On the convergence and complexity of the cubic Newton method

Florentin Goyens [1] and Geovani N. Grapiglia[1]

[1]ICTEAM Institute, UCLouvain, Louvain-la-Neuve, Belgium
{florentin.goyens,geovani.grapiglia}@uclouvain.be

November 2025

1 Introduction

Consider the unconstrained problem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad f(x), \tag{1.1}$$

where $f\colon \mathbb{R}^n \to \mathbb{R}$ is smooth and possibly nonconvex. A rich arsenal of iterative methods exists for this task, ranging from first-order methods (which use only gradient information) to second-order methods (which exploit Hessian information). Newton's method is the classic second-order algorithm, it minimizes a quadratic approximation of the function f, typically achieving rapid local convergence near a solution. However, vanilla Newton steps can be unstable or undefined when the Hessian is singular or not positive-definite, and the method may diverge if started far from a minimizer.

To overcome these issues, researchers have developed modified second-order methods that ensure global convergence. Two prominent paradigms are trust-region methods and regularization methods. In a trust-region method, one restricts each step to a region where the quadratic model is trusted (preventing steps that are too large), while in regularization methods one modifies the model to tame its behavior.

These globablisation strategies allow to show that the sequence of iterates asymptotically converges to a stationary point of f, that is, a point where $\nabla f(x)$ is zero. However, in practice, algorithms terminate when the gradient norm is small, that is, they target a point where $\|\nabla f(x)\| \le \varepsilon$ for some $\varepsilon > 0$. It is therefore desireable to quantify how many iterations of an algorithm are necessary to ensure that ε-stationarity is reached.

For example, on functions with L_1-Lipschitz continuous gradients, the gradient descent method with stepsize $1/L_1$ requires at most $\mathcal{O}\left(L_1\varepsilon^{-2}\right)$ iterations to reach an ε-stationary point (Nesterov, 2004). Second-order methods can outperform this. In their seminal work, Nesterov and Polyak (2006) introduced the cubic-regularized Newton method for functions with L_2-Lipschitz continuous Hessians, i.e.,

$$\left\|\nabla^2 f(x) - \nabla^2 f(y)\right\| \le L_2 \|x - y\| \quad \text{for all } x, y \in \mathbb{R}^n, \tag{1.2}$$

and proved that it requires at most $\mathcal{O}\left(L_2^{1/2}\varepsilon^{-3/2}\right)$ iterations to find an ε-approximate critical point. The original cubic Newton method requires to know the Lipschitz constant L_2, but Cartis et al. (2010b) proposed an adaptive version in which a regularization parameter acts as an estimation of the Lipschitz constant of the Hessian, with similar complexity guarantees.

Unfortunately, the assumption that the Hessian of f is Lipschitz continuous excludes many smooth functions f of interest. The assumption can be relaxed somewhat: it is enough that the Hessian of f be Lipschitz continuous on a convex set containing the iterates $\{x_k\}$ and trial points $\{x_k + s_k\}$ visited by the algorithm. However, it is in general not possible to guarantee a priori that the trial points belong to a specific region. In particular, since they may increase the objective function, trial points may lie outside the sublevelset of f corresponding to the initial iterate $x_0 \in \mathbb{R}^n$,

$$\mathcal{L}_f(x_0) := \{x \in \mathbb{R}^n | f(x) \leq f(x_0)\}. \tag{1.3}$$

In this work, we show a $\mathcal{O}\left(L_2^{1/2}\varepsilon^{-3/2}\right)$ iteration complexity bound for the adaptive cubic Newton method without assuming that the Hessian of f is Lipschitz continuous on a particular set. Instead, we merely assume that f is three times continuously differentiable and that the sublevelset $\mathcal{L}_f(x_0)$ is bounded (Theorem 3.5). The method that we analyze allows for inexact subproblem minimization (Algorithm 1).

