A Note on Inexact Gradient and Hessian Conditions for Cubic Regularized Newton's Method

Zhe Wang[a,1,*], Yi Zhou[b], Yingbin Liang[a], Guanghui Lan[c]

[a] *Department of Electrical and Computer Engineering, Ohio State University, Columbus, OH, 43210, USA*
[b] *Department of Electrical and Computer Engineering, Duke University, Durham, NC, 27708, USA*
[c] *Department of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 30332, USA*

Abstract

The inexact cubic-regularized Newton's method (CR) proposed by Cartis, Gould and Toint achieves the same convergence rate as exact CR proposed by Nesterov and Polyak, but the inexact condition is not implementable due to its dependence on a future variable. This note establishes the same convergence rate under a similar but implementable inexact condition, which depends on only current variables. Our proof bounds the function-value decrease over total iterations rather than each iteration in the previous studies.

Keywords: Nonconvex, Second-order Methods, Second-order Stationary Points.

1. Introduction

The cubic-regularized (CR) Newton's method [1] is a popular approach that solves the following general nonconvex optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}), \tag{1}$$

where $f:\mathbb{R}^d\to\mathbb{R}$ is a differentiable and possibly nonconvex function. Starting from an arbitrary initial point $\mathbf{x}_0$, the update rule of CR can be written as

(CR):

$$\mathbf{s}_{k+1} = \operatorname*{argmin}_{\mathbf{s}\in\mathbb{R}^d} \nabla f(\mathbf{x}_k)^\top \mathbf{s} + \frac{1}{2}\mathbf{s}^\top \nabla^2 f(\mathbf{x}_k)\mathbf{s} + \frac{M}{6}\|\mathbf{s}\|^3,$$
$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{s}_{k+1}, \tag{2}$$

where M is a positive scalar. [1] showed that CR converges to a second-order stationary point $\mathbf{x}$ of the objective function, i.e.,

$$\nabla f(\mathbf{x}) = \mathbf{0} \quad \text{and} \quad \nabla^2 f(\mathbf{x}) \succcurlyeq \mathbf{0}. \tag{3}$$

Such a desirable property allows CR to escape strict saddle points. However, CR needs to compute a full Hessian at each iteration, and is hence computationally intensive. Thus, [2, 3] proposed an algorithm named Adaptive Regularization using Cubics (ARC) that uses an adaptive regularization scheme as well as an inexact sub-problem solver to reduce the computation complexity. More specifically, [2, 3] proposed to use an inexact approximation $\mathbf{H}_k$ to replace the full Hessian $\nabla^2 f(\mathbf{x}_k)$ in the CR update in order to be computationally more efficient, leading to the following inexact CR algorithm

(Inexact CR):

$$\mathbf{s}_{k+1} = \operatorname*{argmin}_{\mathbf{s}\in\mathbb{R}^d} \nabla f(\mathbf{x}_k)^\top \mathbf{s} + \frac{1}{2}\mathbf{s}^\top \mathbf{H}_k\mathbf{s} + \frac{M}{6}\|\mathbf{s}\|^3, \tag{4}$$
$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{s}_{k+1}. \tag{5}$$

[2, 3] showed that if $\mathbf{H}_k$ satisfies the following inexact condition,

$$\|(\mathbf{H}_k - \nabla^2 f(\mathbf{x}_k))\mathbf{s}_{k+1}\| \leqslant C\|\mathbf{s}_{k+1}\|^2, \tag{6}$$

where $C \geqslant 0$. Then ARC with inexact Hessian achieves the same order-level convergence rate to a second-order

*Corresponding author
Email addresses: wang.10982@osu.edu (Zhe Wang), yi.zhou610@duke.edu (Yi Zhou), liang.889@osu.edu (Yingbin Liang), george.lan@isye.gatech.edu (Guanghui Lan)

stationary point as exact CR for nonconvex optimization.

The inexact CR has been further explored in various situations [4, 5, 6, 7]. However, observe that the above inexact condition involves $\|\mathbf{s}_{k+1}\|$ (and hence $\mathbf{x}_{k+1}$), which is not available at iteration k. Thus, such a condition is not practically implementable. More recent research studies [4, 8] used $\mathbf{s}_k$ to implement inexact CR numerically as follows

$$\|\mathbf{H}_k - \nabla^2 f(\mathbf{x}_k)\| \leqslant C\|\mathbf{s}_k\|. \tag{7}$$

These studies demonstrated that inexact CR still performs well in experiments under (7), but did not provide theoretical convergence guarantee of inexact CR under such a condition.

