"Accelerating the cubic regularization of Newton's method on convex problems"

Nesterov, Yurii

ABSTRACT

In this paper we propose an accelerated version of the cubic regularization of Newton's method [6]. The original version, used for minimizing a convex function with Lipschitz-continuous Hessian, guarantees a global rate of convergence of order O(1/k exp.2), where k is the iteration counter. Our modified version converges for the same problem class with order O(1/k exp.3), keeping the complexity of each iteration unchanged. We study the complexity of both schemes on different classes of convex problems. In particular, we argue that for the second-order schemes, the class of non-degenerate problems is different from the standard class.

Accelerating the cubic regularization of Newton's method on convex problems

Yu. Nesterov *

September 2005

Abstract

In this paper we propose an accelerated version of the cubic regularization of Newton's method [6]. The original version, used for minimizing a convex function with Lipschitz-continuous Hessian, guarantees a global rate of convergence of order $O(\frac{1}{k^2})$, where k is the iteration counter. Our modified version converges for the same problem class with order $O(\frac{1}{k^3})$, keeping the complexity of each iteration unchanged. We study the complexity of both schemes on different classes of convex problems. In particular, we argue that for the second-order schemes, the class of non-degenerate problems is different from the standard class.

Keywords: convex optimization, unconstrained minimization, Newton's method, cubic regularization, worst-case complexity, global complexity bounds, non-degenerate problems, condition number.

*Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL), 34 voie du Roman Pays, 1348 Louvain-la-Neuve, Belgium; e-mail: nesterov@core.ucl.ac.be.

The research results presented in this paper have been supported by a grant "Action de recherche concertè ARC 04/09-315" from the "Direction de la recherche scientifique - Communautè française de Belgique". The scientific responsibility rests with the author.

1 Introduction

Motivation. Newton's method is one of the oldest schemes in Numerical Analysis [1]. The first theoretical study of its local performance was carried out in [4]. During many years, numerous useful suggestions were developed for stabilizing its local behavior (a detailed description of the results and references can be found in the comprehensive monographs [2, 3]). However, the global worst-case complexity analysis for the second-order schemes was only given recently.

In [6] a cubic regularization of the Newton's method (CNM), which can be applied to a general smooth unconstrained minimization problem, was proposed. The main advantage of this scheme consists in its natural geometrical interpretation: At each iteration we minimize a *cubic model*, which appears to be a *global upper estimate* for the objective function. This key feature ensures predictable global behavior of the process.

In [6], attention was mainly paid to different classes of nonconvex problems. For convex problems, some of the statements of [6] can be strengthened. Moreover, we can employ convex optimization techniques for accelerating the local scheme.

In this paper we assume that the objective function of a convex unconstrained minimization problem has Lipschitz-continuous Hessian (that is the basic problem class under consideration). As was shown in [6], the global rate of convergence of CNM on this problem class is of the order $O(\frac{1}{k^2})$, where k is the iteration counter. However, note that CNM is a *local one-step second-order* method. From the complexity theory of smooth convex optimization (see, for example, Chapter 2 in [5]), it is known that the rate of convergence of the local one-step first-order method (that is just a *gradient* method) can be improved from $O(\frac{1}{k})$ to $O(\frac{1}{k^2})$ by passing to a *multi-step* strategy. In this paper we show that a similar trick works with CNM also. As a result, we get a new method, which converges on the specified class of problems as $O(\frac{1}{k^3})$.

Contents. Section 2 contains all necessary results related to *regular* functions. Namely, we present the main properties of uniformly convex functions and functions with Lipschitz-continuous derivatives of a certain degree. In Section 3 we introduce a cubic regularization of the Newton step and prove several useful inequalities. At the end of this section we show that CNM, as applied to a convex function from the basic problem class, converges globally as $O(\frac{1}{k^2})$. The majority of the results of this section can be found in [6], but in a slightly weaker form.

In Section 4 we derive an accelerated multi-step version of CNM. We prove that on the basic problem class it converges as $O(\frac{1}{k^3})$. This acceleration is achieved by a modification of the technique of *estimate sequences* described in Section 2.2.1 [5].