2 The cubic Newton method

The second-order Taylor polynomial of f at $x \in \mathbb{R}^n$ is given by

$$T_x^{f,2}(y) := f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2}\left\langle \nabla^2 f(x)(y-x), y-x \right\rangle. \tag{2.1}$$

If the Hessian matrix $\nabla^2 f(x)$ is positive definite, the minimizer of the second-order Taylor approximation is given by

$$y = x - \nabla^2 f(x)^{-1}\nabla f(x).$$

This iteration is the Newton method for minimizing f. It has two main drawbacks: if f is nonconvex, the Hessian may not be positive definite; and the Newton method requires a globablization strategy in order to converge from an arbitrary initial point, which could be far away from the solution. If f is three times continuously differentiable and that the Hessian is L_2-Lipschitz continuous (1.2), then $\left\|\mathrm{D}^3 f(x)\right\| \leq L_2$ for all $x \in \mathbb{R}^n$ and Lemma 3.2 gives the cubic upper bound

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2}\left\langle \nabla^2 f(x)(y-x), y-x \right\rangle + \frac{L}{6}\|y-x\|^3,$$

for all $y \in \mathbb{R}^n$. The original cubic Newton method from Nesterov and Polyak (2006) minimizes this upper bound of f which relies on the constant L_2. When the constant is unknown, it makes sense to estimate it on the fly with an adaptive parameter, as introduced by Cartis et al. (2010a). At $x \in \mathbb{R}^n$, the regularized model for some $\sigma > 0$ is

$$m_{x,\sigma}(y) = f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2}\left\langle \nabla^2 f(x)(y-x), y-x \right\rangle + \frac{\sigma}{6}\|y-x\|^3. \tag{2.2}$$

Algorithm 1 describes the adaptive cubic Newton method with inexact subproblem minimization. Note that, at iteration k, the point x_{k+1} is obtained with the regularization parameter $2^{i_k}\sigma_k$, where $i_k + 1$ is the number of trial points computed at iteration k before accepting the candidate step.

Algorithm 1 Adaptive and inexact cubic Newton method

```
Step 0: Initialization Tolerance $\varepsilon > 0$, regularization $\sigma_0 \geq \sigma_{\min} > 0$ and $x_0 \in \mathbb{R}^n$.
Set $k \leftarrow 0$.
Step 1: Test for termination
If $\|\nabla f(x_k)\| \leq \varepsilon$, terminate and return $x_k$.
Step 2: Step computation
for $i = 0, 1, \dots$ do
    Set $\sigma = 2^i \sigma_k$ and compute $x_\sigma^+ \in \mathbb{R}^n$ such that
        $m_{k,\sigma}(x_\sigma^+) \leq m_{k,\sigma}(x_k)$   and   $\left\|\nabla m_{k,\sigma}(x_\sigma^+)\right\| \leq \frac{\sigma}{4} \left\|x_\sigma^+ - x\right\|^2.$   (2.3)
    if $f(x_k) - f(x_\sigma^+) \geq \frac{1}{12\sqrt{\sigma}} \left\|\nabla f(x_\sigma^+)\right\|^{\frac{3}{2}}$ then
        Set $i_k = i$
        $x_{k+1} = x_\sigma^+$ and $\sigma_{k+1} = \max\left(\sigma_{\min}, \frac{\sigma}{2}\right)$
        $k \leftarrow k + 1$ and go to Step 1.
    end if
end for
```

3 Convergence and complexity analysis

First, we show that all the trial points generated by Algorithm 1 belong to a bounded set, which we denote by Ω. This follows from the boundedness of the sublevelset $\mathcal{L}_f(x_0)$ and the algorithm parameter $\sigma_{\min} > 0$.