The main contribution of this note is the establishment of convergence guarantee for inexact CR under the implementable condition (7), and furthermore under a similar inexact condition for gradient (see (12) below). We show that inexact CR under these conditions achieves the same order of convergence rate as the exact CR, i.e., the algorithm passes an ϵ approximate second-order stationary point within $\mathcal{O}(\epsilon^{-3/2})$ iterations. In contrast to existing proof techniques, our proof relies on an idea of the control of the sufficient decrease of the function value over all iterations rather than requiring a sufficient decrease at each iteration. More specifically, the inexact error $\|\mathbf{H}_k - \nabla^2 f(\mathbf{x}_k)\| \leqslant C\|\mathbf{s}_k\|$ at current iteration is incorporated into the bound on the previous iteration, which yields a successful analysis over all iterations under a more relaxed (and practical) condition (see (7)). We note that our theory guarantees that the algorithm must pass an ϵ approximate second-order stationary point within $\mathcal{O}(\epsilon^{-3/2})$ iterations. Although our result does not necessarily guarantee the final iterate at the termination to satisfy the second-order stationary optimality condition, such an issue can be solved by incorporating a breaking step to check the condition $\max\{\|\mathbf{s}_k\|, \|\mathbf{s}_{k+1}\|\} \leqslant \mathcal{O}(\sqrt{\epsilon})$, which implies $\|\nabla f(\mathbf{x}_{k+1})\| = C\epsilon$ and $\nabla^2 f(\mathbf{x}_{k+1}) \succcurlyeq -C\sqrt{\epsilon}$.

We also note that another inexact condition has been proposed by [9], which takes the form

$$\|(\mathbf{H}_k - \nabla^2 f(\mathbf{x}_k))\mathbf{s}_{k+1}\| \leqslant \epsilon\|\mathbf{s}_{k+1}\|, \tag{8}$$

where ϵ is a pre-defined small constant and is related to the required accuracy. Similar conditions have been used in [10, 11, 12, 13, 14]. Compared to (8), (7) is made to be adaptive to $\|\mathbf{s}_k\|$ so that more progress towards the convergence point can be made during the most phase of the algorithm when the increment $\|s_k\|$ is larger than ϵ.

We note that our focus here is on the implementability of the inexact Hessian condition, and the results are obtained using a fixed regularization, which depends on the knowledge of the Lipschitz constant L, as well as the exact solution to the sub-problem. These issues have been addressed in [2, 3, 9] and several others by employing adaptive regularization as well as inexact sub-problem solvers (in addition to the inexact Hessian conditions as aforementioned). In fact, our algorithm can further be made adaptive by incorporating the idea in [2, 3, 9] to estimate L and can be shown to have the same convergence rate.

Notation: For a vector $\mathbf{x}$, $\|\mathbf{x}\|$ denotes its ℓ_2 norm. For a matrix $\mathbf{H}$, $\|\mathbf{H}\|$ denotes its spectral norm. We let $\mathbf{I}$ denote the identity matrix. For a function $f : \mathbb{R}^d \to \mathbb{R}$, ∇f and $\nabla^2 f$ denote its gradient and Hessian, respectively. $\mathbb{R}$, $\mathbb{R}^+$ and $\mathbb{R}^d$ denote the set of all real numbers, non-negative real numbers and d-dimensional real vectors, respectively. The symbol $\mathbb{S}$ denotes the set of all symmetric matrices.

2. Main Result

Our analysis adopts the following standard assumption as in the previous studies of CR.

Assumption 1. *The objective function in eq. (1) satisfies:*

1. *f is twice-continuously differentiable and bounded below, i.e., $f^* \triangleq \inf_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) > -\infty$;*
2. *The Hessian $\nabla^2 f$ is L-Lipschitz continuous.*

In Assumption 1, we assume that the Hessian is Lipschitz continuous the Lipschitz parameter L is known a priori. We note that such a parameter can be estimated via adaptive line-search methods in practice [2, 3].