In Section 5 we introduce a new class of problems, which can be seen as non-degenerate problems for the second-order methods. This class is composed of the functions from the basic problem class, which are uniformly convex of *degree three*. On these problems both CNM and its accelerated version exhibit a global linear rate of convergence, which is proportional to the product of a certain power of the *second-order condition number* and the logarithm of the required accuracy. We show that the accelerated scheme with an appropriately chosen restarting strategy works better than the pure CNM. The results of this section suggest that the notion of non-degeneracy is *method-dependent*. The standard classification of non-degenerate problems based on the usual condition number works

properly only for the first-order schemes.

In Section 6 we analyze the performance of the proposed schemes on strongly convex functions from the basic problem class. We show that the main computational effort is spent at the first stage of the process, aiming to enter the region of quadratic convergence. In some sense, the complexity of such problems is almost independent on the required accuracy. This situation can lead to erroneous conclusions on the efficiency of some numerical schemes. In Section 7 we give an example of a method which, when applied to strongly convex functions, formally converges as $O(\frac{1}{k^8})$. However, its actual performance appears to be worse than that of accelerated CNM.

Finally, in the last Section 8 we discuss the results.

Notation. In what follows E denotes a finite-dimensional real vector space, and E^* the dual space, which is formed by all linear functions on E. The value of function $s \in E^*$ at $x \in E$ is denoted by $\langle s, x \rangle$.

Let us fix a positive definite self-adjoint operator $B : E \to E^*$. Define the following norms:

$$\begin{aligned} \|h\| &= \langle Bh, h \rangle^{1/2}, \quad h \in E, \\ \|s\|_* &= \langle s, B^{-1}s \rangle^{1/2}, \quad s \in E^*, \\ \|A\| &= \max_{\|h\| \leq 1} \|Ah\|_*, \quad A : E \to E^*. \end{aligned}$$

For a self-adjoint operator $A = A^*$, the same norm can be defined as

$$\|A\| = \max_{\|h\| \leq 1} |\langle Ah, h \rangle|. \tag{1.1}$$

Any $s \in E^*$ generates a rank-one self-adjoint operator $ss^* : E \to E^*$ acting as follows

$$ss^* \cdot x = \langle s, x \rangle \cdot s, \quad x \in E.$$

We extend operator $A(s) = \frac{ss^*}{\|s\|_*}$ onto the origin in a continuous way: $A(0) = 0$.

Further, for function $f(x)$, $x \in E$, we denote by $\nabla f(x)$ its *gradient* at x:

$$f(x+h) = f(x) + \langle \nabla f(x), h \rangle + o(\|h\|), \quad h \in E.$$

Clearly $\nabla f(x) \in E^*$. Similarly, we denote by $\nabla^2 f(x)$ the *Hessian* of f at x:

$$\nabla f(x+h) = \nabla f(x) + \nabla^2 f(x)h + \mathbf{o}(\|h\|), \quad h \in E.$$

Of course, $\nabla^2 f(x)$ is a self-adjoint linear operator from E to E^*.

Acknowledgements. The author would like to thank Laurence Wolsey for very useful comments on the content of the paper.

2 Regular functions

In this section, for the sake of completeness, we include all necessary results on convex functions possessing a certain type of regularity. We start from well known properties of *uniformly convex functions* (see, for example, [8]).

Let a function $f(x)$ be differentiable on E. We call it *uniformly convex on E of degree $p \geq 2$* if there exists a constant $\sigma_p = \sigma_p(f) > 0$ such that

$$f(y) \quad \geq \quad f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{1}{p}\sigma_p \|y - x\|^p, \quad \forall x, y \in E. \tag{2.1}$$

The pair (p, σ_p) is called the *pair of parameters* of the uniformly convex function. Adding such a function to an arbitrary convex function gives a uniformly convex function with the same pair of parameters. Recall that the degree $p = 2$ corresponds to *strongly convex* functions.

Lemma 1 *Assume that for some $p \geq 2$, $\sigma > 0$, and all $x, y \in E$ the following inequality holds:*

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \quad \geq \quad \sigma \|x - y\|^p, \quad x, y \in E. \tag{2.2}$$

Then the function f is uniformly convex on E with parameters p and σ.

