there exist constants $H_f > 0$ and $\nu \in (0, 1]$ such that $\|\nabla^2 f(y) - \nabla^2 f(x)\| \leq H_f \|y - x\|^\nu$ for all $x, y \in \mathcal{X}$. In the case $\nu = 1$, this condition corresponds to the standard Lipschitz-continuous Hessian assumption, which has been extensively studied in the literature. In contrast, the Hölder-continuous regime $\nu < 1$ has received comparatively less attention. In this work, we consider the full range $\nu \in (0, 1]$.

Under the Lipschitz-Hessian setting ($\nu = 1$), second-order methods are among the most powerful tools for solving (1) at small to medium-sized problems, due to their local superlinear convergence near

*Department of Industrial Systems Engineering and Management, National University of Singapore, Singapore (email: ziyangzeng@u.nus.edu, junyuz@nus.edu.sg). The work of Junyu Zhang was partially supported by the Singapore Ministry of Education Academic Research Fund Tier 2 (MOE-T2EP20125-0007).

[†]Department of Mathematics, Linköping University, Sweden (email: chuan.he@liu.se). The work of Chuan He was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. Corresponding author.

nondegenerate solutions. However, it is well known that the classic Newton method may fail to converge globally. To address this, a popular globalization strategy is the cubic-regularized Newton method [23], which not only enjoys global convergence with an optimal $\mathcal{O}(\epsilon^{-3/2})$ iteration complexity for nonconvex functions [6, 4], but also retains local superlinear convergence. Yet the need to repeatedly solve the cubic subproblems often constitutes a computational bottleneck in practical applications. To mitigate this cost, one may solve the subproblems inexactly with first-order methods [3]. As a more mature and practically stable globalization strategy, the Levenberg-Marquardt (quadratic) regularization [16, 20] is also widely used, with the resulting linear systems solved using the conjugate gradient (CG) method:

$$\left(\nabla^2 f(x^k) + \varepsilon_k I\right) d^k = -\nabla f(x^k), \quad (2)$$

where d^k is the search direction at iteration k . The regularization (damping) parameter ε_k plays a critical role in both theory and performance. For example, by setting $\varepsilon_k \equiv \sqrt{\epsilon}$, i.e., equal to the square root of the target accuracy ϵ , [25] proposes a capped CG procedure to solve (2) and obtain an $\mathcal{O}(\epsilon^{-3/2})$ iteration complexity for finding an ϵ -stationary point. While this strategy achieves the optimal global rates, its practical behavior can be unsatisfactory when the target precision ϵ is small. Specifically, if $\varepsilon_k \equiv \sqrt{\epsilon}$ is small, the method may repeatedly select negative curvature directions in early iterations, resulting in slow early progress. Moreover, using a non-adaptive ε_k can destroy the local superlinear convergence that one expects from second-order methods. To address this issue, [28] proposes a gradient-norm-based regularization rule that retains both fast global and local convergence. A more complete discussion of the literature can be found in later sections.

As for the Hölder-Hessian regime $\nu < 1$, results are still quite limited. Recently, He et al. [13] proposed a framework that sets $\varepsilon_k \equiv \epsilon^{\nu/(1+\nu)}$ and uses the capped CG to solve the corresponding linear systems. This method finds an ϵ -stationary point within $\mathcal{O}(\epsilon^{-(2+\nu)/(1+\nu)})$ iterations, matching the lower bound established in [6]. However, the same two limitations persist: (i) a too small and non-adaptive ε_k can lead to unnecessarily excessive reliance on negative curvature directions in early stages, producing weak descent sometimes even worse than gradient descent, and (ii) the desirable local superlinear convergence near nondegenerate local minimizers is missing. A remedy for this is to adapt the regularization parameter ε_k , at the cost of introducing a nested line-search procedure that is agnostic to the problem parameters. More precisely, one needs to solve (2) multiple times per iteration in the outer line-search loop for an appropriate ε_k , while another inner line-search loop is required for each ε_k to check whether a sufficient descent is obtained. This results in a computationally expensive nested double-loop line-search structure.

Contributions. Motivated by these considerations, our main contributions are highlighted below.

- We propose two Newton-CG methods, depending on the availability of ν , both of which adaptively regularize the Newton system using the current gradient magnitudes. To resolve the nested double line-search loop issue, we extend the auto-conditioning technique from first-order methods [15, 19], which approximates unknown problem parameters using historical local information, thereby eliminating the outer line-search loop for selecting ε_k . This allows our algorithms to solve (2) only once per iteration, improving practical efficiency.
- Both methods achieve the best-known iteration complexity $\mathcal{O}(H_f^{1/(1+\nu)} \epsilon^{-(2+\nu)/(1+\nu)})$. For both algorithms, we establish local superlinear convergence near nondegenerate stationary points, which is generally unattainable under non-adaptive damping schemes. A technically interesting point is that when developing a fully parameter-free variant, a direct application of existing auto-conditioning techniques does not work as the parameters controlled by the auto-conditioning

mechanism can become unbounded in the Hölderian regime ($\nu < 1$). Furthermore, a capture theorem is derived to assist the establishment of local superlinear rates.

Related literature. Next, let us review the closely related literature on second-order methods that has yet to be discussed. We begin with the representative choices of damping (regularization) parameters in Newton-type methods under the classical Lipschitz-continuous Hessian assumption. We then summarize the more limited body of work that develops second-order methods for (1) under Hölder continuity of the Hessian.

The Lipschitz-Hessian regime. In the classical Lipschitz-Hessian setting ($\nu = 1$), substantial work has investigated the choice of the damping parameter ε_k in the regularized Newton system. For convex functions, early works [18, 24] proposed gradient-based regularization of the form $\varepsilon_k \propto \|\nabla f(x^k)\|$, yielding quadratic local convergence but not providing satisfactory global complexity guarantees. More recently, Mishchenko [21] analyzed the choice $\varepsilon_k \propto \|\nabla f(x^k)\|^{1/2}$, establishing a global $\mathcal{O}(1/k^2)$ rate while successfully preserving local superlinear convergence. In the nonconvex setting, however, beyond the challenge of Hessian indefiniteness, a tension exists between achieving the optimal global rate and maintaining local superlinear convergence. Ueda and Yamashita [26] proposed a regularization scheme based on the minimum eigenvalue of the Hessian and the gradient norm. While preserving a local superlinear rate, it yields a suboptimal global complexity of $\mathcal{O}(\epsilon^{-2})$. To improve global performance, Gratton et al. [12] designed a trust-region-based method that alternates between regularized Newton and negative curvature steps, which improves the global rate to nearly optimal. Alternatively, Royer et al. [25] fixed ε_k to the target accuracy and employed the capped CG procedure, achieving the optimal global rate, while Curtis et al. [9] further improved this method by leveraging trust-region techniques. However, the non-adaptive regularization does not lead to local superlinear convergence. Along a similar line, Zhu and Xiao [29] achieved the optimal global rate together with local superlinear convergence using capped CG under additional error bound or global strong convexity assumptions. These results imply an inherent trade-off: gradient-norm-based regularization generally favors local behavior, as the regularization term vanishes asymptotically ($\|\nabla f(x^k)\| \rightarrow 0$), allowing the algorithm to behave like pure Newton steps near the solution. On the other hand, fixed-accuracy regularization, while globally optimal, often sacrifices this local rate. Very recently, Zhou et al. [28] bridged this gap by introducing novel gradient-norm-based regularization schemes that achieve both the best-known global rate and local superlinear convergence.

The Hölder-Hessian regime. Second-order methods for (1) under a Hölder-continuous Hessian have received relatively limited attention. The regularized Newton frameworks in [5, 7, 8, 11, 13, 27] constitute the main existing approaches in this setting. Specifically, the early work [5] address (1) by solving a sequence of $(2 + \nu)$ -th regularized subproblems, which can be viewed as a natural generalization of cubic regularization from the Lipschitz-Hessian case. This line was subsequently extended in [7, 8] to high-order regularized methods for nonconvex optimization under Hölder continuity of higher-order derivatives. In a related direction, [11] tackles (1) by solving a sequence of $(2 + \nu)$ -th or cubic regularized Newton subproblems, while [27] develops adaptive regularized Newton schemes on Riemannian manifolds ((1) is a special case) that inexactly solve either $(2 + \nu)$ -th order regularized Newton or trust-region subproblems. All these methods achieve optimal worst-case iteration complexity, but they typically require solving higher-order regularized subproblems or nontrivial polynomial optimization problems, which can be prohibitively expensive. Building on the quadratic regularized Newton and capped CG framework of [25], He et al. [13] proposed a parameter-free framework for

solving (1) with optimal global rate. However, its non-adaptive regularization scheme fails to preserve the appealing local superlinear convergence. In fact, to the best of our knowledge, no existing work that simultaneously attains optimal global rate and guarantees local superlinear rate for nonconvex problem with Hölder-continuous Hessian.

Organization. In Section 2, we introduce notation and standing assumptions used throughout the paper. In Sections 3 and 4, depending on the availability of the Hölder exponent ν , we develop two Newton-CG methods for (1) that adaptively regularize the Newton system while estimating the H_f in a line-search-free, auto-conditioned manner. Global iteration complexity and local superlinear convergence guarantees are established for both methods. Sections 5 and 6 present numerical results and conclusion, respectively. Appendix A contains the proofs of the main results. Appendix B briefly introduces the capped CG procedure used in our algorithms.

Notations. We use $\|\cdot\|$ to denote the ℓ_2 norm of a vector or the spectral norm of a matrix. We denote the level set of f at $u \in \mathbb{R}^n$ by $\mathcal{L}_f(u) := \{x : f(x) \leq f(u)\}$, and we denote its r -neighborhood by $\mathcal{L}_f(u; r) := \{x : \|x - x'\| \leq r, x' \in \mathcal{L}_f(u)\}$ for all $r \geq 0$. For any $z \in \mathbb{R}$, we denote the sign function as $\text{sgn}(z)$, which returns 1 when $z \geq 0$ and -1 when $z < 0$.

2 Preliminaries and basic assumptions

By default, we assume the objective function $f \in \mathcal{C}^2$ to be twice continuously differentiable and has bounded level set, as formalized below.

Assumption 1. *The level set $\mathcal{L}_f(x^0)$ at initial iterate x^0 is compact.*

As a direct consequence of Assumption 1, there exist $f_{\text{low}} \in \mathbb{R}$, $U_g > 0$ and $U_H > 0$ such that

$$f(x) \geq f_{\text{low}}, \quad \|\nabla f(x)\| \leq U_g, \quad \|\nabla^2 f(x)\| \leq U_H \quad \forall x \in \mathcal{L}_f(x^0). \quad (3)$$

Define $\Delta_f := f(x^0) - f_{\text{low}}$. Also, according to the algorithmic design of this paper, all analysis can be restricted to an r_d -neighborhood of the level set $\mathcal{L}_f(x^0)$, with r_d defined as

$$r_d := \max\{1.1, 1.1U_g, U_H\}. \quad (4)$$

The factor 1.1 here can be replaced by any constant greater than 1. We next make the assumption of Hölder continuity on $\nabla^2 f$ over the compact region $\mathcal{L}_f(x^0; r_d)$.

Assumption 2. *The Hessian $\nabla^2 f$ is (H_f, ν) -Hölder continuous on $\mathcal{L}_f(x^0; r_d)$ such that*

$$\|\nabla^2 f(y) - \nabla^2 f(x)\| \leq H_f \|y - x\|^\nu \quad \forall x, y \in \mathcal{L}_f(x^0; r_d),$$

for some $\nu \in (0, 1]$ and $H_f > 0$. Without loss of generality, we default $H_f \geq 1$.

Denote the Taylor residuals of f and ∇f , respectively, as

$$\begin{aligned} \mathcal{R}_0(y, x) &:= \left| f(y) - f(x) - \nabla f(x)^\top (y - x) - \frac{1}{2} (y - x)^\top \nabla^2 f(y, x) (y - x) \right|, \\ \mathcal{R}_1(y, x) &:= \|\nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y - x)\|. \end{aligned} \quad (5)$$

Then Assumption 2 immediately implies the following residual bounds:

$$\mathcal{R}_0(y, x) \leq \frac{H_f \|y - x\|^{2+\nu}}{(1+\nu)(2+\nu)}, \quad \mathcal{R}_1(y, x) \leq \frac{H_f \|y - x\|^{1+\nu}}{1+\nu} \quad \forall x, y \in \mathcal{L}_f(x^0; r_d), \quad (6)$$

see [11]. As a consequence, we can define two handy *lower* estimators of the constant H_f as

$$\mathcal{H}_0(y, x) := \frac{2\mathcal{R}_0(y, x)}{\|y - x\|^{2+\nu}}, \quad \mathcal{H}_1(y, x) := \frac{\mathcal{R}_1(y, x)}{\|y - x\|^{1+\nu}} \quad \forall x \neq y. \quad (7)$$

It is immediate that $H_f \geq \mathcal{H}_0(y, x)$ and $H_f \geq \mathcal{H}_1(y, x)$ for all $x, y \in \mathcal{L}_f(x^0; r_d)$ and $x \neq y$.

3 An adaptive regularized Newton-CG method

In this section, we propose an adaptive Newton-CG method for problem (1), under the basic setting where the Hölder exponent ν is known. This method incorporates the auto-conditioning technique used in developing parameter-free first-order methods (e.g., [15, 19]) to adaptively estimate the local Hölder constant H_f in $\mathcal{L}_f(x^0; r_d)$.

3.1 Algorithm framework

Based on the previous discussion, we present Algorithm 1. At each iteration $k \geq 0$ of this algorithm, we use a **CappedCG** subroutine (Algorithm 3, Appendix B) proposed by [25] to solve the damped Newton system (2):

$$\left(\nabla^2 f(x^k) + 2(\gamma_k \|\nabla f(x^k)\|^\nu)^{\frac{1}{1+\nu}} I \right) d = -\nabla f(x^k). \quad (8)$$

where γ_k is an adaptive estimation of H_f . Rather than approximating H_f through a backtracking search at each iteration, we employ the auto-conditioning mechanism of [15, 19]. It is worth mentioning that all the information for updating the local estimate σ_k is based solely on historical information of the algorithm that has already been computed in the previous steps, no extra Hessian/gradient/function evaluations are needed. Then, depending on whether **CappedCG** returns an approximate solution (SOL) of (8) or a negative curvature (NC) direction of $\nabla^2 f(x^k)$, Algorithm 1 will exploit the damped Newton or negative curvature directions, respectively.

Algorithm 1: An adaptive regularized Newton-CG method

```

1 Input: initial point  $x^0 \in \mathbb{R}^n$ , parameter  $\gamma_0 \geq 1$ , line-search parameters  $\eta, \theta \in (0, 1)$ , exponent  $\nu \in (0, 1]$ .
2 while  $\nabla f(x^k) \neq 0$  do
3   Set  $H_k = \nabla^2 f(x^k)$ ,  $g_k = \nabla f(x^k)$ ,  $\varepsilon_k = (\gamma_k \|g_k\|^\nu)^{1/(1+\nu)}$ , and  $\zeta_k = \min\{1/2, \|g_k\|^{\nu/(1+\nu)}\}$ . Then call
       $(d, \text{d\_type}) \leftarrow \text{CappedCG}(H_k, g_k, \varepsilon_k, \zeta_k, U)$ .
4   if  $\text{d\_type} == \text{NC}$  then
5     Set  $d^k \leftarrow -\text{sgn}(d^T g_k) \frac{|d^T H_k d|}{\|d\|^3} d$ , and find  $\alpha_k = \theta^{j_k}$ , where  $j_k$  is the smallest nonnegative integer  $j$  such
       that
       
$$f(x^k + \theta^j d^k) < f(x^k) - \frac{\eta}{2} \theta^{2j} \|d^k\|^3. \quad (9)$$

6     if  $j_k \geq 1$  then set  $\sigma_k = \mathcal{H}_0(x^k + \theta^{j_k-1} d^k, x^k)$ .
7   else //  $\text{d\_type} = \text{SOL}$ 
8     Set  $d^k \leftarrow d$ , and find  $\alpha_k = \theta^{j_k}$ , where  $j_k$  is the smallest nonnegative integer  $j$  such that
       
$$f(x^k + \theta^j d^k) < f(x^k) - \eta \varepsilon_k \theta^j \|d^k\|^2. \quad (10)$$

9     if  $j_k == 0$  then set  $\sigma_k = \mathcal{H}_1(x^k + d^k, x^k)$  else set  $\sigma_k = \max\{\mathcal{H}_0(x^k + d^k, x^k), \mathcal{H}_0(x^k + \theta^{j_k-1} d^k, x^k)\}$ .
10  Set  $x^{k+1} = x^k + \alpha_k d^k$ ,  $\gamma_{k+1} = \max\{\gamma_k, \sigma_k\}$ , and  $k \leftarrow k + 1$ .

```

As discussed in the introduction, a key distinction from the existing parameter-free Newton-CG methods [13] for problem (1) is that, instead of setting the damping parameter to $\varepsilon_k \propto \epsilon^{\nu/(1+\nu)}$, we set $\varepsilon_k \propto \|\nabla f(x^k)\|^{\nu/(1+\nu)}$. On the one hand, in early iterations when the gradients are still large, the algorithm is more likely to exploit the scaled gradient steps, suppressing the negative curvature steps that are usually slower in early stages. On the other hand, when the iterates get close to a nondegenerate local solution of the problem, the adaptively selected ε_k automatically diminishes as gradient decreases. By setting the relative accuracy of `CappedCG` to $\zeta_k = O(\|\nabla f(x^k)\|^{\nu/(1+\nu)})$, a local superlinear convergence can also be expected.

3.2 Global complexity bound

The following lemma shows that the main iterates and trial iterates of Algorithm 1 lie within a suitable neighborhood of the level set, with the proof given in Appendix A.2.

Lemma 1. *Given Assumption 1, the sequences $\{x^k\}_{k \geq 0}$ and $\{d^k\}_{k \geq 0}$ generated by Algorithm 1 satisfy $x^k + \alpha d^k \in \mathcal{L}_f(x^0; r_d)$ for all $k \geq 0$ and $\alpha \in [0, 1]$, where r_d is defined in (4).*

This lemma informs us that all analysis can be performed under the constants introduced in (3) and we can activate Assumption 2 to utilize the (H_f, ν) -Hölder continuity of $\nabla^2 f$ in $\mathcal{L}_f(x^0; r_d)$.