In our analysis, we allow both the gradient and the Hessian to be replaced by their approximations, and hence the CR iterate becomes

(Inexact gradient and Hessian CR):

$$\mathbf{s}_{k+1} = \operatorname*{argmin}_{\mathbf{s}\in\mathbb{R}^d} \mathbf{g}_k^\top \mathbf{s} + \frac{1}{2}\mathbf{s}^\top \mathbf{H}_k \mathbf{s} + \frac{M}{6}\|\mathbf{s}\|^3, \tag{9}$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{s}_{k+1}. \tag{10}$$

We assume that the approximations $\mathbf{g}_k$ and $\mathbf{H}_k$ satisfy the following inexact conditions, which depend on current information only, and are hence implementable.

Assumption 2. *There exist two constants $\alpha, \beta \in \mathbb{R}^+$, such that the inexact gradient $\mathbf{g}_k$ and inexact Hessian $\mathbf{H}_k$ satisfy, for all $k \geqslant 0$,*

$$\|\mathbf{H}_k - \nabla^2 f(\mathbf{x}_k)\| \leqslant \alpha \|\mathbf{s}_k\|, \tag{11}$$

$$\|\mathbf{g}_k - \nabla f(\mathbf{x}_k)\| \leqslant \beta \|\mathbf{s}_k\|^2. \tag{12}$$

We note that Assumption 2 can be modified as $\|\mathbf{H}_k - \nabla^2 f(\mathbf{x}_k)\| \leqslant \alpha \max\{\|\mathbf{s}_k\|, \epsilon\}$ in order for the algorithm to perform better if $\|\mathbf{s}_k\| \leqslant \epsilon$ (so that the overall performance can be improved). Similar modification can be applied to the gradient as well.

In our analysis, we assume the cubic subproblem in eq. (9) can be solved exactly. This is to simplify the analysis and focus on the inexact conditions in Assumption 2, and we refer to [2, 3, 15, 9] for the analysis of CR under inexact sub-problem solvers. Next, we state our main theorem on convergence of CR under the inexact conditions in Assumption 2.

Theorem 1. *Let Assumptions 1 and 2 hold. Then, after k iterations, the sequence $\{\mathbf{x}_i\}_{i\geqslant 1}$ generated by inexact CR contains a point $\tilde{\mathbf{x}}$ such that*

$$\|\nabla f(\tilde{\mathbf{x}})\| \leqslant \frac{C_1}{(k-1)^{2/3}} \quad \text{and} \quad \nabla^2 f(\tilde{\mathbf{x}}) \succcurlyeq -\frac{C_2}{(k-1)^{1/3}}\mathbf{I}.$$

where $k > 1$, and C_1 and C_2 are universal constants, and are specified in the proof.

Theorem 1 guarantees that after k iterations, inexact CR must pass an approximate second order stationary point with error within $O(1/k^{2/3})$ and $O(1/k^{1/3})$ for the gradient and Hessian, respectively, under the inexact conditions in Assumption 2. The proof of Theorem 1 is based on the following two useful lemmas.

Lemma 2 ([1], Lemma 1)**.** *Let the Hessian $\nabla^2 f$ of the function f be L-Lipschitz continuous with $L > 0$. Then, for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have*

$$\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}) - \nabla^2 f(\mathbf{x})(\mathbf{y}-\mathbf{x})\| \leqslant \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|^2, \tag{13}$$

$$\Big| f(\mathbf{y}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) - \frac{1}{2}(\mathbf{y}-\mathbf{x})^T \nabla^2 f(\mathbf{x})(\mathbf{y}-\mathbf{x}) \Big| \leqslant \frac{L}{6}\|\mathbf{y}-\mathbf{x}\|^3. \tag{14}$$

We then establish Lemma 3, which provides the properties of the minimizer of (9) for a more general setting.