A1. *The sublevelset $\mathcal{L}_f(x_0)$ is bounded.*

Lemma 3.1. *Given $x_0 \in \mathbb{R}^n$, assume that the sublevel set $\mathcal{L}_f(x_0)$ is compact (A1) and let $\sigma_{\min} > 0$. For $\sigma \geq \sigma_{\min}$ and $x \in \mathcal{L}_f(x_0)$, consider the set of trial points*

$$S_\sigma(x) := \left\{ y \in \mathbb{R}^n : \ m_{x,\sigma}(y) \leq f(x) \ \text{ and } \ \|\nabla m_{x,\sigma}(y)\| \leq \frac{\sigma}{4}\|y - x\|^2 \right\}. \tag{3.1}$$

There exists a bounded set Ω such that for all $x \in \mathcal{L}_f(x_0)$ and $\sigma \geq \sigma_{\min}$, we have

$$S_\sigma(x) \subseteq \Omega.$$

Proof. Let $x \in \mathcal{L}_f(x_0)$ and $x_\sigma^+ \in S_\sigma(x)$, the condition $m_{x,\sigma}(x_\sigma^+) \leq f(x)$ gives

$$\tfrac{\sigma}{6}\|x_\sigma^+ - x\|^3 \leq -\langle \nabla f(x), x_\sigma^+ - x\rangle - \tfrac{1}{2}\langle \nabla^2 f(x)(x_\sigma^+ - x), x_\sigma^+ - x\rangle \tag{3.2}$$

$$\leq \|\nabla f(x)\| \, \|x_\sigma^+ - x\| + \tfrac{1}{2}\|\nabla^2 f(x)\| \, \|x_\sigma^+ - x\|^2 \tag{3.3}$$

Thus, letting $r_\sigma(x) = \|x_\sigma^+ - x\|$, we obtain the quadratic inequality

$$\tfrac{\sigma}{6} r_\sigma(x)^2 - \tfrac{1}{2}\|\nabla^2 f(x)\| r_\sigma(x) - \|\nabla f(x)\| \leq 0.$$

The quadratic has two real roots, therefore we must have

$$r_\sigma(x) \leq \frac{\frac{3}{2}\|\nabla^2 f(x)\| + 3\sqrt{\frac{1}{4}\|\nabla^2 f(x)\|^2 + \frac{2}{3}\sigma\|\nabla f(x)\|}}{\sigma} = \frac{3\|\nabla^2 f(x)\|}{2\sigma} + 3\sqrt{\frac{\|\nabla^2 f(x)\|^2}{4\sigma^2} + \frac{2\|\nabla f(x)\|}{3\sigma}}.$$

Therefore, for any $\sigma \geq \sigma_{\min} > 0$ we have

$$\|x_\sigma^+ - x\| \leq \frac{3\|\nabla^2 f(x)\|}{2\sigma_{\min}} + 3\sqrt{\frac{\|\nabla^2 f(x)\|^2}{4{\sigma_{\min}}^2} + \frac{2\|\nabla f(x)\|}{3\sigma_{\min}}} =: R(x). \tag{3.4}$$

That is,

$$x_\sigma^+ \in B\big[x; R(x)\big], \qquad \text{for all } \sigma > \sigma_{\min}.$$

Since $\mathcal{L}_f(x_0)$ is bounded, let us define

$$R_{x_0} := \sup_{x \in \mathcal{L}_f(x_0)} R(x),$$

which is finite because $R(\,\cdot\,)$ is a continuous function on the bounded set $\mathcal{L}_f(x_0)$. This establishes that for all $x \in \mathcal{L}_f(x_0)$ and $\sigma \geq \sigma_{\min}$ we have

$$S_\sigma(x) \subseteq \Omega,$$

with

$$\Omega := \overline{\text{conv}}\left\{x + d \in \mathbb{R}^n | x \in \mathcal{L}_f(x_0) \text{ and } \|d\|_2 \leq R_{x_0}\right\}, \tag{3.5}$$

where $\overline{\text{conv}}$ denotes the closure of the convex hull. The set Ω is compact since $\mathcal{L}_f(x_0)$ is compact and $R_{x_0} < \infty$. □

Our proof relies on the straightforward assumption that the third derivative of f is continuous.