Next, we proceed with the global complexity analysis for finding ϵ -stationary points of problem (1). As the auto-conditioning mechanism always *underestimates* H_f , a sufficient decrease is not necessarily guaranteed even if `CappedCG` successfully returns an approximate solution to (8), a standard analysis may therefore fail. For Algorithm 1, we divide its iterations before reaching an ϵ -stationary point ($\mathbb{K}_\epsilon := \{k : \|\nabla f(x^t)\| > \epsilon, \forall t \leq k\}$) into three subsets:

$$\begin{aligned} \mathbb{K}_{\epsilon,1} &:= \{k \in \mathbb{K}_\epsilon : \|\nabla f(x^{k+1})\| \geq \|\nabla f(x^k)\|/2, \sigma_k \leq 2\gamma_k\}, \\ \mathbb{K}_{\epsilon,2} &:= \{k \in \mathbb{K}_\epsilon : \|\nabla f(x^{k+1})\| \geq \|\nabla f(x^k)\|/2, \sigma_k > 2\gamma_k\}, \\ \mathbb{K}_{\epsilon,3} &:= \{k \in \mathbb{K}_\epsilon : \|\nabla f(x^{k+1})\| < \|\nabla f(x^k)\|/2\}, \end{aligned}$$

where iterations in $\mathbb{K}_{\epsilon,1}$ generate sufficient descent, while $|\mathbb{K}_{\epsilon,2}|$ and $|\mathbb{K}_{\epsilon,3}|$ are provably small. In the following, we provide two lemmas that establish the sufficient descent for $\mathbb{K}_{\epsilon,1}$, depending on the output of `CappedCG`. The proofs of the two lemmas are deferred to Appendix A.2.

Lemma 2. *Given Assumptions 1 and 2, for all $k \in \mathbb{K}_{\epsilon,1}$ with d^k being the output of `CappedCG` with `d.type=SOL`, the following two statements hold.*

- (i) *The step size α_k is well-defined and it satisfies $\alpha_k \geq \min\{1, (2(1-\eta)/(1.1^\nu H_f))^{1/(1+\nu)}\theta\}$.*
- (ii) *Let $c_{\text{sol},\nu} := \eta(1-\eta)^{2/\nu}\theta/100$. The next iterate $x^{k+1} = x^k + \alpha_k d^k$ satisfies*

$$f(x^k) - f(x^{k+1}) \geq c_{\text{sol},\nu} \gamma_k^{-\frac{1}{1+\nu}} \|\nabla f(x^k)\|^{\frac{2+\nu}{1+\nu}}. \quad (11)$$

Moreover, the gradient at the next iterate is bounded by $\|\nabla f(x^{k+1})\| \leq (2H_f + 5)\|\nabla f(x^k)\|$.

Lemma 3. *Given Assumptions 1 and 2, for all $k \in \mathbb{K}_{\epsilon,1}$ with d^k being the output of `CappedCG` with `d.type=NC`, the following two statements hold.*

- (i) *The step size α_k is well-defined, and $\alpha_k \geq \min\{1, \theta((1-\eta)/H_f)^{1/\nu} \|\nabla f(x^k)\|^{(1-\nu)/(1+\nu)}\}$.*

(ii) Let $c_{\text{nc},\nu} := \eta(1-\eta)^{\frac{2}{\nu}}\theta^2/2^{\frac{2+\nu}{\nu}}$. The next iterate $x^{k+1} = x^k + \alpha_k d^k$ satisfies

$$f(x^k) - f(x^{k+1}) \geq c_{\text{nc},\nu} \gamma_k^{-\frac{1}{1+\nu}} \min \left\{ \|\nabla f(x^k)\|^{\frac{2+\nu}{1+\nu}}, 1 \right\}. \quad (12)$$

Moreover, we have $\|\nabla f(x^{k+1})\| \leq 2\alpha_k U_H \|d^k\|$ whenever $\|d^k\| \leq M := \min \{U_H^\nu, \theta(1-\eta)^{\frac{1}{\nu}} U_H / H_f^{\frac{1}{\nu}}\}$.

In either case, an $\Omega(\epsilon^{(2+\nu)/(1+\nu)})$ descent can be achieved for $k \in \mathbb{K}_{\epsilon,1}$ whenever $\|\nabla f(x^k)\| \geq \epsilon$. Recall that $\Delta_f = f(x^0) - f_{\text{low}}$ denotes the function value gap. It then follows directly that $|\mathbb{K}_{\epsilon,1}| \leq \mathcal{O}(\Delta_f \epsilon^{-(2+\nu)/(1+\nu)})$. Since the sequence $\{\gamma_k\}_{k \geq 0} \subseteq [\gamma_0, H_f]$ is nondecreasing in Algorithm 1, it is also straightforward to bound $|\mathbb{K}_{\epsilon,2}| \leq \lceil \log_2(H_f/\gamma_0) \rceil = \mathcal{O}(1)$ as γ_k doubles for each $k \in \mathbb{K}_{\epsilon,2}$. Note that $\mathbb{K}_{\epsilon,3}$ can be divided into at most $|\mathbb{K}_{\epsilon,1}| + |\mathbb{K}_{\epsilon,2}| + 1$ consecutive subsets with each containing at most $\lceil \log_2(U_g/\epsilon) \rceil$ iterations as the gradient halves for each $k \in \mathbb{K}_{\epsilon,3}$. This straightforward analysis immediately yields an iteration complexity of $\mathcal{O}(\epsilon^{-(2+\nu)/(1+\nu)} \ln(1/\epsilon))$. In the next theorem, we show that the logarithmic factor $\ln(1/\epsilon)$ can be removed with a more careful analysis; see the detailed proof in Appendix A.2.4.

Theorem 1. *Given Assumptions 1 and 2, we have $|\mathbb{K}_\epsilon| \leq \mathcal{O}(\Delta_f H_f^{\frac{1}{1+\nu}} \epsilon^{-\frac{2+\nu}{1+\nu}})$ for Algorithm 1.*

The iteration complexity of Theorem 1 matches the best known upper bound in [13] (non-adaptive damping scheme), and it also matches the lower bound provided [6]. In the Lipschitz-Hessian special case ($\nu = 1$), the iteration complexity of Theorem 1 reduces to $\mathcal{O}(\Delta_f H_f^{1/2} \epsilon^{-3/2})$, outperforming the $\mathcal{O}(\Delta_f H_f^2 \epsilon^{-3/2})$ bound by [14, 25, 29], and matching the best known upper bounds in [28].

3.3 Local superlinear convergence

By choosing $\varepsilon_k \propto \|\nabla f(x^k)\|^{\nu/(1+\nu)}$, Algorithm 1 allows that $\varepsilon_k \rightarrow 0$ as $\|\nabla f(x^k)\| \rightarrow 0$. Then, automatically, the damped Newton system (8) asymptotically reduces to the exact Newton system as the algorithm converges to a nondegenerate solution x^* , indicating local superlinear convergence. In the next theorem, we formally state this result.

Theorem 2. *Given Assumptions 1 and 2, let x^* be an arbitrary nondegenerate local minimizer of f such that $\nabla^2 f(x^*) \succeq \mu I$ for some $\mu > 0$. Then there exists $\delta > 0$ such that, for the sequence $\{x^k\}_{k \geq 0}$ generated by Algorithm 1, if $x^{k_0} \in B_\delta(x^*)$ for some $k_0 \geq 0$, then $\{x^k\}_{k \geq k_0} \subseteq B_\delta(x^*)$ and $\{x^k\}_{k \geq k_0}$ converges to x^* superlinearly in the sense that $\|x^{k+1} - x^*\| \leq \mathcal{O}(\|x^k - x^*\|^{(1+2\nu)/(1+\nu)})$.*

The key idea in the analysis is to show that the **CappedCG** subroutine will always successfully return an approximate solution to the damped Newton system (`d_type = SOL`) with $\alpha_k = 1$. Once the iterates x^k are sufficiently close to x^* , then the remaining analysis will be standard. The detailed proof of this theorem is moved to Appendix A.2.5. To our best knowledge, this is the first analysis for Newton-CG methods that simultaneously attains optimal global complexity and local superlinear convergence under the Hölder-Hessian setting, highlighting the advantage of adaptive damping as opposed to non-adaptive damping schemes like the one in [13].

4 A universal adaptive regularized Newton-CG method

Notice that Algorithm 1 requires prior knowledge of the Hölder exponent ν to determine the damping parameter. Although ν is often known in many applications, it is still desirable to have a completely parameter-free method that is universally optimal for all $\nu \in (0, 1]$, which will be the focus of this section.

4.1 Algorithm framework

Due to the lack of ν , we cannot form the ν -dependent adaptive damping scheme (8). Instead, we view Hölder continuity as an approximate Lipschitz continuity with controllable error, whose underlying idea can be traced back to [10, 22].

Proposition 1 ([13], Lemmas 1–2). *Given Assumption 2, for any $z > 0$ and $a \geq 2$, we have*

$$\mathcal{R}_1(y, x) \leq \frac{a \gamma_\nu(z)}{16} \|y - x\|^2 + \frac{z}{a} \quad \forall x, y \in \mathcal{L}_f(x^0; r_d), \quad (13)$$

where the inexact Lipschitz constant is a function of the error level $\gamma_\nu(z) := 4H_f^{\frac{2}{1+\nu}} z^{-\frac{1-\nu}{1+\nu}}$.

We remark that the function $\gamma_\nu(\cdot)$ is used only in the analysis and is not required in the actual implementation of the algorithm. As detailed in Algorithm 2, we solve

$$\left(\nabla^2 f(x^k) + 2(\gamma_k \|\nabla f(x^k)\|)^{1/2} I \right) d = -\nabla f(x^k)$$

by pretending $\nu = 1$ in the subproblem (8) of Algorithm 1. The sequence $\{\gamma_k\}$ adaptively estimates $\{\gamma_\nu(\|\nabla f(x^k)\|)\}$. However, we should also note an important difference that, unlike Algorithm 1 where $\{\gamma_k\}$ is always upper bounded by H_f , the inexact Lipschitz constant $\gamma_\nu(\|\nabla f(x^k)\|) \rightarrow +\infty$ when $\|\nabla f(x^k)\| \rightarrow 0$, posing new challenges in the analysis of the algorithm. Also, due to such potential unboundedness, Algorithm 2 may fail to guarantee sufficient descent. When this occurs, we increase γ_k by a fixed factor, instead of using local smoothness modulus estimation as adopted in Lines 6 and 9 of Algorithm 1.

Algorithm 2: A universal adaptive regularized Newton-CG method

```

1 Input: initial point  $x^0 \in \mathbb{R}^n$ , parameter  $\gamma_0 \geq 1$ , line-search parameters  $\eta \in (0, 1/2]$  and  $\theta \in (0, 1)$ .
2 Set  $c_{\text{sol}} = \frac{\eta(1-\eta)\theta}{400}$ .
3 while  $\nabla f(x^k) \neq 0$  do
4   Set  $H_k = \nabla^2 f(x^k)$ ,  $g_k = \nabla f(x^k)$ ,  $\varepsilon_k = (\gamma_k \|g_k\|)^{1/2}$ , and  $\zeta_k = \min\{1/2, \|g_k\|^{1/2}\}$ . Then call
            $(d, \text{d.type}) \leftarrow \text{CappedCG}(H_k, g_k, \varepsilon_k, \zeta_k)$ .
5   if  $\text{d.type} == \text{NC}$  then
6     Set  $d^k \leftarrow -\text{sgn}(d^T \nabla g_k) \frac{|d^T H_k d|}{\|d\|^3} d$ , and find  $\alpha_k = \theta^{j_k}$ , where  $j_k$  is the smallest nonnegative integer  $j$  such
           that
           
$$f(x^k + \theta^j d^k) < f(x^k) - \frac{\eta}{2} \theta^{2j} \|d^k\|^3. \quad (14)$$

7     If  $\|\nabla f(x^k + \alpha_k d^k)\| > \|g_k\|/2$  and  $\alpha_k < \theta/\gamma_k$  then set  $\gamma_{k+1} = 2\gamma_k$ .
8   else // d.type = SOL
9     Set  $d^k \leftarrow d$ .
10    if  $f(x^k + d^k) \leq f(x^k)$  and  $\|\nabla f(x^k + d^k)\| \leq \|g_k\|/2$  then set  $\alpha_k = 1$ .
11    else find  $\alpha_k = \theta^{j_k}$ , where  $j_k$  is the smallest nonnegative integer  $j$  such that
           
$$f(x^k + \theta^j d^k) < f(x^k) - \eta \varepsilon_k^{1/2} \theta^j \|d^k\|^2. \quad (15)$$

12    if  $\|\nabla f(x^k + \alpha_k d^k)\| > \|g_k\|/2$  and  $f(x^k) - f(x^k + \alpha_k d^k) < c_{\text{sol}} \gamma_k^{-1/2} \|g_k\|^{3/2}$  then set  $\gamma_{k+1} = 2\gamma_k$ .
13   Set  $x^{k+1} = x^k + \alpha_k d^k$  and  $k \leftarrow k + 1$ .

```

4.2 Global complexity bound

Similar to Lemma 1, the main iterates and trial iterates of Algorithm 2 also remain in a properly bounded region, as shown below.

Lemma 4. *Given Assumption 1, the sequences $\{x^k\}_{k \geq 0}$ and $\{d^k\}_{k \geq 0}$ generated by Algorithm 2 satisfy $x^k + \alpha d^k \in \mathcal{L}_f(x^0; r_d)$ for all $k \geq 0$ and $\alpha \in [0, 1]$, where r_d is defined by (4).*

The proof of this lemma is identical to that of Lemma 1 and is therefore omitted. In light of Lemma 4, we carry out the global complexity analysis of Algorithm 2 under the (H_f, ν) -Hölder continuity of $\nabla^2 f$ in $\mathcal{L}_f(x^0; r_d)$, as assumed in Assumption 2. Again, we partition the iterations of Algorithm 2 before finds an ϵ -stationary point ($\mathbb{K}_\epsilon := \{k : \|\nabla f(x^t)\| > \epsilon, \forall t \leq k\}$) into three parts:

$$\begin{aligned}\mathbb{K}_{\epsilon,1} &:= \{k \in \mathbb{K}_\epsilon : \|\nabla f(x^{k+1})\| > \|\nabla f(x^k)\|/2, \gamma_k > \gamma_\nu(\|\nabla f(x^k)\|)\}, \\ \mathbb{K}_{\epsilon,2} &:= \{k \in \mathbb{K}_\epsilon : \|\nabla f(x^{k+1})\| > \|\nabla f(x^k)\|/2, \gamma_k \leq \gamma_\nu(\|\nabla f(x^k)\|)\}, \\ \mathbb{K}_{\epsilon,3} &:= \{k \in \mathbb{K}_\epsilon : \|\nabla f(x^{k+1})\| \leq \|\nabla f(x^k)\|/2\}.\end{aligned}$$

Among the three sets, the iterations in $\mathbb{K}_{\epsilon,1}$ generate sufficient descent. To upper bound the cardinality of this subset, we establish sufficient descent when `CappedCG` outputs directions with `d_type=SOL`, and `NC`, respectively. The proofs of the two lemmas are deferred to Appendix A.3.

Lemma 5. *Given Assumptions 1 and 2, for all $k \in \mathbb{K}_{\epsilon,1}$ with d^k being the output of `CappedCG` with `d_type=SOL`, the following two statements hold.*

- (i) *The step length α_k is well defined and it satisfies $\alpha_k \geq (1 - \eta)\theta/3$.*
- (ii) *Let c_{sol} be defined in Algorithm 2. The next iterate $x^{k+1} = x^k + \alpha_k d^k$ satisfies*

$$f(x^k) - f(x^{k+1}) \geq c_{\text{sol}} \gamma_k^{-\frac{1}{2}} \|\nabla f(x^k)\|^{\frac{3}{2}}. \quad (16)$$

Moreover, the next gradient satisfies $\|\nabla f(x^{k+1})\| \leq 5\|\nabla f(x^k)\|$.

Lemma 6. *Given Assumptions 1 and 2, for all $k \in \mathbb{K}_{\epsilon,1}$ with d^k being the output of `CappedCG` with `d_type=NC`, the following two statements hold.*

- (i) *The step length α_k is well-defined, and $\alpha_k \geq \theta/\gamma_k$.*
- (ii) *Let $c_{\text{nc}} := \eta\theta^2/2$. The next iterate $x^{k+1} = x^k + \alpha_k d^k$ satisfies*

$$f(x^k) - f(x^{k+1}) \geq c_{\text{nc}} \gamma_k^{-\frac{1}{2}} \|\nabla f(x^k)\|^{\frac{3}{2}}. \quad (17)$$

Moreover, we have $\|\nabla f(x^{k+1})\| \leq 2\alpha_k U_H \|d^k\|$ whenever $\|d^k\| \leq \theta U_H$.

As discussed above, the sequence $\{\gamma_k\}$ may be unbounded. Consequently, the order (w.r.t. ϵ) of the $\mathcal{O}(\gamma_k^{-1/2} \|\nabla f(x^k)\|^{3/2})$ descent is not immediately clear from the two lemmas above. The next lemma provides a guaranteed upper bound on $\{\gamma_k\}$, whose proof can be found in Appendix A.3.

Lemma 7. *Suppose that Assumptions 1 and 2 hold. Let $\{\gamma_k\}$ be generated by Algorithm 2. Then, $\gamma_k \leq \max\{\gamma_0, 2\gamma_\nu(\epsilon)\}$ for all $k \in \mathbb{K}_\epsilon$.*

With the sufficient descent established above, the bound on $|\mathbb{K}_{\epsilon,1}|$ can be readily obtained. We now provide a brief proof overview for deriving bounds on $|\mathbb{K}_{\epsilon,2}|$ and $|\mathbb{K}_{\epsilon,3}|$, respectively. A key observation for bounding $|\mathbb{K}_{\epsilon,2}|$ is that $\{\gamma_k\}$ is updated only when $k \in \mathbb{K}_{\epsilon,2}$, but not every iteration in $\mathbb{K}_{\epsilon,2}$ triggers an update of $\{\gamma_k\}$. Thus, we partition $\mathbb{K}_{\epsilon,2}$ into two subsets, depending on whether γ_k is updated. The number of iterations when γ_k is updated can be bounded using the monotonicity of $\{\gamma_k\}$, while the iterations when γ_k is not updated can be bounded using sufficient descent. On the other hand, similar to Theorem 1, $|\mathbb{K}_{\epsilon,3}|$ can be bounded by exploiting the decrease of $\{\|\nabla f(x^k)\|\}$. The global iteration complexity of Algorithm 2 is stated in the following theorem, whose proof is deferred to Appendix A.3.4.

Theorem 3. *Given Assumptions 1 and 2, then $|\mathbb{K}_{\epsilon}| \leq \mathcal{O}(\Delta_f H_f^{\frac{1}{1+\nu}} \epsilon^{-\frac{2+\nu}{1+\nu}})$ holds for Algorithm 2.*

Theorem 3 shows that the iteration complexity of Algorithm 2 matches the best-known upper bound [13] and the lower bound [6]. In the Lipschitz-Hessian case ($\nu = 1$), it also matches the best-known upper bound [28], improving the dependence on H_f compared to [14, 25, 29].