Lemma 3. *Let $M > 0, \mathbf{g} \in \mathbb{R}^d, \mathbf{H} \in \mathbb{S}^{d\times d}$, and*

$$\mathbf{s} = \operatorname*{argmin}_{\mathbf{u}\in\mathbb{R}^d} \mathbf{g}^\top \mathbf{u} + \frac{1}{2}\mathbf{u}^\top \mathbf{H}\mathbf{u} + \frac{M}{6}\|\mathbf{u}\|^3. \tag{15}$$

Then, the following statements hold:

$$\mathbf{g} + \mathbf{H}\mathbf{s} + \frac{M}{2}\|\mathbf{s}\|\mathbf{s} = \mathbf{0}, \tag{16}$$

$$\mathbf{H} + \frac{M}{2}\|\mathbf{s}\|\mathbf{I} \succcurlyeq \mathbf{0}, \tag{17}$$

$$\mathbf{g}^\top \mathbf{s} + \frac{1}{2}\mathbf{s}^\top \mathbf{H}\mathbf{s} + \frac{M}{6}\|\mathbf{s}\|^3 \leqslant -\frac{M}{12}\|\mathbf{s}\|^3. \tag{18}$$

To further explain, (16) corresponds to the first-order necessary optimality condition, (17) corresponds to the second-order necessary optimality condition but with a tighter form due to the specific form of this optimization problem, and (18) guarantees a sufficient decrease at this minimizer.

Proof of Lemma 3. First, (16) follows from the first-order necessary optimality condition of (15), and (17) follows

from Proposition 1 in [1]. We next prove (18). Following similar steps to those in [1], we obtain that

$$\begin{aligned}
&\mathbf{g}^\top \mathbf{s} + \frac{1}{2}\mathbf{s}^\top \mathbf{H}\mathbf{s} + \frac{M}{6}\|\mathbf{s}\|^3 \\
&\quad \overset{\text{(i)}}{=} \left(-\mathbf{H}\mathbf{s} - \frac{M}{2}\|\mathbf{s}\|\mathbf{s}\right)^\top \mathbf{s} + \frac{1}{2}\mathbf{s}^\top \mathbf{H}\mathbf{s} + \frac{M}{6}\|\mathbf{s}\|^3 \\
&\quad = -\frac{1}{2}\mathbf{s}^\top \left(\mathbf{H} + \frac{M}{2}\|\mathbf{s}\|\mathbf{I}\right)\mathbf{s} - \frac{M}{12}\|\mathbf{s}\|^3 \\
&\quad \overset{\text{(ii)}}{\leqslant} -\frac{M}{12}\|\mathbf{s}\|^3,
\end{aligned}$$

where (i) follows from (16), and (ii) follows from (17), which implies that $-\frac{1}{2}\mathbf{s}^\top \left(\mathbf{H} + \frac{M}{2}\|\mathbf{s}\|\mathbf{I}\right)\mathbf{s} \leqslant 0$. □

Now, we are ready to prove our main theorem.

Proof of Theorem 1. Consider any iteration k, we obtain that

$$\begin{aligned}
&f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \\
&\quad \overset{\text{(i)}}{\leqslant} \nabla f(\mathbf{x}_k)^\top \mathbf{s}_{k+1} + \frac{1}{2}\mathbf{s}_{k+1}^\top \nabla^2 f(\mathbf{x}_k)\mathbf{s}_{k+1} + \frac{L}{6}\|\mathbf{s}_{k+1}\|^3 \\
&\quad = \mathbf{g}_k^\top \mathbf{s}_{k+1} + \frac{1}{2}\mathbf{s}_{k+1}^\top \mathbf{H}_k \mathbf{s}_{k+1} + \frac{M}{6}\|\mathbf{s}_{k+1}\|^3 \\
&\qquad + (\nabla f(\mathbf{x}_k) - \mathbf{g}_k)^\top \mathbf{s}_{k+1} + \frac{L-M}{6}\|\mathbf{s}_{k+1}\|^3 \\
&\qquad + \frac{1}{2}\mathbf{s}_{k+1}^\top (\nabla^2 f(\mathbf{x}_k) - \mathbf{H}_k)\mathbf{s}_{k+1} \\
&\quad \overset{\text{(ii)}}{\leqslant} -\frac{3M-2L}{12}\|\mathbf{s}_{k+1}\|^3 + (\nabla f(\mathbf{x}_k) - \mathbf{g}_k)^\top \mathbf{s}_{k+1} \\
&\qquad + \frac{1}{2}\mathbf{s}_{k+1}^\top (\nabla f(\mathbf{x}_k) - \mathbf{H}_k)\mathbf{s}_{k+1} \\
&\quad \overset{\text{(iii)}}{\leqslant} -\frac{3M-2L}{12}\|\mathbf{s}_{k+1}\|^3 + \beta\|\mathbf{s}_k\|^2\|\mathbf{s}_{k+1}\| \\
&\qquad + \alpha\|\mathbf{s}_k\|\|\mathbf{s}_{k+1}\|^2 \\
&\quad \overset{\text{(vi)}}{\leqslant} -\frac{3M-2L}{12}\|\mathbf{s}_{k+1}\|^3 + \beta(\|\mathbf{s}_k\|^3 + \|\mathbf{s}_{k+1}\|^3) \\
&\qquad + \alpha(\|\mathbf{s}_k\|^3 + \|\mathbf{s}_{k+1}\|^3) \\
&\quad = -\left(\frac{3M-2L}{12} - \alpha - \beta\right)\|\mathbf{s}_{k+1}\|^3 + (\alpha+\beta)\|\mathbf{s}_k\|^3.
\end{aligned} \tag{19}$$