A2. *The function f is three times continuously differentiable.*

For such functions, we recall the property of the Taylor polynomial with residual, along with one of its immediate consequence.

Lemma 3.2 (Taylor residual)**.** *Let $x, y \in \mathbb{R}^n$, if $f \in \mathcal{C}^3$, there exists $\xi = x + \tau(y - x)$ with $\tau \in (0,1)$ such that*

$$f(y) = T_x^{f,2}(y) + \frac{1}{3!}\mathrm{D}^3 f(\xi)[y - x]^3, \tag{3.6}$$

and

$$\left\|\nabla f(y) - \nabla T_x^{f,2}(y)\right\| \leq \frac{\left\|\mathrm{D}^3 f(\xi)\right\|}{2} \|y - x\|^2. \tag{3.7}$$

Corollary 3.3. *Under A1,A2, let Ω be the compact set given by Lemma 3.1, we define*

$$L_\Omega := \max_{y \in \Omega} \left\|\mathrm{D}^3 f(y)\right\|, \tag{3.8}$$

which is finite by continuity of $\mathrm{D}^3 f$ over the compact set Ω. Thus, for any $\sigma \geq \sigma_{\min}$ and $x \in \mathcal{L}_f(x_0)$, we have for all $x_\sigma^+ \in S_\sigma(x)$ that

$$f(x_\sigma^+) \leq T_x^{f,2}(x_\sigma^+) + \tfrac{1}{6} L_\Omega \left\|x_\sigma^+ - x\right\|^3. \tag{3.9}$$

Proof. Lemma 3.2 implies that there exists $\xi \in [x, x_\sigma^+]$ such that

$$f(x_\sigma^+) = T_x^{f,2}(x_\sigma^+) + \tfrac{1}{6} D^3 f(\xi)[x_\sigma^+ - x]^3, \tag{3.10}$$

$$\leq T_x^{f,2}(x_\sigma^+) + \tfrac{1}{6} L_\Omega \left\|x_\sigma^+ - x\right\|^3, \tag{3.11}$$

where the upper bound on the third derivative follows from $\xi \in \Omega$—since the convex set Ω contains the segment between x and x_σ^+. □

Lemma 3.4. *Let $x \in \mathcal{L}_f(x_0)$. If $\sigma \geq 2L_\Omega$, any trial step $x_\sigma^+ \in S_\sigma(x)$ is accepted and*

$$f(x) - f(x_\sigma^+) \;\geq\; \frac{1}{12\sqrt{\sigma}} \left\|\nabla f(x_\sigma^+)\right\|^{3/2}.$$

Proof. First note that for $s \in \mathbb{R}^n$,

$$\nabla \left(\frac{\sigma}{3!} \|y\|^3\right)_{|y=s} = \frac{\sigma}{2} \|s\| \, s.$$

We have

$$\nabla f(x_\sigma^+) = \nabla f(x_\sigma^+) - \nabla T_x^{f,2}(x_\sigma^+) + \nabla T_x^{f,2}(x_\sigma^+) + \nabla \left(\frac{\sigma}{3!} \|y\|^3\right)_{|y=x_\sigma^+-x} \tag{3.12}$$
$$- \nabla \left(\frac{\sigma}{3!} \|y\|^3\right)_{|y=x_\sigma^+-x}.$$