4.3 Local superlinear convergence

Based on the global analysis of Algorithm 2, $\{\gamma_k\}$ may grow unbounded as the gradient vanishes, which precludes the standard local analysis. In this section, we first show that such potential unboundedness is a consequence of degeneracy. If Algorithm 2 enters a certain neighborhood of a nondegenerate stationary point x^* , $\{\gamma_k\}$ will stop growing and hence a common upper bound for $\{\gamma_k\}$ exists for all sufficiently large k , enabling the analysis of fast local superlinear convergence.

Lemma 8. *Suppose that Assumptions 1 and 2 hold. Let x^* be an arbitrary nondegenerate local minimizer of f such that $\nabla^2 f(x^*) \succeq \mu I$ for some $\mu > 0$. Then there exists $\delta > 0$ such that, if $x^k \in B_\delta(x^*)$, then `CappedCG` will return $\alpha_k = 1$ and `d.type = SOL`, and Algorithm 2 will enter the next iteration with $\gamma_{k+1} = \gamma_k$.*

This lemma shows that if an iterate falls in a δ -neighborhood of x^* , the parameter γ_k will stop growing in this step. However, it does not immediately guarantee that the next iterate will still stay in this neighborhood. In fact, the iterates may still, possibly, cross over the boundary of $B_\delta(x^*)$ and return an exploding sequence of $\{\gamma_k\}$. This causes a “chicken-and-egg” difficulty in the standard local superlinear analysis: (i) the existence of a non-expansive region requires a finite upper bound on $\{\gamma_k\}$; (ii) the upper bound on $\{\gamma_k\}$ relies on the existence of a non-expansive region (in $B_\delta(x^*)$). Therefore, the standard analysis strategy represented by Theorem 2 no longer works. Inspired by the Capture Theorem in [1], we show in the next lemma that a level-set based capture region of the form $B_\delta(x^*) \cap \{x : f(x) \leq f(x^*) + c\}$, $c > 0$ exists, so that once the iterates of a descent method enter this region, with the updates bounded by gradient magnitudes, all future iterates will be captured by this region due to a function value barrier.

Lemma 9. *Under the setting of Lemma 8, there exists a subset $S \subseteq B_\delta(x^*)$ such that if $x^{k_0} \in S$ for some $k_0 \geq 0$, then all future iterates will stay in S .*

Combining the above two lemmas, we observe that there exists a neighborhood S of x^* such that once an iterate of Algorithm 2 enters S , then all future iterates remain in S and the γ_k remains a finite constant, and consequently, a local superlinear rate can be established. The proof of Lemma 8 and 9, as well as Theorem 4 are all deferred to Appendix A.3.

Theorem 4. *Suppose the conditions of Lemma 8 hold. Let S be defined as in Lemma 9. For $\{x^k\}_{k \geq 0}$ generated by Algorithm 2, if $x^{k_0} \in S$ for some $k_0 \geq 0$, then $\{x^k\}_{k \geq k_0} \subseteq S$ and $\{x^k\}_{k \geq k_0}$ converges to x^* superlinearly in the sense that $\|x^{k+1} - x^*\| \leq O(\|x^k - x^*\|^{\min\{1+\nu, \frac{3}{2}\}})$.*

Note that the local convergence rate of Algorithm 2 is at least as fast as that of Algorithm 1. This gap follows from the orders of the damping terms, i.e., $\frac{1}{2} \geq \frac{\nu}{1+\nu}$ for $\nu \in (0, 1]$. Intuitively, when an iterate x^k is close enough to a nondegenerate stationary point with $\|\nabla f(x^k)\| \ll 1$, a larger exponent yields a smaller damping magnitude. Consequently, the corresponding linear system is closer to the classical Newton system, which leads to a faster local rate.

5 Numerical results

We conduct numerical experiments to evaluate the performance of our universal adaptive regularized Newton-CG method (Algorithm 2, abbreviated as ANCG), and compare it with a parameter-free Newton-CG (abbreviated as HNCG) [13, Algorithm 2] and an adaptive cubic regularized Newton method (abbreviated as ACRN) [11, Universal Method II].

For all methods, we initialize with $x^0 = (1, \dots, 1)^T$, and choose the following parameter settings, which provide numerically stable and efficient performance in practice:

- For ANCG, we set $(\gamma_0, \theta, \eta) = (10, 0.5, 0.01)$;
- For HNCG, we set $(\gamma_{-1}, \theta, r, \eta) = (10, 0.5, 2, 0.01)$;
- For ACRN, we set $H_0 = 10$. To solve its cubic regularized subproblems, we employ the gradient descent approach suggested in [3], with the initial point uniformly selected from the unit sphere.

All the algorithms are coded in Matlab, and all the computations are performed on a laptop with a 2.60 GHz Intel Core i5-14500 processor and 16 GB of RAM.

5.1 Infeasibility detection problem

Consider the infeasibility detection model from [2]:

$$\min_{x \in \mathbb{R}^n} \frac{1}{m} \sum_{i=1}^m (x^T A_i x + b_i^T x + c_i)_+^p,$$

where $p > 2$, $A_i \in \mathbb{R}^{n \times n}$, $b_i \in \mathbb{R}^n$, and $c_i \in \mathbb{R}$ for $1 \leq i \leq m$. For each triple (n, m, p) , we generate 10 random instances and aim to compute a 10^{-4} -stationary point using HNCG, ANCG, and ACRN.

Table 5.1 reports the average runtime, the average number of subproblems solved, and the average number of Hessian-vector products for the three methods. A “subproblem” refers to a cubic subproblem for ACRN and to a damped Newton system for HNCG and ANCG.

Overall, the results indicate that ANCG consistently achieves the lowest runtime across all tested dimensions, often reducing runtime by more than half relative to HNCG and by a even larger margin relative to ACRN. Moreover, ANCG requires substantially fewer subproblems and Hessian-vector products, demonstrating improved computational efficiency without compromising the quality of the final solution.

Dimension			Runtime (seconds)			Total subproblems			Hessian-vector products		
n	m	p	HNCG	ANCG	ACRN	HNCG	ANCG	ACRN	HNCG	ANCG	ACRN
100	10	2.25	0.03	0.01	1.37	49.3	17.0	44.0	1567.4	405.7	707717.9
100	10	2.50	0.02	0.01	1.44	50.6	21.0	43.8	1840.0	558.9	730280.6
100	10	2.75	0.02	0.01	1.54	54.0	24.2	49.0	2045.4	696.4	793014.5
100	10	3.00	0.02	0.01	2.07	56.6	28.0	50.2	2221.6	833.9	872891.8
500	50	2.25	1.30	0.83	20.47	63.0	20.0	54.3	4190.4	841.3	813247.3
500	50	2.50	1.62	0.97	19.89	72.7	25.0	56.4	5120.6	1170.5	902732.9
500	50	2.75	1.73	1.14	19.29	72.8	30.0	62.4	5758.3	1511.1	797625.4
500	50	3.00	1.85	1.26	14.89	75.0	34.0	65.5	6255.3	1746.5	629736.3
1000	100	2.25	14.94	8.28	84.02	70.1	21.0	58.0	5403.1	1013.7	874021.2
1000	100	2.50	18.16	10.59	88.39	80.3	27.0	61.3	7078.0	1569.0	941388.9
1000	100	2.75	20.49	12.37	62.12	84.0	32.0	64.0	7917.6	1919.5	611309.0
1000	100	3.00	22.07	14.19	59.87	85.1	37.0	67.7	8595.4	2295.5	552507.6

Table 1: Comparison of HNCG, ANCG, and ACRN across different dimensions for the infeasibility detection problem.

5.2 Single-layer neural networks problem

We consider the problem of training single-layer RePU neural networks [17]:

$$\min_{x \in \mathbb{R}^n} \frac{1}{m} \sum_{i=1}^m \phi((a_i^\top x)_+^p - b_i),$$

where $\phi(t) = t^2$ and $p > 2$. For each triple (n, m, p) , we generate 10 random instances by independently sampling $a_i \sim \mathcal{N}(0, I_n)$, and setting $b_i = |\bar{b}_i|$ where $\bar{b}_i \sim \mathcal{N}(0, 1)$.

Dimension			Runtime (seconds)			Total subproblems			Hessian-vector products		
n	m	p	HNCG	ANCG	ACRN	HNCG	ANCG	ACRN	HNCG	ANCG	ACRN
100	20	2.25	0.02	0.01	0.31	16.2	17.0	6.2	528.6	346.6	198021.6
100	20	2.50	0.00	0.00	0.38	18.4	18.3	8.9	611.3	397.2	241037.9
100	20	2.75	0.00	0.00	0.42	19.4	19.6	10.6	678.4	431.6	272028.1
100	20	3.00	0.01	0.00	0.50	22.3	21.1	13.3	810.8	469.7	316404.2
500	100	2.25	0.19	0.05	5.76	36.9	21.7	12.7	6385.5	1154.0	305064.4
500	100	2.50	0.24	0.06	6.36	40.9	24.2	13.8	10044.5	1470.4	332043.9
500	100	2.75	0.33	0.07	7.29	41.4	26.2	16.5	13660.2	1830.5	379043.2
500	100	3.00	0.43	0.08	7.94	44.9	28.5	18.4	18813.1	2180.7	419560.4
1000	200	2.25	1.53	0.31	27.09	42.8	23.4	15.7	16185.7	1566.9	351082.2
1000	200	2.50	2.60	0.38	30.94	45.4	25.6	18.3	29214.5	2091.2	401068.5
1000	200	2.75	4.40	0.42	36.02	48.6	27.3	21.3	50089.2	2632.8	457057.8
1000	200	3.00	7.05	0.51	38.82	50.3	30.1	23.0	83362.3	3450.8	492060.8

Table 2: Comparison of HNCG, ANCG, and ACRN across different dimensions for the single-layer neural network problem.

Our goal is to obtain a 10^{-4} -stationary point. The numerical results are summarized in Table 5.2. As before, the table reports the average runtime, the average number of subproblems, and the average number of Hessian-vector products for HNCG, ANCG, and ACRN. The results show a clear pattern across all tested dimensions: ANCG is consistently much faster than both HNCG and ACRN. In

addition, ANCG uses the fewest Hessian-vector products.

The two experiments indicate that ANCG solves these problems effectively, achieving the shortest runtime and using the fewest Hessian-vector products. Moreover, both HNCG and ANCG are consistently faster and require fewer Hessian-vector products than ACRN, suggesting a computational advantage of quadratic regularization over higher-order (cubic) regularization schemes that typically require solving more complex subproblems. Furthermore, relative to HNCG, ANCG consistently solves fewer total subproblems (i.e., damped Newton systems), often requiring less than half as many. This reduction highlights the practical benefit of eliminating regularization parameter line search from the algorithmic structure.

6 Conclusion

In this paper, we study Newton-CG methods for finding an ϵ -stationary point of a nonconvex function f whose Hessian is Hölder continuous with modulus $H_f > 0$ and exponent $\nu \in (0, 1]$. We identify two limitations in existing methods: non-adaptive regularization and a line-search based regularization parameter tuning procedure. The former can lead to inefficiency in the early stage and preclude local superlinear convergence, while the latter may make the algorithms computationally expensive. To circumvent these limitations, we propose two Newton-CG algorithms, depending on the availability of ν , that adaptively regularize the Newton system through auto-conditioning mechanisms, thereby eliminating the need for line search of the regularization parameters. To the best of our knowledge, this work proposes the first Newton-CG method that attains the best-known iteration complexity $\mathcal{O}(H_f^{1/(1+\nu)} \epsilon^{-(2+\nu)/(1+\nu)})$ for nonconvex problems with Hölder-continuous Hessians, while simultaneously enjoying local superlinear convergence. Numerical experiments further validate the practical advantages of our method.

References

- [1] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 3 edition, 2016.
- [2] R. H. Byrd, F. E. Curtis, and J. Nocedal. Infeasibility detection and SQP methods for nonlinear optimization. *SIAM J. Optim.*, 20(5):2281–2299, 2010.
- [3] Y. Carmon and J. Duchi. Gradient descent finds the cubic-regularized nonconvex Newton step. *SIAM J. Optim.*, 29(3):2146–2178, 2019.
- [4] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. Lower bounds for finding stationary points I. *Math. Program.*, 184(1-2):71–120, 2020.
- [5] C. Cartis, N. I. Gould, and P. L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part II: worst-case function- and derivative-evaluation complexity. *Math. Program.*, 130(2):295–319, 2011.
- [6] C. Cartis, N. I. Gould, and P. L. Toint. Worst-case evaluation complexity and optimality of second-order methods for nonconvex smooth optimization. *Proc. Int. Congr. Math. (ICM 2018)*, 3:3711–3750, 2018.

- [7] C. Cartis, N. I. Gould, and P. L. Toint. Universal regularization methods: varying the power, the smoothness and the accuracy. *SIAM J. Optim.*, 29(1):595–615, 2019.
- [8] C. Cartis, N. I. Gould, and P. L. Toint. Sharp worst-case evaluation complexity bounds for arbitrary-order nonconvex optimization with inexpensive constraints. *SIAM J. Optim.*, 30(1):513–541, 2020.
- [9] F. E. Curtis, D. P. Robinson, C. W. Royer, and S. J. Wright. Trust-region Newton-CG with strong second-order complexity guarantees for nonconvex optimization. *SIAM J. Optim.*, 31(1):518–544, 2021.
- [10] O. Devolder, F. Glineur, and Y. Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Math. Program.*, 146(1):37–75, 2014.
- [11] G. N. Grapiglia and Y. Nesterov. Regularized Newton methods for minimizing functions with Hölder continuous Hessians. *SIAM J. Optim.*, 27(1):478–506, 2017.
- [12] S. Gratton, S. Jerad, and P. L. Toint. Yet another fast variant of Newton’s method for nonconvex optimization. *IMA J. Numer. Anal.*, 2024.
- [13] C. He, H. Huang, and Z. Lu. Newton-CG methods for nonconvex unconstrained optimization with Hölder continuous Hessian. *Math. Oper. Res.*, 2025.
- [14] C. He, Z. Lu, and T. K. Pong. A Newton-CG based augmented Lagrangian method for finding a second-order stationary point of nonconvex equality constrained optimization with complexity guarantees. *SIAM J. Optim.*, 33(3):1734–1766, 2023.
- [15] G. Lan, T. Li, and Y. Xu. Projected gradient methods for nonconvex and stochastic optimization: new complexities and auto-conditioned stepsizes. *arXiv preprint arXiv:2412.14291*, 2024.
- [16] K. Levenberg. A method for the solution of certain non-linear problems in least squares. *Quart. Appl. Math.*, 2(2):164–168, 1944.
- [17] B. Li, S. Tang, and H. Yu. Better approximations of high dimensional smooth functions by deep neural networks with rectified power units. *Commun. Comput. Phys.*, 27(2):379–411, Feb. 2020.
- [18] D. H. Li, M. Fukushima, L. Qi, and N. Yamashita. Regularized Newton methods for convex minimization problems with singular solutions. *Comput. Optim. Appl.*, 28(2):131–147, 2004.
- [19] T. Li and G. Lan. A simple uniformly optimal method without line search for convex optimization. *Math. Program.*, pages 1–38, 2025.
- [20] D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *J. Soc. Indust. Appl. Math.*, 11(2):431–441, 1963.
- [21] K. Mishchenko. Regularized Newton method with global $\mathcal{O}(1/k^2)$ convergence. *SIAM J. Optim.*, 32(4):2671–2692, 2022.
- [22] Y. Nesterov. Universal gradient methods for convex optimization problems. *Math. Program.*, 152(1):381–404, 2015.

- [23] Y. Nesterov and B. T. Polyak. Cubic regularization of Newton method and its global performance. *Math. Program.*, 108(1):177–205, 2006.
- [24] R. A. Polyak. Regularized Newton method for unconstrained convex optimization. *Math. Program.*, 120(1):125–145, 2009.
- [25] C. W. Royer, M. O’Neill, and S. J. Wright. A Newton-CG algorithm with complexity guarantees for smooth unconstrained optimization. *Math. Program.*, 180(1):451–488, 2020.
- [26] K. Ueda and N. Yamashita. Convergence properties of the regularized Newton method for the unconstrained nonconvex optimization. *Appl. Math. Optim.*, 62(1):27–46, 2010.
- [27] C. Zhang and R. Jiang. Riemannian adaptive regularized Newton methods with Hölder continuous Hessians. *Comput. Optim. Appl.*, 92(1):29–79, Sept. 2025.
- [28] Y. Zhou, J. Xu, B. Li, C. Bao, C. Ding, and J. Zhu. A regularized Newton method for nonconvex optimization with global and local complexity guarantees. In *Advances in Neural Information Processing Systems*, 2025.
- [29] H. Zhu and Y. Xiao. A hybrid inexact regularized Newton and negative curvature method. *Comput. Optim. Appl.*, 88(3):849–870, 2024.

A Appendix

A.1 Supporting lemmas

We start with a technical lemma that will be applied in the analysis of Theorem 1 and 3 to remove the $\ln(1/\epsilon)$ factor in the iteration complexities.

Lemma 10. *For any nonempty set $\mathcal{I} \subset \mathbb{N}$, and $p, q, c, M > 0$, let $\{z_i\}_{i \in \mathcal{I}}$ be a positive sequence s.t. $\sum_{i \in \mathcal{I}} z_i^p \leq M$. Then it holds that $\sum_{i \in \mathcal{I}} \ln \frac{z_i}{c} \leq \frac{\max\{M/p, M/q\}}{ec^q}$, regardless of the cardinality of \mathcal{I} .*

Proof. Proof. First, applying the Jensen’s inequality to $\ln(\cdot)$ function, we obtain:

$$\begin{aligned} \sum_{i \in \mathcal{I}} \ln \frac{z_i}{c} &= \frac{|\mathcal{I}|}{p} \frac{\sum_{i \in \mathcal{I}} \ln(z_i^p)}{|\mathcal{I}|} + \frac{|\mathcal{I}|}{q} \ln \left(\frac{1}{c^q} \right) \leq \frac{|\mathcal{I}|}{p} \ln \left(\frac{\sum_{i \in \mathcal{I}} z_i^p}{|\mathcal{I}|} \right) + \frac{|\mathcal{I}|}{q} \ln \left(\frac{1}{c^q} \right) \\ &\leq \frac{|\mathcal{I}|}{p} \ln \left(\frac{M}{|\mathcal{I}|} \right) + \frac{|\mathcal{I}|}{q} \ln \left(\frac{1}{c^q} \right) \leq \max \left\{ \frac{1}{p}, \frac{1}{q} \right\} |\mathcal{I}| \ln \left(\frac{M}{c^q |\mathcal{I}|} \right). \end{aligned}$$

Note that for any $a, b > 0$, we have $b \ln(a/b) \leq a/e$, where $e \approx 2.718$ is the base of natural logarithm. Applying this fact to the above bound (with $a = Mc^{-q}$ and $b = |\mathcal{I}|$) proves the lemma. \square

A.2 Proof of Section 3

A.2.1 Proof of Lemma 1.