where (i) follows from Lemma 2 with $\mathbf{y} = \mathbf{x}_{k+1}, \mathbf{x} = \mathbf{x}_k$ and $\mathbf{s}_{k+1} = \mathbf{x}_{k+1} - \mathbf{x}_k$, (ii) follows from (18) in Lemma 3 with $\mathbf{g} = \mathbf{g}_k, \mathbf{H} = \mathbf{H}_k$ and $\mathbf{s} = \mathbf{s}_{k+1}$, (iii) follows from Assumption 2, and (vi) follows from the inequality that for $a, b \in \mathbb{R}^+$, $a^2 b \leqslant a^3 + b^3$, which can be verified by checking the cases with $a < b$ and $a \geqslant b$, respectively.

Summing (19) from 0 to $k-1$, we obtain that

$$\begin{aligned}
&f(\mathbf{x}_k) - f(\mathbf{x}_0) \\
&\leqslant -\sum_{i=0}^{k-1}\left(\frac{3M-2L}{12} - \alpha - \beta\right)\|\mathbf{s}_{i+1}\|^3 + \sum_{i=0}^{k-1}(\alpha+\beta)\|\mathbf{s}_i\|^3 \\
&\leqslant -\sum_{i=1}^{k}\left(\frac{3M-2L}{12} - \alpha - \beta\right)\|\mathbf{s}_i\|^3 + \sum_{i=0}^{k}(\alpha+\beta)\|\mathbf{s}_i\|^3 \\
&= -\sum_{i=1}^{k}\left(\frac{3M-2L}{12} - 2\alpha - 2\beta\right)\|\mathbf{s}_i\|^3 + (\alpha+\beta)\|\mathbf{s}_0\|^3.
\end{aligned}$$

Therefore, we have

$$\sum_{i=1}^{k}\gamma\|\mathbf{s}_i\|^3 \leqslant f(\mathbf{x}_0) - f^* + (\alpha+\beta)\|\mathbf{s}_0\|^3, \tag{20}$$

where $\gamma \triangleq \frac{3M-2L}{12} - 2\alpha - 2\beta$. We note that we set $M > \frac{2}{3}L + 8\alpha + 8\beta$ to have $\gamma > 0$, which is needed to conclude that $\sum_{i=1}^{k}\|\mathbf{s}_i\|^3$ is upper bounded from (20). One way to satisfy the requirement of M is to adopt a similar adaptive scheme proposed by [2, 3].

Let $m \triangleq \operatorname{argmin}_{i\in\{1,\cdots,k-1\}}\|\mathbf{s}_i\|^3 + \|\mathbf{s}_{i+1}\|^3$. We obtain that

$$\begin{aligned}
&\|\mathbf{s}_m\|^3 + \|\mathbf{s}_{m+1}\|^3 \\
&\quad = \min_{i\in\{1,\cdots,k-1\}}\|\mathbf{s}_i\|^3 + \|\mathbf{s}_{i+1}\|^3 \\
&\quad \leqslant \frac{1}{k-1}\sum_{i=1}^{k-1}\left(\|\mathbf{s}_i\|^3 + \|\mathbf{s}_{i+1}\|^3\right) \\
&\quad \overset{\text{(i)}}{\leqslant} \frac{2}{\gamma(k-1)}\left(f(\mathbf{x}_0) - f^* + (\alpha+\beta)\|\mathbf{s}_0\|^3\right).
\end{aligned}$$

where (i) follows (20).