By Lemma 3.2, there exists $\xi = x + \tau(x_\sigma^+ - x)$ with $\tau \in (0,1)$ such that (3.7) holds. Combining with $L_\Omega \leq \frac{\sigma}{2}$ and (2.3) gives

$$\left\|\nabla f(x_\sigma^+)\right\| \leq \left\|\nabla f(x_\sigma^+) - \nabla T_x^{f,2}(x_\sigma^+)\right\| + \left\|\nabla T_x^{f,2}(x_\sigma^+) + \nabla \left(\frac{\sigma}{3!} \|y\|^3\right)_{|y=x_\sigma^+-x}\right\| \tag{3.13}$$
$$+ \frac{\sigma}{2} \left\|x_\sigma^+ - x\right\|^2$$
$$\leq \frac{\left\|D^3 f(\xi)\right\|}{2} \left\|x_\sigma^+ - x\right\|^2 + \left\|\nabla m_x(x_\sigma^+)\right\| + \frac{\sigma}{2} \left\|x_\sigma^+ - x\right\|^2 \tag{3.14}$$
$$\leq \frac{L_\Omega}{2} \left\|x_\sigma^+ - x\right\|^2 + \frac{\sigma}{4} \left\|x_\sigma^+ - x\right\|^2 + \frac{\sigma}{2} \left\|x_\sigma^+ - x\right\|^2 \tag{3.15}$$
$$\leq \left[\frac{\sigma}{4} + \frac{\sigma}{4} + \frac{\sigma}{2}\right] \left\|x_\sigma^+ - x\right\|^2 = \sigma \left\|x_\sigma^+ - x\right\|^2. \tag{3.16}$$

By Corollary 3.3, we have for all $x_\sigma^+ \in S_\sigma(x)$ that

$$f(x_\sigma^+) \leq T_x^{f,2}(x_\sigma^+) + \tfrac{1}{6} L_\Omega \left\|x_\sigma^+ - x\right\|^3. \tag{3.17}$$
$$= T_x^{f,2}(x_\sigma^+) + \frac{\sigma}{6} \left\|x_\sigma^+ - x\right\|^3 + \frac{(L_\Omega - \sigma)}{6} \left\|x_\sigma^+ - x\right\|^3 \tag{3.18}$$
$$= m_x(x_\sigma^+) + \frac{(L_\Omega - \sigma)}{6} \left\|x_\sigma^+ - x\right\|^3 \tag{3.19}$$
$$\leq f(x) + \frac{(L_\Omega - \sigma)}{6} \left\|x_\sigma^+ - x\right\|^3. \tag{3.20}$$

Using $\sigma - L_\Omega \geq \frac{\sigma}{2}$ and (3.16), gives

$$f(x) - f(x_\sigma^+) \geq \frac{\sigma}{12} \left\|x_\sigma^+ - x\right\|^3 \geq \frac{1}{12} \sigma^{-1/2} \left\|\nabla f(x_\sigma^+)\right\|^{3/2}. \qquad \square$$

Using the function decrease provided by the previous lemma, we establish our complexity result for the cubic Newton method (Algorithm 1). We do not assume that the Hessian of f is Lipschitz continuous, we merely require that f is three times differentiable over the bounded sublevelset $\mathcal{L}_f(x_0)$.

Theorem 3.5 (Complexity of ARC). *Let $f \in \mathcal{C}^3$ (A2) with $f(x) \geq f_{\text{low}}$ for all $x \in \mathbb{R}^n$ and assume that the sublevelset $\mathcal{L}_f(x_0)$ is bounded for some $x_0 \in \mathbb{R}^n$ (A1). The adaptive cubic Newton method (Algorithm 1) requires at most*

$$2 + 24(f(x_0) - f_{\text{low}}) \max\{\sigma_{\min}, 4L_\Omega\}^{1/2} \varepsilon^{-3/2} + \log_2\left(\frac{\max\{\sigma_{\min}, 2L_\Omega\}}{\sigma_0}\right) \tag{3.21}$$

evaluations of f, ∇f and $\nabla^2 f$ to find a point that satisfies $\|\nabla f(x)\| \leq \varepsilon$, where L_Ω is such that $\left\|D^3 f(y)\right\| \leq L_\Omega$ for all y in the compact set Ω containing $\mathcal{L}_f(x_0)$—defined in Lemma 3.1 .