Proof. Because the line-search steps guarantee Algorithm 1 to be a descent method, we have $\{x^k\}_{k \geq 0} \subseteq \mathcal{L}_f(x^0)$. Then fix any $k \geq 0$, to show $x^k + \alpha d^k \in \mathcal{L}_f(x^0; r_d)$ for all $\alpha \in [0, 1]$, it suffices to show $\|d^k\| \leq r_d$, which has two possibilities based on the d-type outputs.

Case 1. `d_type=SOL`. In this case, applying the second inequality of Lemma 12(i) with $g = \nabla f(x^k)$ and $\sigma = \varepsilon_k = (\gamma_k \|\nabla f(x^k)\|^\nu)^{1/(1+\nu)}$, we obtain

$$\|d^k\| \leq 1.1(\|\nabla f(x^k)\|/\gamma_k)^{1/(1+\nu)}. \quad (18)$$

Together with the fact that $\gamma_k \geq \gamma_0 \geq 1$ and (3), we prove $\|d^k\| \leq 1.1 \max\{U_g, 1\} \leq r_d$.

Case 2. `d_type=NC`. Line 5 of Algorithm 1 and (3) yield $\|d^k\| = (d^k)^\top \nabla^2 f(x^k) d^k / \|d^k\|^2 \leq U_H \leq r_d$. Combining these two cases, we complete the proof of this lemma. \square

A.2.2 Proof of Lemma 2.

Proof. Throughout this proof, we will frequently use the shorthand $\varepsilon_k = (\gamma_k \|\nabla f(x^k)\|^\nu)^{1/(1+\nu)}$ defined in Algorithm 1 to simplify the notation. As the algorithm does not terminate, we have $\|\nabla f(x^k)\| \neq 0$ and hence $\varepsilon_k > 0$. Because `d_type = SOL`, we can apply the fourth inequality of Lemma 12(i) to yield $d^k \neq 0$. Moreover, for $k \in \mathbb{K}_{\varepsilon,1}$, we also have $\|\nabla f(x^{k+1})\| \geq \|\nabla f(x^k)\|/2$ and $\sigma_k \leq 2\gamma_k$. With the above information, we can start the proof.

Statement (i). If $\alpha_k = 1$, the statement clearly holds. If $\alpha_k < 1$, then we know $j_k \geq 1$. In this case, using the definition of σ_k (Algorithm 1 Line 9) for $j \in \{0, j_k - 1\}$ that violates (10), we have

$$\begin{aligned} & -\eta \varepsilon_k \theta^j \|d^k\|^2 \leq f(x^k + \theta^j d^k) - f(x^k) \\ & \stackrel{(i)}{\leq} \theta^j \nabla f(x^k)^\top d^k + \frac{\theta^{2j}}{2} (d^k)^\top \nabla^2 f(x^k) d^k + \mathcal{R}_0(x^k + \theta^j d^k, x^k) \\ & \stackrel{(ii)}{=} -\theta^j \left(1 - \frac{\theta^j}{2}\right) (d^k)^\top \left(\nabla^2 f(x^k) + 2\varepsilon_k I\right) d^k - \theta^{2j} \varepsilon_k \|d^k\|^2 + \frac{\mathcal{H}_0(x^k + \theta^j d^k, x^k) \|\theta^j d^k\|^{2+\nu}}{2} \\ & \stackrel{(iii)}{\leq} -\theta^j \varepsilon_k \|d^k\|^2 + \frac{\sigma_k \theta^{(2+\nu)j} \|d^k\|^{2+\nu}}{2}, \end{aligned}$$

where (i) is by the triangle inequality and the definition (5), (ii) is by the third relation of Lemma 12(i) and the definition (7), and (iii) is by the first inequality of Lemma 12(i), and the definition of σ_k in Algorithm 1 Line 9, which is effective for $j = 0$ or $j = j_k - 1$. As $d^k \neq 0$, dividing both sides of the above inequality by $\sigma_k \theta^j \|d^k\|^{2+\nu}/2$ yields

$$\theta^{(1+\nu)j} \geq \frac{2(1-\eta)\varepsilon_k}{\sigma_k \|d^k\|^\nu}, \quad j \in \{0, j_k - 1\}. \quad (19)$$

Then setting $j = j_k - 1$ in the above inequality gives

$$\alpha_k = \theta^{j_k} \geq \left(\frac{2(1-\eta)\varepsilon_k}{\sigma_k \|d^k\|^\nu}\right)^{1/(1+\nu)} \theta \geq \left(\frac{2(1-\eta)\gamma_k}{1.1^\nu \sigma_k}\right)^{1/(1+\nu)} \theta \geq \left(\frac{2(1-\eta)\gamma_k}{1.1^\nu H_f}\right)^{1/(1+\nu)} \theta, \quad (20)$$

where the second inequality is by the definition $\varepsilon_k = (\gamma_k \|\nabla f(x^k)\|^\nu)^{1/(1+\nu)}$ and (18), and the last inequality is because σ_k always underestimates H_f .

Statement (ii). Now we prove this statement by considering two separate cases below.

Case 1. $\alpha_k = 1$. In this case, we have $x^{k+1} = x^k + d^k$ and it holds that

$$\begin{aligned} \|\nabla f(x^{k+1})\| & \leq \mathcal{R}_1(x^k + d^k, x^k) + \|(\nabla^2 f(x^k) + 2\varepsilon_k I) d^k + \nabla f(x^k)\| + 2\varepsilon_k \|d^k\| \\ & \leq \sigma_k \|d^k\|^{1+\nu} + \frac{5}{2} \varepsilon_k \|d^k\|, \end{aligned}$$

where the first line is by the triangle inequality, and the second line is by the definition of σ_k (Algorithm 1 Line 9), the definition (7), the fourth relation of Lemma 12(i), and the definition of ζ_k in Line 3 of Algorithm 1. As a consequence, at least one of $\sigma_k \|d^k\|^{1+\nu} \geq \frac{\|\nabla f(x^{k+1})\|}{2}$ and $2.5\varepsilon_k \|d^k\| \geq \frac{\|\nabla f(x^{k+1})\|}{2}$ will hold. Combined with the definition of $\mathbb{K}_{\varepsilon,1}$ that $\sigma_k \leq 2\gamma_k$, and $\|\nabla f(x^{k+1})\| \geq \|\nabla f(x^k)\|/2$, we know $\|d^k\| \geq \gamma_k^{-\frac{1}{1+\nu}} \|\nabla f(x^k)\|^{\frac{1}{1+\nu}}/10$. By this lower bound of $\|d^k\|$ and (10), we complete the proof of (11) when $\alpha_k = 1$.

Case 2. $\alpha_k < 1$. In this case $j_k \geq 1$. Choosing $j = 0$ in (19) lower bounds $\|d^k\|^\nu \geq 2(1 - \eta)\varepsilon_k/\sigma_k$. Together with the first inequality of (20), $\sigma_k \leq 2\gamma_k$, and (10), we obtain that

$$f(x^k) - f(x^{k+1}) \geq \eta\varepsilon_k\alpha_k\|d^k\|^2 \geq \eta(1 - \eta)^{\frac{2}{\nu}}\theta\gamma_k^{-\frac{1}{1+\nu}}\|\nabla f(x^k)\|^{\frac{2+\nu}{1+\nu}},$$

which implies (11) in this case.

Next gradient bound. By the inequalities (6), (18), the fact that $\alpha_k \leq 1$, $\gamma_k \geq 1$, and the definition of ζ_k in Algorithm 1 Line 3, and the first and last inequalities of Lemma 12(i), one has that

$$\begin{aligned} \|\nabla f(x^{k+1})\| &= \|\nabla f(x^k + \alpha_k d^k)\| \\ &\leq \mathcal{R}_1(x^k + \alpha_k d^k, x^k) + \alpha_k \|\nabla^2 f(x^k) + 2\varepsilon_k I\| d^k + \|\nabla f(x^k)\| + 2\alpha_k \varepsilon_k \|d^k\| + (1 - \alpha_k) \|\nabla f(x^k)\| \\ &\leq H_f \|d^k\|^{1+\nu} + \frac{5\alpha_k \varepsilon_k}{2} \|d^k\| + (1 - \alpha_k) \|\nabla f(x^k)\| \\ &\leq (1.1)^{1+\nu} \frac{H_f}{\gamma_k} \|\nabla f(x^k)\| + (1 + 1.75\alpha_k) \|\nabla f(x^k)\| \leq (2H_f + 5) \|\nabla f(x^k)\|. \end{aligned}$$

Hence, we complete the proof of this lemma. \square

A.2.3 Proof of Lemma 3.

Proof. As the algorithm does not terminate, we know $\|\nabla f(x^k)\| \neq 0$. By Lemma 12(ii), we also confirm that $d^k \neq 0$ in this case. We should also note that the search direction d^k in this lemma is not the raw output of `CappedCG`. When `d.type=NC`, it undergoes an additional re-normalization step (Algorithm 1 Line 5) such that

$$\nabla f(x^k)^\top d^k \leq 0 \quad \text{and} \quad \frac{(d^k)^\top \nabla^2 f(x^k) d^k}{\|d^k\|^2} = -\|d^k\| < -\varepsilon_k, \quad (21)$$

where the second inequality is by Lemma 12(ii). Since $k \in \mathbb{K}_{\varepsilon,1}$, we have that $\|\nabla f(x^{k+1})\| \geq \|\nabla f(x^k)\|/2$ and $\sigma_k \leq 2\gamma_k$. With these in mind, let us prove the two statements one by one.

Statement (i). If (9) holds $j = 0$, then $\alpha_k = 1$, which clearly implies this statement. Now let us consider the case $j_k \geq 1$ and $\alpha_k = \theta^{j_k} < 1$. Because (9) fails for $j = j_k - 1$, one has

$$\begin{aligned} -\frac{\eta}{2}\theta^{2j}\|d^k\|^3 &\leq f(x^k + \theta^j d^k) - f(x^k) \leq \theta^j \nabla f(x^k)^\top d^k + \frac{\theta^{2j}}{2} (d^k)^\top \nabla^2 f(x^k) d^k + \mathcal{R}_0(x^k + \theta^j d^k, x^k) \\ &\leq -\frac{\theta^{2j}}{2} \|d^k\|^3 + \frac{\mathcal{H}_0(x^k + \theta^j d^k, x^k) \|\theta^j d^k\|^{2+\nu}}{2} = -\frac{\theta^{2j}}{2} \|d^k\|^3 + \frac{\sigma_k \theta^{(2+\nu)j} \|d^k\|^{2+\nu}}{2}, \end{aligned}$$

where the second line is due to (21) and the definition (7). Next, dividing both sides by $\theta^{2j}\sigma_k\|d^k\|^{2+\nu}/2$ yields that $\theta^{\nu j} \geq (1 - \eta)\|d^k\|^{1-\nu}/\sigma_k$. Combining this with $\alpha_k = \theta^{j_k}$, $\sigma_k \leq H_f$, and $\|d^k\| \geq \varepsilon_k$, and setting $j = j_k - 1$ in the above inequality gives

$$\alpha_k = \theta^{j_k} \geq \theta(1 - \eta)^{\frac{1}{\nu}} \|d^k\|^{\frac{1-\nu}{\nu}} / \sigma_k^{\frac{1}{\nu}} \geq \theta(1 - \eta)^{\frac{1}{\nu}} \gamma_k^{\frac{1-\nu}{\nu(1+\nu)}} \|\nabla f(x^k)\|^{\frac{1-\nu}{1+\nu}} / H_f^{\frac{1}{\nu}}, \quad (22)$$

which along with $\gamma_k \geq 1$ yields statement (i) as desired.

Statement (ii). If $\alpha_k = 1$, it follows from (9), $\|d^k\| \geq \varepsilon_k$, and $\gamma_k \geq 1$ that

$$f(x^k) - f(x^{k+1}) \geq \frac{\eta}{2} \varepsilon_k^3 = \frac{\eta}{2} (\gamma_k \|\nabla f(x^k)\|^\nu)^{\frac{3}{1+\nu}} \geq \frac{\eta}{2} \gamma_k^{-\frac{1}{1+\nu}} \min \left\{ \|\nabla f(x^k)\|^{\frac{2+\nu}{1+\nu}}, 1 \right\},$$

which implies (12). If $\alpha_k < 1$, $\|d^k\| \geq \varepsilon_k$, the first inequality of (22), and (9) imply

$$f(x^k) - f(x^{k+1}) \geq \frac{\eta}{2} \alpha_k^2 \varepsilon_k^3 \geq \frac{\eta}{2} \theta^2 \left(\frac{(1-\eta)\gamma_k}{\sigma_k} \right)^{\frac{2}{\nu}} \gamma_k^{-\frac{1}{1+\nu}} \|\nabla f(x^k)\|^{\frac{2+\nu}{1+\nu}},$$

combined with $\gamma_k/\sigma_k \geq 1/2$, we complete the proof of statement (ii).

Next gradient bound. Given $\|d^k\| \leq M = \min\{U_H^\nu, \theta(1-\eta)^{\frac{1}{\nu}} U_H/H_f^{\frac{1}{\nu}}\}$, let us bound the next gradient. When $\alpha_k = 1$, by $\|d^k\| \leq U_H^\nu$, we have $U_H\|d^k\| \geq \|d^k\|^{\frac{1+\nu}{\nu}}$. When $\alpha_k < 1$, by the first inequality of (22), $\sigma_k \leq H_f$, and $\|d^k\| \leq \theta(1-\eta)^{\frac{1}{\nu}} U_H/H_f^{\frac{1}{\nu}}$, we have $\alpha_k U_H\|d^k\| \geq \|d^k\|^{\frac{1+\nu}{\nu}}$. Overall, using Assumptions 1 and 2, we obtain that ∇f is Lipschitz continuous, which further implies that

$$\|\nabla f(x^{k+1})\| \leq \|\nabla f(x^k)\| + \alpha_k U_H \|d^k\| \stackrel{(i)}{\leq} \|d^k\|^{\frac{1+\nu}{\nu}} + \alpha_k U_H \|d^k\| \leq 2\alpha_k U_H \|d^k\|.$$

where (i) is because $\|d^k\| \geq \varepsilon_k \geq \|\nabla f(x^k)\|^{\frac{\nu}{1+\nu}}$. Hence we complete the proof. \square

A.2.4 Proof of Theorem 1.

Proof. Recall that $\mathbb{K}_\epsilon := \{k : \|\nabla f(x^t)\| > \epsilon, \forall t \leq k\}$ contains all iterations before termination. And by definition, $|\mathbb{K}_{\epsilon,i}|$, $i = 1, 2, 3$, form a partition of \mathbb{K}_ϵ .

Part 1. Bounding $|\mathbb{K}_{\epsilon,1}|$. Because $\gamma_k \leq H_f$ for all $k \in \mathbb{K}_\epsilon$, combining Lemmas 2 and 3, we know that each iteration $k \in \mathbb{K}_{\epsilon,1}$ results in either a constant descent in the objective value $f(x^k) - f(x^{k+1}) \geq c_{\text{nc},\nu} H_f^{-\frac{1}{1+\nu}}$, or a sufficient descent of

$$f(x^k) - f(x^{k+1}) \geq \min\{c_{\text{sol},\nu}, c_{\text{nc},\nu}\} \gamma_k^{-\frac{1}{1+\nu}} \|\nabla f(x^k)\|^{\frac{2+\nu}{1+\nu}} \geq \min\{c_{\text{sol},\nu}, c_{\text{nc},\nu}\} H_f^{-\frac{1}{1+\nu}} \epsilon^{\frac{2+\nu}{1+\nu}}.$$

Because Algorithm 1 is a descent method, and recalling that $\Delta_f = f(x^0) - f_{\text{low}}$, we have

$$|\mathbb{K}_{\epsilon,1}| \leq \frac{\Delta_f H_f^{\frac{1}{1+\nu}} \epsilon^{-\frac{2+\nu}{1+\nu}}}{\min\{c_{\text{sol},\nu}, c_{\text{nc},\nu}\}} + \frac{\Delta_f H_f^{\frac{1}{1+\nu}}}{c_{\text{nc},\nu}} = \mathcal{O} \left(\Delta_f H_f^{\frac{1}{1+\nu}} \epsilon^{-\frac{2+\nu}{1+\nu}} \right). \quad (23)$$

Part 2. Bounding $|\mathbb{K}_{\epsilon,2}|$. According to Line 10 of Algorithm 1, the sequence $\{\gamma_k\}_{k \geq 0}$ is nondecreasing and is always upper bounded by H_f . For every $k \in \mathbb{K}_{\epsilon,2}$, its definition requires $\sigma_k \geq 2\gamma_k$, and Line 10 of Algorithm 1 states that next $\gamma_{k+1} = \max\{\gamma_k, \sigma_k\} \geq 2\gamma_k$ will at least double the size. Therefore, we can bound $|\mathbb{K}_{\epsilon,2}|$ by

$$|\mathbb{K}_{\epsilon,2}| \leq \lceil \ln H_f / \ln 2 \rceil + 1 = \mathcal{O}(1).$$

Part 3. Bounding $|\mathbb{K}_{\epsilon,3}|$. The upper bound of this subset is a bit complicated. For all $k \in \mathbb{K}_{\epsilon,3}$, the next gradient halves: $\|\nabla f(x^{k+1})\| \leq \|\nabla f(x^k)\|/2$. Note that $\mathbb{K}_{\epsilon,1}, \mathbb{K}_{\epsilon,2}$ are finite and $\mathbb{K}_\epsilon = \mathbb{K}_{\epsilon,1} \cup \mathbb{K}_{\epsilon,2} \cup \mathbb{K}_{\epsilon,3}$, one can see that $\mathbb{K}_{\epsilon,3}$ can be partitioned into $|\mathbb{K}_{\epsilon,1}| + |\mathbb{K}_{\epsilon,2}| + 1$ disjoint subsets of *consecutive* nonnegative integers. Now, for those subsets that follows an iteration index $k \in \mathbb{K}_{\epsilon,1}$, we denote them by $\mathbb{K}_{\epsilon,3}^i$ with $i \in \mathcal{I}_1 := \{1, \dots, \ell_1\}$. For those subsets that follows an iteration index $k \in \mathbb{K}_{\epsilon,2}$ or in case the

subset contains 0, we denote then by $\mathbb{K}_{\epsilon,3}^i$ with $i \in \mathcal{I}_2 := \{\ell_1 + 1, \dots, \ell_1 + \ell_2\}$. Then clearly, we have $\ell_1 \leq |\mathbb{K}_{\epsilon,1}|$ and $\ell_2 \leq |\mathbb{K}_{\epsilon,2}| + 1$.