Therefore, we have

$$\begin{aligned}
&\max\{\|\mathbf{s}_m\|, \|\mathbf{s}_{m+1}\|\} \\
&\leqslant \frac{1}{(k-1)^{1/3}}\left(\frac{2}{\gamma}\left(f(\mathbf{x}_0) - f^* + (\alpha+\beta)\|\mathbf{s}_0\|^3\right)\right)^{1/3}.
\end{aligned} \tag{21}$$

Next, we prove the convergence rate of ∇f and $\nabla^2 f$. We first derive

$$\begin{aligned}
&\|\nabla f(\mathbf{x}_{m+1})\| \\
&\overset{\text{(i)}}{=} \left\|\nabla f(\mathbf{x}_{m+1}) - \left(\mathbf{g}_m + \mathbf{H}_m\mathbf{s}_{m+1} + \frac{M}{2}\|\mathbf{s}_{m+1}\|\mathbf{s}_{m+1}\right)\right\|
\end{aligned}$$

$$
\begin{aligned}
&\leqslant \|\nabla f(\mathbf{x}_{m+1}) - (\mathbf{g}_m + \mathbf{H}_m \mathbf{s}_{m+1})\| + \frac{M}{2}\|\mathbf{s}_{m+1}\|^2 \\
&\leqslant \left\|\nabla f(\mathbf{x}_{m+1}) - \nabla f(\mathbf{x}_m) - \nabla^2 f(\mathbf{x}_m)\mathbf{s}_{m+1}\right\| \\
&\quad + \|\nabla f(\mathbf{x}_m) - \mathbf{g}_m\| + \|(\nabla^2 f(\mathbf{x}_m) - \mathbf{H}_m)\mathbf{s}_{m+1}\| \\
&\quad + \frac{M}{2}\|\mathbf{s}_{m+1}\|^2 \\
&\overset{\text{(ii)}}{\leqslant} \frac{L}{2}\|\mathbf{s}_{m+1}\|^2 + \beta\|\mathbf{s}_m\|^2 + \alpha\|\mathbf{s}_m\|\|\mathbf{s}_{m+1}\| + \frac{M}{2}\|\mathbf{s}_{m+1}\|^2 \\
&\overset{\text{(iii)}}{\leqslant} \frac{C_1}{(k-1)^{2/3}},
\end{aligned}
$$

where (i) follows from (16) with $\mathbf{g} = \mathbf{g}_m, \mathbf{H} = \mathbf{H}_m$ and $\mathbf{s} = \mathbf{s}_{m+1}$, (ii) follows from (13) in Lemma 2 and Assumption 2, and (iii) follows from (21) and the definition that $C_1 \triangleq \frac{L+M+2\beta+2\alpha}{2}\left(\frac{2}{\gamma}\left(f(\mathbf{x}_0) - f^* + (\alpha+\beta)\|\mathbf{s}_0\|^3\right)\right)^{2/3}$.

We next prove the the convergence rate of $\nabla^2 f(\cdot)$.

$$
\begin{aligned}
\nabla^2 f(\mathbf{x}_{m+1}) &\overset{(i)}{\succcurlyeq} \mathbf{H}_m - \|\mathbf{H}_m - \nabla^2 f(\mathbf{x}_{m+1})\|\mathbf{I} \\
&\overset{(ii)}{\succcurlyeq} -\frac{M}{2}\|\mathbf{s}_{m+1}\|\mathbf{I} - \|\mathbf{H}_m - \nabla^2 f(\mathbf{x}_{m+1})\|\mathbf{I} \\
&\succcurlyeq -\frac{M}{2}\|\mathbf{s}_{m+1}\|\mathbf{I} - \|\mathbf{H}_m - \nabla^2 f(\mathbf{x}_m)\|\mathbf{I} \\
&\quad - \|\nabla^2 f(\mathbf{x}_m) - \nabla^2 f(\mathbf{x}_{m+1})\|\mathbf{I} \\
&\overset{(iii)}{\succcurlyeq} -\frac{M}{2}\|\mathbf{s}_{m+1}\|\mathbf{I} - \alpha\|\mathbf{s}_m\|\mathbf{I} - L\|\mathbf{s}_{m+1}\|\mathbf{I} \\
&\overset{(iv)}{\succcurlyeq} -\frac{C_2}{(k-1)^{1/3}}\mathbf{I},
\end{aligned}
$$