Proof. Let

$$T = \inf\{k \in \mathbb{N} : \|\nabla f(x_k)\| \leq \varepsilon\}.$$

Thus, if T is finite, we have $\|\nabla f(x_k)\| > \varepsilon$ for all $k \leq T-1$, and $\|\nabla f(x_T)\| \leq \varepsilon$, which means that the algorithm ends in Step 1 of iteration T (the test for termination).

At iteration k, the next iterate x_{k+1} is computed with regularization parameter $\sigma = 2^{i_k}\sigma_k$. Lemma 3.4 ensures that $\sigma \geq 2L_\Omega$ is sufficient to accept the trial point and produce a function decrease, where L_Ω is defined such that $\left\|\mathrm{D}^3 f(x)\right\| \leq L_\Omega$ for all $x \in \Omega$ (see Corollary 3.3).

Thus, under A1 and $f \in \mathcal{C}^3$, Algorithm (1) to minimize f starting from $x_0 \in \mathbb{R}^n$ is well defined, all iterates belong to the set $\mathcal{L}_f(x_0)$ and all trial points belong to Ω.

From $\max(\sigma_{\min}, \frac{\sigma}{2}) = \sigma_{k+1}$ we get

$$2^{i_k-1}\sigma_k = \frac{\sigma}{2} \leq \sigma_{k+1}.$$

From $2^{i_k-1} = \sigma_{k+1}/\sigma_k$, we obtain $i_k + 1 = 2 + \log_2 \sigma_{k+1} - \log \sigma_k$. The total number of subproblems that is solved up to iteration T is given by

$$\sum_{k=0}^{T-1}(i_k+1) = \sum_{k=0}^{T-1} 2 + \log_2 \sigma_{k+1} - \log \sigma_k \tag{3.22}$$

$$= 2T + \log_2 \sigma_T - \log_2 \sigma_0 \tag{3.23}$$

$$\leq 2T + \log_2\left(\frac{\max\{\sigma_{\min}, 2L_\Omega\}}{\sigma_0}\right), \tag{3.24}$$

where we used that $\sigma_T \leq \max\{\sigma_{\min}, 2L_\Omega\}$. This follows from the mechanism of the update $\max(\sigma_{\min}, \frac{\sigma}{2}) = \sigma_{k+1}$ and Lemma 3.4 ensuring that any accepted trial step is computed with a regularization parameter $\sigma = 2^{i_k}\sigma_k \leq \max\{\sigma_{\min}, 4L_\Omega\}$.

We sum the function decrease at each iteration to conclude,

$$f(x_0) - f_{\text{low}} \geq \sum_{k=0}^{T-2} f(x_k) - f(x_{k+1}) \tag{3.25}$$

$$\geq \sum_{k=0}^{T-2} \frac{1}{12}(2^{i_k}\sigma_k)^{-1/2}\|\nabla f(x_{k+1})\|^{\frac{3}{2}} \tag{3.26}$$

$$\geq (T-1)\frac{1}{12}\max\{\sigma_{\min}, 4L_\Omega\}^{-1/2}\varepsilon^{\frac{3}{2}}. \tag{3.27}$$

This yields an upper bound on T

$$T - 1 \leq 12(f(x_0) - f_{\text{low}})\max\{\sigma_{\min}, 4L_\Omega\}^{1/2}\varepsilon^{-\frac{3}{2}}. \tag{3.28}$$

□

References

Cartis, C., Gould, N. I., and Toint, P. L. (2010a). On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization problems. *Siam journal on optimization*, 20(6):2833–2852.

Cartis, C., Gould, N. I. M., and Toint, P. L. (2010b). Adaptive cubic regularisation methods for unconstrained optimization. Part II: Worst-case function- and derivative-evaluation complexity. *Mathematical Programming*, 130(2):295–319.

Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization.* Springer US.

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