Because $\epsilon \leq \|\nabla f(x^k)\| \leq U_g$ for all $k \in \mathbb{K}_\epsilon$, each subset has at most $\lceil \ln(U_g/\epsilon)/\ln 2 \rceil + 1$ iterations. Consequently, for those subsets following an $\mathbb{K}_{\epsilon,2}$ iteration, we have

$$\sum_{i \in \mathcal{I}_2} |\mathbb{K}_{\epsilon,3}^i| \leq (|\mathbb{K}_{\epsilon,2}| + 1) \left(\left\lceil \frac{\ln(U_g/\epsilon)}{\ln 2} \right\rceil + 1 \right) \leq \mathcal{O} \left(\ln \frac{1}{\epsilon} \right).$$

Next, for those subsets that follow a $\mathbb{K}_{\epsilon,1}$ iteration (in \mathcal{I}_1), directly applying the above bound would result in an undesirable $\mathcal{O}(\epsilon^{-\frac{2+\nu}{1+\nu}} \ln \frac{1}{\epsilon})$ complexity. To remove the extra logarithmic factor, let us further partition these subsets into $\mathcal{I}_1 = \mathcal{I}_{\text{sol}} \cup \mathcal{I}_{\text{nc}}$ by whether the subset follows an iteration $k \in \mathbb{K}_{\epsilon,1}$ with `d.type=SOL` or `d.type=NC`, respectively. For notational simplicity, let us suppose the initial iteration of each $\mathbb{K}_{\epsilon,3}^i$ has gradient norm δ_i . The last iteration of $\mathbb{K}_{\epsilon,1}$ prior to each $\mathbb{K}_{\epsilon,3}^i$ has gradient norm $\bar{\delta}_i$, search direction norm $\hat{\delta}_i$ and line-search step length $\hat{\alpha}_i$.

To upper bound $\sum_{i \in \mathcal{I}_{\text{sol}}} |\mathbb{K}_{\epsilon,3}^i|$, by the relationship $\delta_i \leq (2H_f + 5)\bar{\delta}_i$ from Lemma 2, we have

$$\begin{aligned} \sum_{i \in \mathcal{I}_{\text{sol}}} |\mathbb{K}_{\epsilon,3}^i| &\leq \sum_{i \in \mathcal{I}_{\text{sol}}} \left(\left\lceil \frac{\ln(\delta_i/\epsilon)}{\ln 2} \right\rceil + 1 \right) \leq 2(1 + \ln(2H_f + 5)) |\mathcal{I}_{\text{sol}}| + 2 \sum_{i \in \mathcal{I}_{\text{sol}}} \ln(\bar{\delta}_i/\epsilon) \\ &\leq 2(1 + \ln(2H_f + 5)) |\mathbb{K}_{\epsilon,1}| + \frac{4\Delta_f H_f^{\frac{1}{1+\nu}} \epsilon^{-\frac{2+\nu}{1+\nu}}}{3ec_{\text{sol},\nu}} = \mathcal{O} \left(\Delta_f H_f^{\frac{1}{1+\nu}} \epsilon^{-\frac{2+\nu}{1+\nu}} \right), \end{aligned}$$

where the third inequality uses $|\mathcal{I}_{\text{sol}}| \leq |\mathbb{K}_{\epsilon,1}|$, $\sum_{i \in \mathcal{I}_{\text{sol}}} \bar{\delta}_i^{\frac{2+\nu}{1+\nu}} \leq H_f^{\frac{1}{1+\nu}} \Delta_f / c_{\text{sol},\nu}$ (due to (11) and $\gamma_k \leq H_f$), and Lemma 10 with $(z_i, p, q, c) = (\bar{\delta}_i, \frac{2+\nu}{1+\nu}, \frac{2+\nu}{1+\nu}, \epsilon)$.

For \mathcal{I}_{nc} , we divide it into $\mathcal{I}_{\text{nc}}^{\leq} := \{i \in \mathcal{I}_{\text{nc}} : \hat{\delta}_i \leq M\}$ with M defined in Lemma 3, and $\mathcal{I}_{\text{nc}}^> := \mathcal{I}_{\text{nc}} \setminus \mathcal{I}_{\text{nc}}^{\leq}$. For each $i \in \mathcal{I}_{\text{nc}}^>$, by (9) we know the last iteration in $\mathbb{K}_{\epsilon,1}$ will guarantee a descent of at least

$$\frac{\eta}{2} \hat{\alpha}_i^2 \hat{\delta}_i^3 \geq \frac{\eta}{2} \hat{\delta}_i^3 \cdot \min \left\{ 1, \theta^2 (1 - \eta)^{\frac{2}{\nu}} \hat{\delta}_i^{\frac{2-2\nu}{\nu}} H_f^{-\frac{2}{\nu}} \right\} \geq \frac{\eta}{2} \min \left\{ M^3, \theta^2 (1 - \eta)^{\frac{2}{\nu}} M^{\frac{2+\nu}{\nu}} H_f^{-\frac{2}{\nu}} \right\} =: C,$$

where the first inequality is due to the first inequality of (22) and the fact that $\sigma_k \leq H_f$. As the total descent is controlled by Δ_f , then we easily upper bound $|\mathcal{I}_{\text{nc}}^>|$ by $\Delta_f / C = \mathcal{O}(1)$ and obtain

$$\sum_{i \in \mathcal{I}_{\text{nc}}^>} |\mathbb{K}_{\epsilon,3}^i| \leq |\mathcal{I}_{\text{nc}}^>| \cdot \left(\left\lceil \frac{\ln(U_g/\epsilon)}{\ln 2} \right\rceil + 1 \right) \leq \mathcal{O} \left(\ln \frac{1}{\epsilon} \right). \quad (24)$$

For $i \in \mathcal{I}_{\text{nc}}^{\leq}$, again, (9) guarantees the last iteration, indexed by k_i , in $\mathbb{K}_{\epsilon,1}$ to generate a descent of at least $\frac{\eta}{2} \hat{\alpha}_i^2 \hat{\delta}_i^3 \geq \frac{\eta}{2} \hat{\alpha}_i^2 \hat{\delta}_i^2 \epsilon^{\frac{\nu}{1+\nu}}$, where we single out one $\hat{\delta}_i$ and lower bound it by (21), under the fact that $\hat{\delta}_i \geq \varepsilon_{k_i} = (\gamma_{k_i} \|\nabla f(x^{k_i})\|^\nu)^{1/(1+\nu)} \geq \epsilon^{\frac{\nu}{1+\nu}}$. Again, using Δ_f to control the total descent gives

$$\sum_{i \in \mathcal{I}_{\text{nc}}^{\leq}} \hat{\alpha}_i^2 \hat{\delta}_i^2 \leq \frac{2\Delta_f}{\eta \cdot \epsilon^{\frac{\nu}{1+\nu}}}, \quad (25)$$

which shall be used later as a summation upper bound in Lemma 10. With this in mind, and use the inequality $\delta_i \leq 2\hat{\alpha}_i U_H \hat{\delta}_i$ provided by Lemma 3, we have

$$\begin{aligned} \sum_{i \in \mathcal{I}_{\text{nc}}^{\leq}} |\mathbb{K}_{\epsilon,3}^i| &\leq \sum_{i \in \mathcal{I}_{\text{nc}}^{\leq}} \left(\left\lceil \frac{\ln(\delta_i/\epsilon)}{\ln 2} \right\rceil + 1 \right) \leq 2(1 + \ln(2U_H)) |\mathcal{I}_{\text{nc}}^{\leq}| + 2 \sum_{i \in \mathcal{I}_{\text{nc}}^{\leq}} \ln \left(\frac{\hat{\alpha}_i \hat{\delta}_i}{\epsilon} \right) \\ &\leq 2(1 + \ln(2U_H)) |\mathbb{K}_{\epsilon,1}| + \frac{4\Delta_f}{e\eta\epsilon^{(1+2\nu)/(1+\nu)}} = \mathcal{O} \left(\Delta_f H_f^{\frac{1}{1+\nu}} \epsilon^{-\frac{2+\nu}{1+\nu}} \right), \end{aligned} \quad (26)$$

where the third relation is by $|\mathcal{I}_{\text{nc}}^{\leq}| \leq |\mathbb{K}_{\epsilon,1}|$, the summation upper bound (25), and Lemma 10 with $(z_i, p, q, c) = (\hat{\alpha}_i \hat{\delta}_i, 2, 1, \epsilon)$. Synthesizing the inequalities (23)–(24) and (26), we obtain

$$|\mathbb{K}_{\epsilon}| = |\mathbb{K}_{\epsilon,1}| + |\mathbb{K}_{\epsilon,2}| + \sum_{i \in \mathcal{I}_2} |\mathbb{K}_{\epsilon,3}^i| + \sum_{i \in \mathcal{I}_{\text{sol}}} |\mathbb{K}_{\epsilon,3}^i| + \sum_{i \in \mathcal{I}_{\text{nc}}^{\geq}} |\mathbb{K}_{\epsilon,3}^i| + \sum_{i \in \mathcal{I}_{\text{nc}}^{\leq}} |\mathbb{K}_{\epsilon,3}^i| \leq \mathcal{O} \left(\Delta_f H_f^{\frac{1}{1+\nu}} \epsilon^{-\frac{2+\nu}{1+\nu}} \right),$$

hence we complete the proof of this theorem. \square

A.2.5 Proof of Theorem 2.

Proof. First, let us define the radius constant δ in Theorem 2 as

$$\delta = \min \left\{ \frac{1}{2L_g H_f^{1/\nu}}, \left(\frac{\mu}{2H_f} \right)^{\frac{1}{\nu}}, \frac{1}{4L_g} \left(\frac{\mu}{2H_f + 4H_f^{\frac{1}{1+\nu}} L_g^{\frac{\nu}{1+\nu}} + 2L_g^{\frac{1+2\nu}{1+\nu}}} \right)^{\frac{1+\nu}{\nu}} \right\},$$

where $L_g := \|\nabla^2 f(x^*)\| + 1$. By $x^{k_0} \in \mathcal{L}_f(x^0)$, and $\delta \leq \frac{1}{2H_f^{1/\nu}} < \frac{r_d}{2}$, we know $B_{\delta}(x^*) \subseteq \mathcal{L}_f(x^0; r_d)$. Consequently, we can apply the (H_f, ν) -Hölder continuity of $\nabla^2 f$ in $B_{\delta}(x^*)$ and obtain

$$\frac{\mu}{2} I \preceq \nabla^2 f(x) \preceq L_g I, \quad \forall x \in B_{\delta}(x^*), \quad (27)$$

where the first half of inequality uses $\delta \leq \left(\frac{\mu}{2H_f} \right)^{1/\nu}$. With this property, when a point $x^k \in B_{\delta}(x^*)$, the subroutine `CappedCG` will always output `d_type=SOL` because $\nabla^2 f(x^k) \succ 0$ has no negative curvature directions. Now we would like to claim that the stepsize output by the `CappedCG` subroutine will always take $\alpha_k = 1$ when $x^k \in B_{\delta}(x^*)$ by verifying the line search condition (10):

$$\begin{aligned} f(x^k + d^k) - f(x^k) &\leq \nabla f(x^k)^{\top} d^k + \frac{1}{2} (d^k)^{\top} \nabla^2 f(x^k) d^k + \frac{H_f}{2} \|d^k\|^{2+\nu} \\ &\stackrel{(i)}{=} -2\varepsilon_k \|d^k\|^2 - \frac{1}{2} d^{k\top} \nabla^2 f(x^k) d^k + \frac{H_f}{2} \|d^k\|^{2+\nu} \\ &\stackrel{(ii)}{\leq} -2\varepsilon_k \|d^k\|^2 - \frac{\mu}{4} \|d^k\|^2 + \frac{\mu}{4} \|d^k\|^2 \leq -\eta \varepsilon_k \|d^k\|^2, \end{aligned}$$

where (i) is by third relation of Lemma 12(i) and (ii) is by (27), and (18) that implies

$$\|d^k\|^{\nu} \leq 1.1^{\nu} (\|\nabla f(x^k)\|)^{\frac{\nu}{1+\nu}} \leq 1.1^{\nu} (L_g \|x^k - x^*\|)^{\frac{\nu}{1+\nu}} \leq 1.1^{\nu} (L_g \delta)^{\frac{\nu}{1+\nu}} \leq \frac{\mu}{2H_f}.$$

Now, we confirm that for Algorithm 1, all once an iteration enters the $B_{\delta}(x^*)$ neighborhood, it will always execute the Newton step. It remains to show that all future iterations will remain in this neighborhood and they will converge superlinearly to the locally optimal solution x^* . Now suppose $x^k \in B_{\delta}(x^*)$ for some $k \geq k_0$, then according to the previous argument, $x^{k+1} = x^k + d^k$, we have

$$\begin{aligned} \|x^{k+1} - x^*\| &\stackrel{(i)}{\leq} \|(\nabla^2 f(x^k) + 2\varepsilon_k I)^{-1} \nabla f(x^k) + x^k - x^*\| + \|d^k + (\nabla^2 f(x^k) + 2\varepsilon_k I)^{-1} \nabla f(x^k)\| \\ &\leq \|(\nabla^2 f(x^k) + 2\varepsilon_k I)^{-1}\| \left(\|\nabla f(x^k) + \nabla^2 f(x^k)(x^k - x^*)\| + 2\varepsilon_k \|x^k - x^*\| + \zeta_k \|\nabla f(x^k)\| \right) \\ &\stackrel{(ii)}{\leq} \frac{2}{\mu} \left(H_f \|x^k - x^*\|^{1+\nu} + 2\gamma_k^{\frac{1}{1+\nu}} \|\nabla f(x^k)\|^{\frac{\nu}{1+\nu}} \|x^k - x^*\| + \|\nabla f(x^k)\|^{\frac{1+2\nu}{1+\nu}} \right) \\ &\stackrel{(iii)}{\leq} \frac{2}{\mu} \left(H_f \delta^{1+\nu} + 2H_f^{\frac{1}{1+\nu}} L_g^{\frac{\nu}{1+\nu}} \delta^{\frac{1+2\nu}{1+\nu}} + (L_g \delta)^{\frac{1+2\nu}{1+\nu}} \right) \leq \delta, \end{aligned}$$

where (i) is by triangle inequality, (ii) is by (6) and the definition of ε_k, ζ_k in Line 3 of Algorithm 1, and (iii) is by (27). Based on this bound, we can conclude by induction that $\{x^k\}_{k \geq k_0} \subseteq B_\delta(x^*)$. Finally, by reusing the step (ii) in the above inequality, we establish the superlinear rate for $k \geq k_0$

$$\begin{aligned} \|\nabla f(x^{k+1})\| &\leq \frac{2Lg}{\mu} \left(H_f \|x^k - x^*\|^{1+\nu} + 2\gamma_k^{\frac{1}{1+\nu}} \|\nabla f(x^k)\|_{1+\nu}^\nu \|x^k - x^*\| + \|\nabla f(x^k)\|_{1+\nu}^{\frac{1+2\nu}{1+\nu}} \right) \\ &\leq \frac{2Lg}{\mu} \left(\frac{2^{1+\nu} H_f}{\mu^{1+\nu}} \|\nabla f(x^k)\|^{1+\nu} + \left(4\mu^{-1} H_f^{\frac{1}{1+\nu}} + 1 \right) \|\nabla f(x^k)\|_{1+\nu}^{\frac{1+2\nu}{1+\nu}} \right) \end{aligned}$$

That is, $\|\nabla f(x^{k+1})\| = \mathcal{O}(\|\nabla f(x^k)\|_{1+\nu}^{\frac{1+2\nu}{1+\nu}})$ for $k \geq k_0$, suggesting a superlinear rate of order $\frac{1+2\nu}{1+\nu}$. \square

A.3 Proof of Section 4.

Before presenting the proofs, we first provide supporting inequalities and lemmas. By direct computation, the following equivalence holds for any $z > 0$:

$$\gamma \geq \gamma_\nu(z) \iff \frac{(\gamma z)^{1/2}}{H_f} \geq 2^{1+\nu} \left(\frac{z}{\gamma}\right)^{\nu/2} \iff \frac{(\gamma z)^{(1-\nu)/2}}{H_f} \geq \frac{2^{1+\nu}}{\gamma^\nu}. \quad (28)$$

Its proof can be found in [13, Lemma 3] and is omitted here. The following lemma, used to establish sufficient descent later, shows that when γ_k is sufficiently large, either the gradient norm can be sufficiently reduced or the search direction is sufficiently large. Throughout Appendix A.3, we will frequently use the shorthand $\varepsilon_k = (\gamma_k \|\nabla f(x^k)\|)^{1/2}$ as defined in Algorithm 2.

Lemma 11. *Suppose Assumptions 1 and 2 hold, and that `CappedCG` outputs `d_type=SOL` at iteration $k \geq 0$ of Algorithm 2. If $\gamma_k \geq \gamma_\nu(\|\nabla f(x^k)\|)$, then either $\|\nabla f(x^{k+1})\| \leq \|\nabla f(x^k)\|/2$ or $11\|d^k\| \geq (\|\nabla f(x^k)\|/\gamma_k)^{1/2}$ holds.*

Proof. Proof. As the algorithm does not terminate, we have $\|\nabla f(x^k)\| \neq 0$ and hence $\varepsilon_k > 0$. Because `d_type = SOL`, we can see from the fourth inequality of Lemma 12(i) that $d^k \neq 0$. If $11\|d^k\| \geq (\|\nabla f(x^k)\|/\gamma_k)^{1/2}$ holds, then the lemma is proved. Now, suppose $11\|d^k\| < (\|\nabla f(x^k)\|/\gamma_k)^{1/2}$. We next prove $\|\nabla f(x^{k+1})\| \leq \|\nabla f(x^k)\|/2$ must hold. By Algorithm 2 Line 10, it suffices to show $f(x^k + d^k) \leq f(x^k)$ and $\|\nabla f(x^k + d^k)\| \leq \|\nabla f(x^k)\|/2$.