where (i) follows from Weyl's inequality, (ii) follows from (17) with $\mathbf{H} = \mathbf{H}_m$ and $\mathbf{s} = \mathbf{s}_{m+1}$, (iii) follows from Assumption 2, the fact that $\nabla^2 f$ is $L-$Lipschitz and the definition of $\mathbf{s}_{m+1}$, and (iv) follows from the definition that $C_2 \triangleq \frac{M+2L+2\alpha}{2}\left(\frac{2}{\gamma}\left(f(\mathbf{x}_0) - f^* + (\alpha+\beta)\|\mathbf{s}_0\|^3\right)\right)^{1/3}$, and (21) . □

Acknowledgement

The work of Z. Wang, Y. Zhou and Y. Liang was supported in part by U.S. National Science Foundation under the grant CCF-1761506, and the work of G. Lan was supported in part by Army Research Office under the grant W911NF-18-1-0223 and National Science Foundation under the grant CMMI-1254446.

References

[1] Y. Nesterov, B. T. Polyak, Cubic regularization of newton method and its global performance, Mathematical Programming 108 (1) (2006) 177–205.

[2] C. Cartis, N. I. M. Gould, P. L. Toint, Adaptive cubic regularization methods for unconstrained optimization. Part I : Motivation, convergence and numerical results, Mathematical Programming 127 (2) (2011) 245–295.

[3] C. Cartis, N. I. M. Gould, P. L. Toint, Adaptive cubic regularization methods for unconstrained optimization. Part II worst-case function- and derivative-evaluation complexity, Mathematical Programming 130 (2) (2011) 295–319.

[4] J. M. Kohler, A. Lucchi, Sub-sampled cubic regularization for non-convex optimization, in: Proc. 34th International Conference on Machine Learning (ICML), Vol. 70, 2017, pp. 1895–1904.

[5] C. Cartis, N. I. M. Gould, P. L. Toint, An adaptive cubic regularization algorithm for nonconvex optimization with convex constraints and its function-evaluation complexity, IMA Journal of Numerical Analysis 32 (4) (2012) 1662 – 1695.

[6] C. Cartis, N. I. M. Gould, P. L. Toint, Complexity bounds for second-order optimality in unconstrained optimization, Journal of Complexity 28 (1) (2012) 93 – 108.

[7] Y. Zhou, Z. Wang, Y. Liang, Convergence of cubic regularization for nonconvex optimization under KL property, in: Proc. 32nd Advances in Neural Information Processing Systems (NeurIPS), 2018.

[8] Z. Wang, Y. Zhou, Y. Liang, G. Lan, Sample complexity of stochastic variance-reduced cubic regularization for nonconvex optimization, in: Proc. 22nd International Conference on Artificial Intelligence and Statistics (AISTATS), 2018.

[9] P. Xu, F. Roosta-Khorasani, M. W. Mahoney, Newton-type methods for non-convex optimization under inexact hessian information, arXiv: 1708.07164.

[10] S. Ghadimi, H. Liu, T. Zhang, Second-order methods with cubic regularization under inexact information, arXiv: 1710.05782.

[11] B. Jiang, T. Lin, S. Zhang, A unified scheme to accelerate adaptive cubic regularization and gradient methods for convex optimization, arXiv:1710.04788.

[12] N. Tripuraneni, M. Stern, C. Jin, J. Regier, M. I. Jordan, Stochastic cubic regularization for fast nonconvex optimization, arXiv: 1711.02838.

[13] Z. Yao, P. Xu, F. Roosta-Khorasani, M. W. Mahoney, Inexact non-convex Newton-type methods, arXiv:1802.06925.

[14] Z. Wang, Y. Zhou, Y. Liang, G. Lan, Cubic regularization with momentum for nonconvex optimization, arXiv:1810.03763.

[15] D. Zhou, P. Xu, Q. Gu, Stochastic variance-reduced cubic regu-

larized newton method, in: Proc. 35th International Conference on Machine Learning (ICML), 2018.