We first prove $f(x^k + d^k) \leq f(x^k)$. Suppose for contradiction that $f(x^k + d^k) > f(x^k)$. Denote $\varphi(\alpha) = f(x^k + \alpha d^k)$. Then, one has $\varphi(1) > \varphi(0)$. Since $d^k \neq 0$, it follows that

$$\varphi'(0) = \nabla f(x^k)^\top d^k \stackrel{(i)}{=} -(d^k)^\top (\nabla^2 f(x^k) + 2\varepsilon_k I) d^k \stackrel{(ii)}{\leq} -\varepsilon_k \|d^k\|^2 < 0.$$

Here, the steps (i) and (ii) follow from the third and first relations of Lemma 12, respectively, with $\sigma = \varepsilon_k = (\gamma_k \|\nabla f(x^k)\|)^{1/2}$, according to the design of Algorithm 2. Together with $\varphi(1) > \varphi(0)$, we know there must exist a local minimizer α_* such that $\varphi'(\alpha_*) = \nabla f(x^k + \alpha_* d^k)^\top d^k = 0$ and $\varphi(\alpha_*) < \varphi(0)$. By Lemma 4, we can then apply (6) to obtain that

$$\begin{aligned} \frac{H_f \alpha_*^{1+\nu} \|d^k\|^{2+\nu}}{1+\nu} &\geq (d^k)^\top (\nabla f(x^k + \alpha_* d^k) - \nabla f(x^k) - \alpha_* \nabla^2 f(x^k) d^k) \\ &\stackrel{(i)}{=} -(d^k)^\top \nabla f(x^k) - \alpha_* (d^k)^\top \nabla^2 f(x^k) d^k \\ &\stackrel{(ii)}{=} (1 - \alpha_*) (d^k)^\top (\nabla^2 f(x^k) + 2\varepsilon_k I) d^k + 2\alpha_* \varepsilon_k \|d^k\|^2 \\ &\stackrel{(iii)}{\geq} \varepsilon_k \|d^k\|^2. \end{aligned}$$

Here, (i) uses $\phi'(\alpha^*) = 0$, (ii) and (iii) use the third and first relations of Lemma 12(i), respectively. As $\|d^k\| \neq 0$, the above inequality further implies that

$$\|d^k\|^\nu \stackrel{(i)}{\geq} \frac{(1+\nu)(\gamma_k \|\nabla f(x^k)\|)^{1/2}}{\alpha_*^{1+\nu} H_f} \stackrel{(ii)}{\geq} \frac{(\gamma_k \|\nabla f(x^k)\|)^{1/2}}{H_f} \stackrel{(iii)}{\geq} 2^{1+\nu} \left(\frac{\|\nabla f(x^k)\|}{\gamma_k} \right)^{\nu/2},$$

where (i) uses the definition of ε_k in Algorithm 2, (ii) uses $\alpha_* \in (0, 1)$ and $\nu \geq 0$, (iii) uses (28) and $\gamma_k \geq \gamma_\nu(\|\nabla f(x^k)\|)$. Together with $11\|d^k\| < (\|\nabla f(x^k)\|/\gamma_k)^{1/2}$, we arrive at a contradiction that $11^\nu 2^{1+\nu} < 1$ with $\nu \in (0, 1]$. Hence, we have proven the claim that $f(x^k + d^k) \leq f(x^k)$.

We next prove $\|\nabla f(x^k + d^k)\| \leq \|\nabla f(x^k)\|/2$. By Lemma 4, we can apply (13), the last inequality of Lemma 12(i), $\zeta_k \in (0, 1)$, $11\|d^k\| < (\|\nabla f(x^k)\|/\gamma_k)^{1/2}$, and $\gamma_k \geq \gamma_\nu(\|\nabla f(x^k)\|)$ to obtain that

$$\begin{aligned} \|\nabla f(x^k + d^k)\| &\leq \mathcal{R}_1(x^k + d^k, x^k) + \|(\nabla^2 f(x^k) + 2\varepsilon_k I)d^k + \nabla f(x^k)\| + 2\varepsilon_k \|d^k\| \\ &\leq \frac{\gamma_\nu(\|\nabla f(x^k)\|)}{4} \|d^k\|^2 + \frac{\|\nabla f(x^k)\|}{4} + \frac{5}{2}\varepsilon_k \|d^k\| \\ &\leq \frac{\|\nabla f(x^k)\|}{484} + \frac{\|\nabla f(x^k)\|}{4} + \frac{5\|\nabla f(x^k)\|}{22} \leq \frac{\|\nabla f(x^k)\|}{2}. \end{aligned}$$

Hence, the proof of this lemma is complete. \square

With the above supporting lemmas, we can proceed to the proof of the main results in Section 4.

A.3.1 Proof of Lemma 5.

Proof. As the algorithm does not terminate, we have $\|\nabla f(x^k)\| \neq 0$ and hence $\varepsilon_k > 0$. Because `d_type` = SOL, we can apply the fourth inequality of Lemma 12(i) to yield $d^k \neq 0$. Moreover, for $k \in \mathbb{K}_{\varepsilon, 1}$, we have $\|\nabla f(x^{k+1})\| > \|\nabla f(x^k)\|/2$ and $\gamma_k > \gamma_\nu(\|\nabla f(x^k)\|)$. Using these and Lemma 11, we obtain that $11\|d^k\| \geq (\|\nabla f(x^k)\|/\gamma_k)^{1/2}$. With the above information, we can start the proof.

Statement (i). If $\alpha_k = 1$, the statement clearly holds. If $\alpha_k < 1$, then we know $j_k \geq 1$. In this case, for $j \in \{0, j_k - 1\}$ that violates (15), we have

$$\begin{aligned} -\eta\varepsilon_k \theta^j \|d^k\|^2 &\leq f(x^k + \theta^j d^k) - f(x^k) \stackrel{(i)}{\leq} \theta^j \nabla f(x^k)^\top d^k + \frac{\theta^{2j}}{2} (d^k)^\top \nabla^2 f(x^k) d^k + \mathcal{R}_0(x^k + \theta^j d^k, x^k) \\ &\stackrel{(ii)}{\leq} -\theta^j \left(1 - \frac{\theta^j}{2}\right) (d^k)^\top (\nabla^2 f(x^k) + 2\varepsilon_k I) d^k - \theta^{2j} \varepsilon_k \|d^k\|^2 + \frac{H_f \|\theta^j d^k\|^{2+\nu}}{2} \\ &\stackrel{(iii)}{\leq} -\theta^j \varepsilon_k \|d^k\|^2 + \frac{H_f \theta^{(2+\nu)j} \|d^k\|^{2+\nu}}{2}, \end{aligned}$$

where (i) is by the triangle inequality and the definition (5), (ii) is by the third relation of Lemma 12(i) and the residual bound (6), and (iii) is by the first inequality of Lemma 12(i). As $d^k \neq 0$, dividing both sides of the above inequality by $H_f \theta^j \|d^k\|^{2+\nu}/2$ yields

$$\theta^{(1+\nu)j} \geq \frac{2(1-\eta)\varepsilon_k}{H_f \|d^k\|^\nu} \geq 2^{2+\nu} (1-\eta) \frac{(\|\nabla f(x^k)\|/\gamma_k)^{\nu/2}}{\|d^k\|^\nu}, \quad j \in \{0, j_k - 1\}. \quad (29)$$

where the last inequality follows from the definition of ε_k , $\gamma_k > \gamma_\nu(\|\nabla f(x^k)\|)$, and (28). Then setting $j = j_k - 1$ gives

$$\begin{aligned} \alpha_k = \theta^{j_k} &\geq 2(1-\eta)\theta \left(\frac{(\|\nabla f(x^k)\|/\gamma_k)^{1/4}}{\|d^k\|^{1/2}} \right)^{\frac{2\nu}{1+\nu}} \left(\frac{(\|\nabla f(x^k)\|/\gamma_k)^{1/4}}{\sqrt{11}\|d^k\|^{1/2}} \right)^{\frac{1-\nu}{1+\nu}} \\ &\geq \frac{(1-\eta)\theta(\|\nabla f(x^k)\|/\gamma_k)^{1/4}}{2\|d^k\|^{1/2}} \geq \frac{(1-\eta)\theta}{3}, \end{aligned} \quad (30)$$

where the first inequality uses $11\|d^k\| \geq (\|\nabla f(x^k)\|/\gamma_k)^{1/2}$ and (29), and the last inequality uses the second inequality in Lemma 12(i). Hence the statement (i) holds.

Statement (ii). We prove this statement by considering two separate cases below.

Case 1. $\alpha_k = 1$. Recall that $11\|d^k\| \geq (\|\nabla f(x^k)\|/\gamma_k)^{1/2}$. Together with (10), this implies $f(x^k) - f(x^{k+1}) \geq (\eta/121)\gamma_k^{-1/2}\|\nabla f(x^k)\|^{3/2}$. Hence, (16) follows by the definition of c_{sol} .

Case 2. $\alpha_k < 1$. In this case, by $11\|d^k\| \geq (\|\nabla f(x^k)\|/\gamma_k)^{1/2}$, (15), and (30), one has that $f(x^k) - f(x^{k+1}) \geq \eta\varepsilon_k\alpha_k\|d^k\|^2 \geq (\eta(1-\eta)\theta/400)\gamma_k^{-1/2}\|\nabla f(x^k)\|^{3/2}$, which together with the definition of c_{sol} proves (16).

Next gradient bound. By (13), $\alpha_k \in (0, 1]$, $\gamma_k \geq 1$, $\zeta_k \leq 1$ (see Algorithm 2), $\gamma_\nu(\|\nabla f(x^k)\|) \leq \gamma_k$, and the first, second and last inequalities of Lemma 12(i), one has

$$\begin{aligned} & \|\nabla f(x^{k+1})\| = \|\nabla f(x^k + \alpha_k d^k)\| \\ & \leq \mathcal{R}_1(x^k + \alpha_k d^k, x^k) + \alpha_k\|(\nabla^2 f(x^k) + 2\varepsilon_k I)d^k + \nabla f(x^k)\| + 2\alpha_k\varepsilon_k\|d^k\| + (1 - \alpha_k)\|\nabla f(x^k)\| \\ & \leq \frac{\gamma_\nu(\|\nabla f(x^k)\|)}{4}\|d^k\|^2 + \frac{5}{4}\|\nabla f(x^k)\| + \frac{4 + \zeta_k}{2}\varepsilon_k\|d^k\| \\ & \leq \frac{\gamma_k}{4}1.1^2\frac{\|\nabla f(x^k)\|}{\gamma_k} + \frac{5}{4}\|\nabla f(x^k)\| + \frac{11}{4}\|\nabla f(x^k)\| \leq 5\|\nabla f(x^k)\|. \end{aligned}$$

Hence, we complete the proof of this lemma. \square

A.3.2 Proof of Lemma 6.

Proof. As the algorithm does not terminate, we know $\|\nabla f(x^k)\| \neq 0$. By Lemma 12(ii), we also confirm that $d^k \neq 0$ in this case. We should also note that the search direction d^k in this lemma is not the raw output of `CappedCG`. When `d.type=NC`, it undergoes an additional re-normalization step (Algorithm 2 Line 6) such that

$$\nabla f(x^k)^\top d^k \leq 0 \quad \text{and} \quad \frac{(d^k)^\top \nabla^2 f(x^k) d^k}{\|d^k\|^2} = -\|d^k\| < -\varepsilon_k, \quad (31)$$

where the second inequality is by Lemma 12(ii). Since $k \in \mathbb{K}_{\varepsilon, 1}$, we have that $\|\nabla f(x^{k+1})\| > \|\nabla f(x^k)\|/2$ and $\gamma_\nu(\|\nabla f(x^k)\|) < \gamma_k$. With these in mind, let us prove the two statements.

Statement (i). If (14) holds $j = 0$, then $\alpha_k = 1$, which immediately implies the statement. Now let us consider the case $j_k \geq 1$ and $\alpha_k = \theta^{j_k} < 1$. Because (14) fails for $j = j_k - 1$, one has

$$\begin{aligned} -\frac{\eta}{2}\theta^{2j}\|d^k\|^3 & \leq f(x^k + \theta^j d^k) - f(x^k) \leq \theta^j \nabla f(x^k)^\top d^k + \frac{\theta^{2j}}{2}(d^k)^\top \nabla^2 f(x^k) d^k + \mathcal{R}_0(x^k + \theta^j d^k, x^k) \\ & \leq -\frac{\theta^{2j}}{2}\|d^k\|^3 + \frac{H_f \theta^{(2+\nu)j}}{2}\|d^k\|^{2+\nu}. \end{aligned}$$

where the second inequality is from (31). Next, dividing both sides by $H_f \theta^{2j}\|d^k\|^{2+\nu}/2$, setting $j = j_k - 1$ in the above inequality and using the fact that $\nu \in (0, 1]$, $\eta \in (0, 1/2]$ and $\|d^k\| \geq \varepsilon_k$ yields

$$\alpha_k = \theta^{j_k} \geq \theta \frac{(1-\eta)^{1/\nu}\|d^k\|^{1-\nu}}{H_f^{1/\nu}} \geq \theta \frac{(1-\eta)^{1/\nu}\varepsilon_k^{1-\nu}}{H_f^{1/\nu}} \geq \theta \frac{(1-\eta)^{1/\nu}2^{\frac{1+\nu}{\nu}}}{\gamma_k} \geq \theta/\gamma_k, \quad (32)$$

where the third inequality follows from $\varepsilon_k = (\gamma_k\|\nabla f(x^k)\|)^{1/2}$ and (28). Hence, statement (i) holds.

Statement (ii). By (14), (31), and (32) we have

$$f(x^k) - f(x^{k+1}) \geq \frac{\eta}{2} \alpha_k^2 \|d^k\|^3 \geq \frac{\eta \theta^2}{2} \gamma_k^{-\frac{1}{2}} \|\nabla f(x^k)\|^{\frac{3}{2}},$$

which together with the definition of c_{nc} implies that (17) holds as desired.

Next gradient bound. Given $\|d^k\| \leq \theta U_H$, we consider two separate cases. When $\alpha_k = 1$, we have $U_H \|d^k\| \geq \theta U_H \|d^k\| \geq \|d^k\|^2 / \gamma_k$. When $\alpha_k < 1$, by (32) we have that $\alpha_k U_H \|d^k\| \geq \frac{\theta}{\gamma_k} U_H \|d^k\| \geq \frac{\|d^k\|^2}{\gamma_k}$. In addition, by Assumption 1 and 2, one has ∇f is Lipschitz continuous, which implies that

$$\|\nabla f(x^{k+1})\| \leq \|\nabla f(x^k)\| + \alpha_k U_H \|d^k\| \leq \frac{\|d^k\|^2}{\gamma_k} + \alpha_k U_H \|d^k\| \leq 2\alpha_k U_H \|d^k\|, \quad (33)$$

where the second inequality is due to (31) and $\varepsilon_k = (\gamma_k \|\nabla f(x^k)\|)^{1/2}$. This completes the proof. \square

A.3.3 Proof of Lemma 7.

Proof. Suppose for contradiction that $\gamma_k > \max\{\gamma_0, 2\gamma_\nu(\epsilon)\}$ for some $k \in \mathbb{K}_\epsilon$. Then there exists $\tilde{k} \leq k - 1$ such that $\max\{\gamma_0, 2\gamma_\nu(\epsilon)\} < \gamma_{\tilde{k}+1} = 2\gamma_{\tilde{k}}$. By this and $\|\nabla f(x^{\tilde{k}})\| > \epsilon$, one has that $\gamma_{\tilde{k}} > \gamma_\nu(\epsilon) \geq \gamma_\nu(\|\nabla f(x^{\tilde{k}})\|)$. In addition, since $\gamma_{\tilde{k}+1} = 2\gamma_{\tilde{k}}$, it follows from Algorithm 2 that $\|\nabla f(x^{\tilde{k}+1})\| > \|\nabla f(x^{\tilde{k}})\|/2$, and either $f(x^{\tilde{k}}) - f(x^{\tilde{k}+1}) < c_{\text{sol}} \gamma_{\tilde{k}}^{-1/2} \|\nabla f(x^{\tilde{k}})\|^{3/2}$ when $\text{d.type} = \text{SOL}$, or $\alpha_{\tilde{k}} < \theta/\gamma_{\tilde{k}}$ when $\text{d.type} = \text{NC}$. Note that $\tilde{k} \in \mathbb{K}_{\epsilon,1}$, which leads to a contradiction with the sufficient descent established in Lemma 5 and the lower bound for $\alpha_{\tilde{k}}$ established in Lemma 6. Hence, we have $\gamma_k \leq \max\{\gamma_0, 2\gamma_\nu(\epsilon)\}$ for all $k \in \mathbb{K}_\epsilon$. \square

A.3.4 Proof of Theorem 3.

Proof. Recall that $\mathbb{K}_\epsilon = \{k : \|\nabla f(x^t)\| > \epsilon, \forall t \leq k\}$ contains all iterations before finding an ϵ -stationary point. Also, $|\mathbb{K}_{\epsilon,i}|$, $i = 1, 2, 3$, form a partition of \mathbb{K}_ϵ .

Part 1. Bounding $|\mathbb{K}_{\epsilon,1}|$. Combining Lemmas 5, 6 and 7, we know that each iteration $k \in \mathbb{K}_{\epsilon,1}$ results in a sufficient descent of

$$f(x^k) - f(x^{k+1}) \geq \min\{c_{\text{sol}}, c_{\text{nc}}\} \epsilon^{\frac{3}{2}} / \gamma_k^{\frac{1}{2}} \geq \min\{c_{\text{sol}}, c_{\text{nc}}\} \epsilon^{\frac{3}{2}} / \max\{\gamma_0, 2\gamma_\nu(\epsilon)\}^{\frac{1}{2}},$$

As Algorithm 2 is a descent method and $\Delta_f = f(x^0) - f_{\text{low}}$, by the definition of $\gamma_\nu(\cdot)$, one has

$$|\mathbb{K}_{\epsilon,1}| \leq \frac{\Delta_f \epsilon^{-\frac{3}{2}} \max\{\gamma_0, 2\gamma_\nu(\epsilon)\}^{\frac{1}{2}}}{\min\{c_{\text{sol}}, c_{\text{nc}}\}} = \frac{\Delta_f \max\{\gamma_0^{\frac{1}{2}} \epsilon^{-\frac{3}{2}}, 2\sqrt{2} H_f^{\frac{1}{1+\nu}} \epsilon^{-\frac{2+\nu}{1+\nu}}\}}{\min\{c_{\text{sol}}, c_{\text{nc}}\}} = \mathcal{O}\left(\Delta_f H_f^{\frac{1}{1+\nu}} \epsilon^{-\frac{2+\nu}{1+\nu}}\right).$$

Part 2. Bounding $|\mathbb{K}_{\epsilon,2}|$. It follows from the update rule of $\{\gamma_k\}$, and Lemma 5 and 6 that γ_k increases only for some $k \in \mathbb{K}_{\epsilon,2}$. Thus, we further divide $\mathbb{K}_{\epsilon,2}$ into $\mathbb{K}_{\epsilon,2}^1 := \{k \in \mathbb{K}_{\epsilon,2} : \gamma_{k+1} = 2\gamma_k\}$ and $\mathbb{K}_{\epsilon,2}^2 := \mathbb{K}_{\epsilon,2} \setminus \mathbb{K}_{\epsilon,2}^1$. For $k \in \mathbb{K}_{\epsilon,2}^1$, $\gamma_{k+1} = 2\gamma_k$. Since $\{\gamma_k\}_{k \geq 0}$ is nondecreasing and is always bounded above by $\max\{\gamma_0, 2\gamma_\nu(\epsilon)\}$, we have $|\mathbb{K}_{\epsilon,2}^1| \leq \lceil \log_2(\max\{\gamma_0, 2\gamma_\nu(\epsilon)\}/\gamma_0) \rceil \leq \mathcal{O}(\ln(1/\epsilon))$. For $k \in \mathbb{K}_{\epsilon,2}^2$, one has $\|\nabla f(x^{k+1})\| > \|\nabla f(x^k)\|/2$. In addition, since $\gamma_{k+1} = \gamma_k$, by Lines 7 and 12 of Algorithm 2, we have either a sufficient descent $f(x^k) - f(x^{k+1}) \geq c_{\text{sol}} \gamma_k^{-1/2} \|\nabla f(x^k)\|^{1/2}$ for $\text{d.type} = \text{SOL}$, or a lower step size bound $\alpha_k \geq \theta/\gamma_k$ for $\text{d.type} = \text{NC}$, which further implies a sufficient descent $f(x^k) - f(x^{k+1}) \geq c_{\text{nc}} \gamma_k^{-1/2} \|\nabla f(x^k)\|^{1/2}$ due to (14) and (31). Therefore, each iteration $k \in \mathbb{K}_{\epsilon,2}^2$ results in a sufficient descent in the objective value. Using the same argument in Part 1,

we obtain that $|\mathbb{K}_{\epsilon,2}^2| \leq \mathcal{O}\left(\Delta_f H_f^{\frac{1}{1+\nu}} \epsilon^{-\frac{2+\nu}{1+\nu}}\right)$. Combining the upper bounds of $|\mathbb{K}_{\epsilon,2}^1|$ and $|\mathbb{K}_{\epsilon,2}^2|$ yields $|\mathbb{K}_{\epsilon,2}| \leq \mathcal{O}\left(\Delta_f H_f^{\frac{1}{1+\nu}} \epsilon^{-\frac{2+\nu}{1+\nu}}\right)$.

Part 3. Bounding $|\mathbb{K}_{\epsilon,3}|$. For each $k \in \mathbb{K}_{\epsilon,3}$, the next gradient halves: $\|\nabla f(x^{k+1})\| \leq \|\nabla f(x^k)\|/2$. Note that $\mathbb{K}_{\epsilon,1}$, $\mathbb{K}_{\epsilon,2}^1$, $\mathbb{K}_{\epsilon,2}^2$ are finite and $\mathbb{K}_{\epsilon} = \mathbb{K}_{\epsilon,1} \cup \mathbb{K}_{\epsilon,2}^1 \cup \mathbb{K}_{\epsilon,2}^2 \cup \mathbb{K}_{\epsilon,3}$, one can see that $\mathbb{K}_{\epsilon,3}$ can be partitioned into $|\mathbb{K}_{\epsilon,1}| + |\mathbb{K}_{\epsilon,2}^1| + |\mathbb{K}_{\epsilon,2}^2| + 1$ disjoint subsets of *consecutive* nonnegative integers. Now, for those subsets that follow an iteration index $k \in \mathbb{K}_{\epsilon,1}$, we denote them by $\mathbb{K}_{\epsilon,3}^i$ with $i \in \mathcal{I}_1 := \{1, \dots, \ell_1\}$. For those subsets that follow an iteration index $k \in \mathbb{K}_{\epsilon,2}^2$, we denote them by $\mathbb{K}_{\epsilon,3}^i$ with $i \in \mathcal{I}_2 = \{\ell_1+1, \dots, \ell_1+\ell_2\}$. For those subsets that follow an iteration index $k \in \mathbb{K}_{\epsilon,2}^1$, or in the case where the subset contains 0, we denote them by $\mathbb{K}_{\epsilon,3}^i$ with $i \in \mathcal{I}_3 := \{\ell_1 + \ell_2 + 1, \dots, \ell_1 + \ell_2 + \ell_3\}$. Then clearly, $\ell_1 \leq |\mathbb{K}_{\epsilon,1}|$, $\ell_2 \leq |\mathbb{K}_{\epsilon,2}^2|$, and $\ell_3 \leq |\mathbb{K}_{\epsilon,2}^1| + 1$.

Part 3 (i). Bounding $\sum_{i \in \mathcal{I}_3} |\mathbb{K}_{\epsilon,3}^i|$. Since $\epsilon \leq \|\nabla f(x^k)\| \leq U_g$ for all $k \in \mathbb{K}_{\epsilon}$, each subset has at most $\lceil \ln(U_g/\epsilon)/\ln 2 \rceil + 1$ iterations. Consequently, for those subsets following an $\mathbb{K}_{\epsilon,2}^1$ iteration, we have

$$\sum_{i \in \mathcal{I}_3} |\mathbb{K}_{\epsilon,3}^i| \leq (|\mathbb{K}_{\epsilon,2}^1| + 1) \left(\left\lceil \frac{\ln(U_g/\epsilon)}{\ln 2} \right\rceil + 1 \right) \leq \mathcal{O} \left(\left(\ln \frac{1}{\epsilon} \right)^2 \right).$$

For the rest subsets, directly applying the above bound would result in an undesirable $\mathcal{O}(\epsilon^{-\frac{2+\nu}{1+\nu}} \ln \frac{1}{\epsilon})$ complexity. To remove the extra logarithmic factor, a more careful analysis is required. For notational simplicity, let us suppose that the initial iteration of each $\mathbb{K}_{\epsilon,3}^i$ has gradient δ_i , and the iteration prior to each $\mathbb{K}_{\epsilon,3}^i$ has gradient norm $\bar{\delta}_i$, direction norm $\hat{\delta}_i$ and line-search step length $\hat{\alpha}_i$.

Part 3 (ii). Bounding $\sum_{i \in \mathcal{I}_1} |\mathbb{K}_{\epsilon,3}^i|$. We further partition $\mathcal{I}_1 = \mathcal{I}_{\text{sol}} \cup \mathcal{I}_{\text{nc}}$ according to whether the subset follows an iteration $k \in \mathbb{K}_{\epsilon,1}$ with `d_type=SOL` or `NC`, respectively.

To bound $\sum_{i \in \mathcal{I}_{\text{sol}}} |\mathbb{K}_{\epsilon,3}^i|$, by the relationship $\delta_i \leq 5\bar{\delta}_i$ from Lemma 5, we have

$$\begin{aligned} \sum_{i \in \mathcal{I}_{\text{sol}}} |\mathbb{K}_{\epsilon,3}^i| &\leq \sum_{i \in \mathcal{I}_{\text{sol}}} \left(\left\lceil \frac{\ln(\delta_i/\epsilon)}{\ln 2} \right\rceil + 1 \right) \leq (2 + 2 \ln(5)) |\mathcal{I}_{\text{sol}}| + \sum_{i \in \mathcal{I}_{\text{sol}}} 2 \ln(\bar{\delta}_i/\epsilon) \\ &\leq 6|\mathbb{K}_{\epsilon,1}| + \frac{4\Delta_f \max\{\gamma_0, 2\gamma_\nu(\epsilon)\}^{1/2}}{3ec_{\text{sol}}\epsilon^{3/2}} \leq \mathcal{O} \left(\Delta_f H_f^{\frac{1}{1+\nu}} \epsilon^{-\frac{2+\nu}{1+\nu}} \right) \end{aligned}$$

and the third inequality uses $|\mathcal{I}_{\text{sol}}| \leq |\mathbb{K}_{\epsilon,1}|$, $\sum_{i \in \mathcal{I}_{\text{sol}}} \bar{\delta}_i^{3/2} \leq \Delta_f \max\{\gamma_0, 2\gamma_\nu(\epsilon)\}^{1/2}/c_{\text{sol}}$ (due to (16) and Lemma 7), and Lemma 10 with $(z_i, p, q, c) = (\bar{\delta}_i, 3/2, 3/2, \epsilon)$.

For $\sum_{i \in \mathcal{I}_{\text{nc}}} |\mathbb{K}_{\epsilon,3}^i|$, we can further divide \mathcal{I}_{nc} into $\mathcal{I}_{\text{nc}}^{\leq} := \{i \in \mathcal{I}_{\text{nc}} : \hat{\delta}_i \leq \theta U_H\}$, and $\mathcal{I}_{\text{nc}}^{>} := \mathcal{I}_{\text{nc}} \setminus \mathcal{I}_{\text{nc}}^{\leq}$. For each $i \in \mathcal{I}_{\text{nc}}^{>}$, by (14), we know the last iteration in $|\mathbb{K}_{\epsilon,1}|$ will guarantee a descent of at least

$$\frac{\eta}{2} \hat{\alpha}_i^2 \hat{\delta}_i^3 \geq \frac{\eta \theta^5 U_H^3}{2\gamma_k^2} \geq \frac{\eta \theta^5 U_H^3}{2 \max\{\gamma_0, 2\gamma_\nu(\epsilon)\}^2} \geq \frac{\eta \theta^5 U_H^3}{2 \max\{\gamma_0^2, 64H_f^{4/(1+\nu)} \epsilon^{-(2-2\nu)/(1+\nu)}\}},$$

where the first inequality follows from $\hat{\alpha}_k \geq \theta/\gamma_k$, and the second inequality is because of $\gamma_k \leq \max\{\gamma_0, 2\gamma_\nu(\epsilon)\}$. As the total descent is controlled by Δ_f , we have $|\mathcal{I}_{\text{nc}}^{>}| \leq \mathcal{O}(\Delta_f \epsilon^{-\frac{2-2\nu}{1+\nu}})$ and

$$\sum_{i \in \mathcal{I}_{\text{nc}}^{>}} |\mathbb{K}_{\epsilon,3}^i| \leq |\mathcal{I}_{\text{nc}}^{>}| \cdot \left(\left\lceil \frac{\ln(U_g/\epsilon)}{\ln 2} \right\rceil + 1 \right) \leq o \left(\epsilon^{-\frac{2+\nu}{1+\nu}} \right).$$

For $i \in \mathcal{I}_{\text{nc}}^{\leq}$, by (14) we know the last iteration, indexed by k_i , in $|\mathbb{K}_{\epsilon,1}|$ will guarantee a descent of at least $\frac{\eta}{2} \hat{\alpha}_i^2 \hat{\delta}_i^3 \geq \frac{\eta}{2} \hat{\alpha}_i^2 \hat{\delta}_i^2 \epsilon^{\frac{1}{2}}$, where we single out one $\hat{\delta}_i$ and lower bound it by (31) under the fact that

$\hat{\delta}_i \geq \varepsilon_{k_i} = (\gamma_{k_i} \|\nabla f(x^{k_i})\|)^{\frac{1}{2}} \geq \epsilon^{\frac{1}{2}}$. Again, using Δ_f to control the total descent gives

$$\sum_{i \in \mathcal{I}_{\text{nc}}^{\leq}} \hat{\alpha}_i^2 \hat{\delta}_i^2 \leq \frac{2\Delta_f}{\eta \cdot \epsilon^{\frac{1}{2}}}, \quad (34)$$

which shall be used later as a summation upper bound in Lemma 10. With this in mind and using the inequality $\delta_i \leq 2\hat{\alpha}_i U_H \hat{\delta}_i$ provided by Lemma 6, we have

$$\begin{aligned} \sum_{i \in \mathcal{I}_{\text{nc}}^{\leq}} |\mathbb{K}_{\epsilon,3}^i| &\leq \sum_{i \in \mathcal{I}_{\text{nc}}^{\leq}} \left(\left\lceil \frac{\ln(\delta_i/\epsilon)}{\ln 2} \right\rceil + 1 \right) \leq (2 + 2 \ln(2U_H)) |\mathcal{I}_{\text{nc}}^{\leq}| + 2 \sum_{i \in \mathcal{I}_{\text{nc}}^{\leq}} \ln((\hat{\alpha}_i \hat{\delta}_i)/\epsilon) \\ &\leq (2 + 2 \ln(2U_H)) |\mathbb{K}_{\epsilon,1}| + \frac{4\Delta_f}{e\eta\epsilon} \leq \mathcal{O} \left(\Delta_f H_f^{\frac{1}{1+\nu}} \epsilon^{-\frac{2+\nu}{1+\nu}} \right) \end{aligned}$$

where the third inequality follows from $|\mathcal{I}_{\text{nc}}^{\leq}| \leq |\mathbb{K}_{\epsilon,1}|$, the summation bound (34), and Lemma 10 with $(z_i, p, q, c) = (\hat{\alpha}_i \hat{\delta}_i, 2, 1/2, \epsilon)$.

Part 3 (iii). Bounding $\sum_{i \in \mathcal{I}_2} |\mathbb{K}_{\epsilon,3}^i|$. We partition $\mathcal{I}_2 = \mathcal{I}'_{\text{sol}} \cup \mathcal{I}'_{\text{nc}}$ according to whether the subset follows an iteration $k \in \mathbb{K}_{\epsilon,2}^2$ with d.type=SOL or NC, respectively. Before proceeding, we establish an upper bound for the next gradient norm with iteration index $k \in \mathbb{K}_{\epsilon,2}^2$. For d.type=SOL, by (13), the fact that $\alpha_k \leq 1$, $\gamma_k \geq 1$, and $\zeta_k \leq 1/2$ in Algorithm 2, and the first, second and last relations of Lemma 12(i), one has

$$\begin{aligned} \|\nabla f(x^{k+1})\| &= \|\nabla f(x^k + \alpha_k d^k)\| \\ &\leq \mathcal{R}_1(x^k + \alpha_k d^k, x^k) + \alpha_k \|\nabla^2 f(x^k) + 2\varepsilon_k I\| d^k + \|\nabla f(x^k)\| + 2\alpha_k \varepsilon_k \|d^k\| + (1 - \alpha_k) \|\nabla f(x^k)\| \\ &\leq \frac{\gamma_\nu (\|\nabla f(x^k)\|)}{4} \|d^k\|^2 + \frac{5}{4} \|\nabla f(x^k)\| + \frac{4 + \zeta_k}{2} \varepsilon_k \|d^k\| \leq 2H_f^{\frac{2}{1+\nu}} \|\nabla f(x^k)\|^{\frac{2\nu}{1+\nu}} + 6\|\nabla f(x^k)\|. \end{aligned} \quad (35)$$

Thus, we have $\|\nabla f(x^{k+1})\| \leq 4H_f^{\frac{2}{1+\nu}} \|\nabla f(x^k)\|^{\frac{2\nu}{1+\nu}}$ if $\|\nabla f(x^k)\| \leq (H_f^2/3^{1+\nu})^{1/(1-\nu)}$ and $\nu \in (0, 1)$. For d.type = NC, the same reasoning in (33) gives $\|\nabla f(x^{k+1})\| \leq 2\alpha_k U_H \|d^k\|$ whenever $\|d^k\| \leq \theta U_H$. With these next gradient bounds, we are ready to bound $\sum_{i \in \mathcal{I}'_{\text{sol}}} |\mathbb{K}_{\epsilon,3}^i|$ and $\sum_{i \in \mathcal{I}'_{\text{nc}}} |\mathbb{K}_{\epsilon,3}^i|$.

For $\sum_{i \in \mathcal{I}'_{\text{sol}}} |\mathbb{K}_{\epsilon,3}^i|$, we consider the following two cases:

Case 1. When $\nu \in (0, 1)$, we further divide $\mathcal{I}'_{\text{sol}}$ into $\mathcal{I}'_{\text{sol}}^{\leq} := \{i \in \mathcal{I}'_{\text{sol}} : \bar{\delta}_i \leq (H_f^2/3^{1+\nu})^{1/(1-\nu)}\}$ and $\mathcal{I}'_{\text{sol}}^> := \mathcal{I}'_{\text{sol}} \setminus \mathcal{I}'_{\text{sol}}^{\leq}$. For $\mathcal{I}'_{\text{sol}}^>$, we have $\bar{\delta}_i > (H_f^2/3^{1+\nu})^{1/(1-\nu)}$. By Line 12 of Algorithm 2, we know the last iteration in $|\mathbb{K}_{\epsilon,2}^2|$, indexed by k , will guarantee a descent of at least

$$c_{\text{sol}} \gamma_k^{-1/2} \|\nabla f(x^k)\|^{3/2} \geq c_{\text{sol}} \max\{\gamma_0, 2\gamma_\nu(\epsilon)\}^{-1/2} H_f^{3/(1-\nu)} / 3^{(3+3\nu)/(2-2\nu)} \geq \Omega(\epsilon^{\frac{1-\nu}{2+2\nu}}).$$

Using Δ_f to control the total descent yields $|\mathcal{I}'_{\text{sol}}^>| \leq \mathcal{O}(\Delta_f \epsilon^{-\frac{1-\nu}{2+2\nu}})$, which implies

$$\sum_{i \in \mathcal{I}'_{\text{sol}}^>} |\mathbb{K}_{\epsilon,3}^i| \leq |\mathcal{I}'_{\text{sol}}^>| \cdot \left(\left\lceil \frac{\ln(U_g/\epsilon)}{\ln 2} \right\rceil + 1 \right) \leq o\left(\epsilon^{-\frac{2+\nu}{1+\nu}}\right).$$

For $\mathcal{I}'_{\text{sol}}^{\leq}$, the next gradient bound for $|\mathbb{K}_{\epsilon,2}^2|$ implies $\delta_i \leq 4H_f^{2/(1+\nu)} \bar{\delta}_i^{2\nu/(1+\nu)}$. It then follows that

$$\begin{aligned} \sum_{i \in \mathcal{I}'_{\text{sol}}^{\leq}} |\mathbb{K}_{\epsilon,3}^i| &\leq \sum_{i \in \mathcal{I}'_{\text{sol}}^{\leq}} \left(\left\lceil \frac{\ln(\delta_i/\epsilon)}{\ln 2} \right\rceil + 1 \right) \leq \left(2 + 2 \ln(4H_f^{2/(1+\nu)}) \right) |\mathcal{I}'_{\text{sol}}^{\leq}| + 2 \sum_{i \in \mathcal{I}'_{\text{sol}}^{\leq}} \ln(\bar{\delta}_i^{2\nu/(2+\nu)}/\epsilon) \\ &\leq (2 + 2 \ln(4H_f^{2/(1+\nu)})) |\mathbb{K}_{\epsilon,2}^2| + \frac{4\Delta_f \max\{\gamma_0, 2\gamma_\nu(\epsilon)\}^{1/2}}{e c_{\text{sol}} \epsilon^{1/2}} \leq \mathcal{O}(\Delta_f H_f^{\frac{1}{1+\nu}} \epsilon^{-\frac{2+\nu}{1+\nu}}), \end{aligned}$$

where the third inequality follows from $|\mathcal{I}'_{\text{sol}}{}^{\leq}| \leq |\mathbb{K}_{\epsilon,2}^2|$, $\sum_{i \in \mathcal{I}'_{\text{sol}}{}^{\leq}} \bar{\delta}_i^{3/2} \leq \Delta_f \max\{\gamma_0, 2\gamma_\nu(\epsilon)\}^{1/2}/c_{\text{sol}}$ (due to the definitions of \mathcal{I}_2 and $\mathbb{K}_{\epsilon,2}^2$) and Lemma 10 with $(z_i, p, q, c) = (\bar{\delta}_i^{\frac{2\nu}{2+\nu}}, \frac{3(2+\nu)}{4\nu}, \frac{1}{2}, \epsilon)$.

Case 2. When $\nu = 1$, (35) implies $\delta_i \leq (2H_f + 6)\bar{\delta}_i$. Then we have

$$\begin{aligned} \sum_{i \in \mathcal{I}'_{\text{sol}}} |\mathbb{K}_{\epsilon,3}^i| &\leq \sum_{i \in \mathcal{I}'_{\text{sol}}} \left(\left\lceil \frac{\ln(\delta_i/\epsilon)}{\ln 2} \right\rceil + 1 \right) \leq 2(1 + \ln(2H_f + 6))|\mathcal{I}'_{\text{sol}}| + 2 \sum_{i \in \mathcal{I}'_{\text{sol}}} \ln(\bar{\delta}_i/\epsilon) \\ &\leq 2(1 + \ln(2H_f + 6))|\mathbb{K}_{\epsilon,2}^2| + \frac{2\Delta_f \max\{\gamma_0, 8H_f\}^{1/2}}{e c_{\text{sol}} \epsilon} = \mathcal{O}(\Delta_f H_f^{\frac{1}{2}} \epsilon^{-\frac{3}{2}}), \end{aligned}$$

where the third inequality uses $|\mathcal{I}'_{\text{sol}}| \leq |\mathbb{K}_{\epsilon,2}^2|$, $\sum_{i \in \mathcal{I}'_{\text{sol}}{}^{\leq}} \bar{\delta}_i^{3/2} \leq \Delta_f \max\{\gamma_0, 8H_f\}^{1/2}/c_{\text{sol}}$ (due to the definitions of \mathcal{I}_2 and $\mathbb{K}_{\epsilon,2}^2$) and Lemma 10 with $(z_i, p, q, c) = (\bar{\delta}_i, 3/2, 1, \epsilon)$.

Combining the above two cases, we obtain that $\sum_{i \in \mathcal{I}'_{\text{sol}}} |\mathbb{K}_{\epsilon,3}^i| \leq \mathcal{O}(\Delta_f H_f^{\frac{1}{1+\nu}} \epsilon^{-\frac{2+\nu}{1+\nu}})$ for all $\nu \in (0, 1]$. Note from the proof of Lemma 6 that the result $\|\nabla f(x^{k+1})\| \leq 2\alpha_k U_H \|d^k\|$ whenever $\|d^k\| \leq \theta U_H$ is also valid at iteration $k \in \mathbb{K}_{\epsilon,2}^2$ because $\alpha_k \geq \theta/\gamma_k$. Moreover, by the definition of \mathcal{I}'_{nc} , one sees that $\mathbb{K}_{\epsilon,2}^i$, $i \in \mathcal{I}'_{\text{nc}}$ follows an iteration $k \in \mathbb{K}_{\epsilon,2}^2$. Therefore, for \mathcal{I}'_{nc} , we can use the same arguments as those for bounding $\sum_{i \in \mathcal{I}_{\text{nc}}} |\mathbb{K}_{\epsilon,3}^i|$ in Part 3(ii). That is, we start by dividing \mathcal{I}'_{nc} into two disjoint subsets: $\mathcal{I}'_{\text{nc}}{}^{\leq} = \{i \in \mathcal{I}'_{\text{nc}} : \hat{\delta}_i \leq \theta U_H\}$ and $\mathcal{I}'_{\text{nc}}{}^{>} = \mathcal{I}'_{\text{nc}}/\mathcal{I}'_{\text{nc}}{}^{\leq}$. We then bound $\sum_{i \in \mathcal{I}'_{\text{nc}}{}^{\leq}} |\mathbb{K}_{\epsilon,3}^i|$ and $\sum_{i \in \mathcal{I}'_{\text{nc}}{}^{>}} |\mathbb{K}_{\epsilon,3}^i|$, respectively, following the same derivations used in Part 3(ii). As a result, we obtain $\sum_{i \in \mathcal{I}'_{\text{nc}}} |\mathbb{K}_{\epsilon,3}^i| \leq \mathcal{O}(\Delta_f H_f^{\frac{1}{1+\nu}} \epsilon^{-\frac{2+\nu}{1+\nu}})$.

Lastly, summarizing Parts 1-3, we obtain

$$\begin{aligned} |\mathbb{K}_\epsilon| &= |\mathbb{K}_{\epsilon,1}| + |\mathbb{K}_{\epsilon,2}| + \sum_{i \in \mathcal{I}_3} |\mathbb{K}_{\epsilon,3}^i| + \sum_{i \in \mathcal{I}_{\text{sol}}} |\mathbb{K}_{\epsilon,3}^i| + \sum_{i \in \mathcal{I}_{\text{nc}}} |\mathbb{K}_{\epsilon,3}^i| + \sum_{i \in \mathcal{I}'_{\text{sol}}} |\mathbb{K}_{\epsilon,3}^i| + \sum_{i \in \mathcal{I}'_{\text{nc}}} |\mathbb{K}_{\epsilon,3}^i| \\ &\leq \mathcal{O} \left(\Delta_f H_f^{\frac{1}{1+\nu}} \epsilon^{-\frac{2+\nu}{1+\nu}} \right). \end{aligned}$$

Hence, we complete the proof of this theorem. \square

A.3.5 Proof of Lemma 8.

Proof. First, let us define the radius constant δ in Theorem 4 as

$$\delta = \min \left\{ \frac{1}{2L_g H_f^{1/\nu}}, \left(\frac{\mu}{2H_f} \right)^{\frac{1}{\nu}}, \frac{1}{2L_g} \left(\frac{\mu}{2H_f} \right)^{\frac{2}{\nu}}, \left[\frac{6L_g}{\mu} \left(\frac{2H_f}{\mu} + 8 \left(\frac{H_f}{\mu} \right)^{\frac{1}{1+\nu}} + L_g^{\frac{1}{2}} \right) \right]^{-\frac{1+\nu}{\nu}} \right\}, \quad (36)$$

where $L_g := \|\nabla^2 f(x^*)\| + 1$. The same argument at the beginning of A.2.5 yields that $B_\delta(x^*) \subseteq \mathcal{L}_f(x^0; r_d)$, the relation (27) holds for all $x \in B_\delta(x^*)$, once an iteration enters the $B_\delta(x^*)$ neighborhood, Algorithm 2 will always execute the Newton step with $\alpha_k = 1$ and $\text{d_type} = \text{SOL}$.

We now show that $\gamma_{k+1} = \gamma_k$ for any $x^k \in B_\delta(x^*)$. From Algorithm 2 we know that γ_k increases only when $k \in \mathbb{K}_{\epsilon,2}$. To prove this statement, it is suffice to show that $k \notin \mathbb{K}_{\epsilon,2}$ when $x^k \in B_\delta(x^*)$. Suppose for contradiction that $x^k \in B_\delta(x^*)$ for some $k \in \mathbb{K}_{\epsilon,2}$. Then, we have

$$\begin{aligned} \|x^{k+1} - x^*\| &\stackrel{(i)}{\leq} \|(\nabla^2 f(x^k) + 2\varepsilon_k I)^{-1} \nabla f(x^k) + x^k - x^*\| + \|d^k + (\nabla^2 f(x^k) + 2\varepsilon_k I)^{-1} \nabla f(x^k)\| \\ &\leq \|(\nabla^2 f(x^k) + 2\varepsilon_k I)^{-1}\| \left(\|\nabla f(x^k) + \nabla^2 f(x^k)(x^k - x^*)\| + 2\varepsilon_k \|x^k - x^*\| + \zeta_k \|\nabla f(x^k)\| \right) \end{aligned}$$

$$\stackrel{(ii)}{\leq} \frac{2}{\mu} \left(H_f \|x^k - x^*\|^{1+\nu} + 2\gamma_k^{\frac{1}{2}} \|\nabla f(x^k)\|^{\frac{1}{2}} \|x^k - x^*\| + \|\nabla f(x^k)\|^{\frac{3}{2}} \right), \quad (37)$$

where (i) is by the triangle inequality, and (ii) is by (6) and the definition of ε_k, ζ_k in Algorithm 2. By reusing the step (ii) in the above inequality and repeatedly using (27), we have

$$\begin{aligned} \|\nabla f(x^{k+1})\| &\stackrel{(i)}{\leq} \frac{2Lg}{\mu} \left(H_f \|x^k - x^*\|^{1+\nu} + 2(\gamma_\nu(\|\nabla f(x^k)\|)\|\nabla f(x^k)\|)^{\frac{1}{2}} \|x^k - x^*\| + \|\nabla f(x^k)\|^{\frac{3}{2}} \right) \\ &\leq \frac{2Lg}{\mu} \left(2H_f \mu^{-1} \|x^k - x^*\|^\nu + 4H_f^{\frac{1}{1+\nu}} (2\mu^{-1})^{\frac{1}{1+\nu}} \|x^k - x^*\|^{\frac{\nu}{1+\nu}} + L_g^{\frac{1}{2}} \|x^k - x^*\| \right) \|\nabla f(x^k)\| \\ &\stackrel{(ii)}{\leq} \frac{2Lg}{\mu} \left(2H_f \mu^{-1} + 4H_f^{\frac{1}{1+\nu}} (2\mu^{-1})^{\frac{1}{1+\nu}} + L_g^{\frac{1}{2}} \right) \|x^k - x^*\|^{\frac{\nu}{1+\nu}} \cdot \|\nabla f(x^k)\| \stackrel{(iii)}{\leq} \frac{1}{3} \|\nabla f(x^k)\|, \end{aligned}$$

where (i) uses $\gamma_k \leq \gamma_\nu(\|\nabla f(x^k)\|)$ for $k \in \mathbb{K}_{\varepsilon,2}$, (ii) and (iii) are due to $x^k \in B_\delta(x^*)$. This contradicts with $\|\nabla f(x^{k+1})\| > \frac{1}{2} \|\nabla f(x^k)\|$ for $k \in \mathbb{K}_{\varepsilon,2}$. Thus, when $x^k \in B_\delta(x^*)$, we have $k \notin \mathbb{K}_{\varepsilon,2}$, which implies that $\gamma_{k+1} = \gamma_k$. \square

A.3.6 Proof of Lemma 9.

Proof. Recall δ defined in (36), let $c = 2(1 + \frac{4Lg}{\mu})$, and we define S as

$$S := \left\{ x : \|x - x^*\| \leq \delta, f(x) - f(x^*) \leq \frac{\mu}{4} \left(\frac{\delta}{c} \right)^2 \right\}. \quad (38)$$

Clearly, $S \subseteq B_\delta(x^*)$. By the previous discussion in Appendix A.3.5, we conclude that (27) holds for all $x \in S$, and Algorithm 2 will always execute the Newton step with $\gamma_k = \gamma_{k+1}$ if $x_k \in S$.

We next prove that once $x_k \in S$, all future iterates will stay in S . For any $x^k \in S$, we have

$$\|\nabla f(x^k)\| \stackrel{(i)}{\geq} \frac{\zeta_k \varepsilon_k}{2} \|d^k\| \stackrel{(ii)}{\geq} \|(\nabla^2 f(x^k) + 2\varepsilon_k I)d^k + \nabla f(x^k)\| \stackrel{(iii)}{\geq} \frac{\mu}{2} \|d^k\| - \|\nabla f(x^k)\|,$$

where (i) is by $\zeta_k \leq 1$ and the second inequality of Lemma 12(i), (ii) is by the fourth inequality of Lemma 12(i), and (iii) uses the triangle inequality and (27). This implies that $\|d^k\| \leq \frac{4}{\mu} \|\nabla f(x^k)\|$. And by (27), we have $\frac{\mu}{4} \|x^k - x^*\|^2 \leq f(x^k) - f(x^*) \leq \frac{\mu}{4} \left(\frac{\delta}{c} \right)^2$, which implies that $\|x^k - x^*\| \leq \frac{\delta}{c}$. Thus,

$$\|x^{k+1} - x^*\| = \|x^k + d^k - x^*\| \leq \|x^k - x^*\| + \|d^k\| \stackrel{(i)}{\leq} \left(1 + \frac{4Lg}{\mu} \right) \|x^k - x^*\| < \delta,$$

where (i) uses (27) and $\|d^k\| \leq \frac{4}{\mu} \|\nabla f(x^k)\|$. Besides, since Algorithm 2 is a descent method, one has

$$f(x^{k+1}) - f(x^*) \leq f(x^k) - f(x^*) \leq \frac{\mu}{4} \left(\frac{\delta}{c} \right)^2$$

Thus, $x^{k+1} \in S$. Now suppose $x^{k_0} \in S$ for some $k_0 \geq 0$, by induction we have that $\{x^k\}_{k \geq k_0} \subseteq S$. \square

A.3.7 Proof of Theorem 4.

Proof. By Lemma 8 and Lemma 9, suppose $x^{k_0} \in S$, where S is defined in (38), for some $k_0 \geq 0$, we have that Algorithm 2 will always execute the Newton step, $\{x^k\}_{k \geq k_0} \subseteq S$, and $\gamma_{k+1} = \gamma_k$ for $k \geq k_0$. Therefore, $\gamma_k = \gamma_{k_0}$ for all $k \geq k_0$. Then, by (37) we have

$$\begin{aligned} \|x^{k+1} - x^*\| &\leq \frac{2}{\mu} \left(H_f \|x^k - x^*\|^{1+\nu} + 2\gamma_k^{\frac{1}{2}} \|\nabla f(x^k)\|^{\frac{1}{2}} \|x^k - x^*\| + \|\nabla f(x^k)\|^{\frac{3}{2}} \right) \\ &\leq \frac{2}{\mu} \left(H_f \|x^k - x^*\|^{1+\nu} + (2\gamma_{k_0}^{\frac{1}{2}} + L_g) L_g^{\frac{1}{2}} \|x^k - x^*\|^{\frac{3}{2}} \right). \end{aligned}$$

This shows $\|x^{k+1} - x^*\| = \mathcal{O}(\|x^k - x^*\|^{\min\{1+\nu, \frac{3}{2}\}})$ for $k \geq k_0$, i.e., a superlinear rate of order $\min\{1+\nu, \frac{3}{2}\}$. \square

B Capped conjugate gradient method

In this work, we use the capped CG (`CappedCG`) method [25], presented in Algorithm 3, as a subroutine to solve the possibly *indefinite* damped Newton systems. Consider the linear system $(H + 2\sigma I)d = -g$, where $g \neq 0$ is a nonzero vector in \mathbb{R}^n , $\sigma > 0$, and $H \in \mathbb{S}^{n \times n}$ is a possibly indefinite symmetric matrix. `CappedCG` can efficiently return either (i) an approximate solution d and `d.type=SOL` with well controlled relative error or (ii) a negative curvature direction d and `d.type=NC`. The following lemma captures a few useful properties of `CappedCG`, which are adopted from [25, Lemma 3].

Lemma 12. *Let $(d, d_type) = \text{CappedCG}(H, g, \zeta, \sigma)$, then the following statements hold:*

(i) *If $d_type=\text{SOL}$, then d satisfies*

$$\begin{aligned} \sigma \|d\|^2 &\leq d^\top (H + 2\sigma I)d, & \|d\| &\leq 1.1\sigma^{-1}\|g\|, \\ d^\top g &= -d^\top (H + 2\sigma I)d, & \|(H + 2\sigma I)d + g\| &\leq \zeta\sigma\|d\|/2. \end{aligned}$$

(ii) *If $d_type=\text{NC}$, then d satisfies $d^\top g \leq 0$ and $d^\top H d / \|d\|^2 < -\sigma$.*

Algorithm 3: Capped conjugate gradient method $(d, d_type) = \text{CappedCG}(H, g, \zeta, \sigma)$

```

1 Input: symmetric matrix  $H \in \mathbb{R}^{n \times n}$ , vector  $g \neq 0$ , damping parameter  $\sigma > 0$ , desired relative accuracy
    $\zeta \in (0, 1)$ .
2 Output:  $(d, d\_type)$ . // d.type=SOL: solved, d.type=NC: negative curvature
3 Initialize:
    $U \leftarrow 0, \bar{H} \leftarrow H + 2\sigma I, \kappa \leftarrow \frac{U+2\sigma}{\sigma}, \hat{\zeta} \leftarrow \frac{\zeta}{3\kappa}, \tau \leftarrow \frac{\sqrt{\kappa}}{\sqrt{\kappa+1}}, T \leftarrow \frac{4\kappa^4}{(1-\sqrt{\tau})^2}, y^0 \leftarrow 0, r^0 \leftarrow g, p^0 \leftarrow -g, j \leftarrow 0$ .
4 if  $(p^0)^\top \bar{H} p^0 < \sigma \|p^0\|^2$  then return  $(p^0, \text{NC})$ .
5 while TRUE do
6   Compute  $\alpha_j \leftarrow (r^j)^\top r^j / (p^j)^\top \bar{H} p^j$ ,  $y^{j+1} \leftarrow y^j + \alpha_j p^j$ , and  $r^{j+1} \leftarrow r^j + \alpha_j \bar{H} p^j$ ; // Begin Standard CG
7   Compute  $\beta_{j+1} \leftarrow \|r^{j+1}\|^2 / \|r^j\|^2$ , and update  $p^{j+1} \leftarrow -r^{j+1} + \beta_{j+1} p^j$ ; // End Standard CG
8   Set  $j \leftarrow j + 1$ , and update  $U \leftarrow \max\{U, \|H p^0\| / \|p^0\|, \|H p^j\| / \|p^j\|, \|H y^j\| / \|y^j\|, \|H r^j\| / \|r^j\|\}$ ;
9   Update  $\kappa, \hat{\zeta}, \tau, T$  by Line 3 accordingly.
10  if  $(y^j)^\top \bar{H} y^j < \sigma \|y^j\|^2$  then return  $(y^j, \text{NC})$ .
11  else if  $\|r^j\| \leq \hat{\zeta} \|r^0\|$  then return  $(y^j, \text{SOL})$ .
12  else if  $(p^j)^\top \bar{H} p^j < \sigma \|p^j\|^2$  then return  $(p^j, \text{NC})$ .
13  else if  $\|r^j\| > \sqrt{T} \tau^{j/2} \|r^0\|$  then
14    Compute  $\alpha_j, y^{j+1}$  as in the main loop above;
15    Find  $i \in \{0, \dots, j-1\}$  such that  $(y^{j+1} - y^i)^\top \bar{H} (y^{j+1} - y^i) < \sigma \|y^{j+1} - y^i\|^2$ ;
16    return  $(y^{j+1} - y^i, \text{NC})$ .

```