Proof:
Indeed,

$$\begin{aligned} f(y) - f(x) - \langle \nabla f(x), y - x \rangle \quad &= \quad \int\limits_0^1 \langle f(x + \tau(y - x)) - \nabla f(x), y - x \rangle d\tau \\ &= \quad \int\limits_0^1 \tfrac{1}{\tau} \langle f(x + \tau(y - x)) - \nabla f(x), \tau(y - x) \rangle d\tau \\ &\overset{(2.2)}{\geq} \quad \int\limits_0^1 \sigma \tau^{p-1} \|y - x\|^p d\tau \;=\; \tfrac{1}{p}\sigma \|y - x\|^p. \end{aligned}$$

□

In our analysis we often use the following inequality.

Lemma 2 *For any $h \in E$, and $s \in E^*$ we have*

$$\langle s, h \rangle + \tfrac{1}{p}\sigma \|h\|^p \quad \geq \quad -\tfrac{p-1}{p} \left(\tfrac{1}{\sigma}\right)^{\frac{1}{p-1}} \|s\|_*^{\frac{p}{p-1}}. \tag{2.3}$$

Proof:
Denote by h the minimum of the left-hand side in (2.3). It satisfies the first-order optimality condition

$$s + \sigma \|h\|^{p-2} Bh = 0.$$

Hence, $\langle s, h \rangle = -\sigma \|h\|^p$ and $\|s\|_* = \sigma \|h\|^{p-1}$. Therefore

$$\langle s, h \rangle + \tfrac{1}{p}\sigma \|h\|^p \;=\; -\tfrac{p-1}{p}\sigma \|h\|^p \;=\; -\tfrac{p-1}{p}\sigma \left(\tfrac{1}{\sigma}\|s\|_*\right)^{\frac{p}{p-1}},$$

and (2.3) follows. □

Lemma 3 *Let f be uniformly convex on E of degree $p \geq 2$. Then*

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle \;\leq\; \tfrac{p-1}{p} \left(\tfrac{1}{\sigma_p}\right)^{\frac{1}{p-1}} \|\nabla f(y) - \nabla f(x)\|_*^{\frac{p}{p-1}}. \tag{2.4}$$

Proof:
Assume that f attains its global minimum at some point x^*. Then

$$\begin{aligned} f(x^*) \;&=\; \min_y f(y) \overset{(2.1)}{\geq} \min_{x \in E} \left[f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{1}{p}\sigma_p \|y - x\|^p \right] \\ &\overset{(2.3)}{=} f(x) - \tfrac{p-1}{p} \left(\tfrac{1}{\sigma}\right)^{\frac{1}{p-1}} \|\nabla f(x)\|_*^{\frac{p}{p-1}}. \end{aligned}$$

Let us fix x and consider the convex function $\phi(y) = f(y) - \langle \nabla f(x), y \rangle$. Note that it is uniformly convex with parameters p and σ_p. Moreover, it attains its minimum at $y = x$. Hence, applying the above inequality to $\phi(y)$, we get (2.4). $\square$

Note that for $p = 2$ conditions (2.2) and (2.4) are necessary and sufficient for a function f to be strongly convex with parameter $\sigma_2 = \sigma$. Twice differentiable strongly convex functions admit also the following convenient characterization:

$$\langle \nabla^2 f(x) h, h \rangle \;\geq\; \sigma_2 \|h\|^2, \quad \forall x, h \in E. \tag{2.5}$$

Finally, let us give an important example of a uniformly convex function. Let us fix an arbitrary $x_0 \in E$. Define the function $d_p(x) = \frac{1}{p}\|x - x_0\|^p$. Then

$$\nabla d_p(x) \;=\; \|x - x_0\|^{p-2} \cdot B(x - x_0), \quad x \in E.$$

Lemma 4 *For any x and y from E we have*

$$\langle \nabla d_p(x) - \nabla d_p(y), x - y \rangle \;\geq\; \left(\tfrac{1}{2}\right)^{p-2} \|x - y\|^p, \tag{2.6}$$

$$d_p(x) - d_p(y) - \langle \nabla d_p(y), x - y \rangle \;\geq\; \tfrac{1}{p} \left(\tfrac{1}{2}\right)^{p-2} \|x - y\|^p. \tag{2.7}$$

Proof:
Without loss of generality assume $x_0 = 0$. Then

$$\begin{aligned} \langle \nabla d_p(x) - \nabla d_p(y), x - y \rangle \;&=\; \langle \|x\|^{p-2} \cdot Bx - \|y\|^{p-2} \cdot By, x - y \rangle \\ &=\; \|x\|^p + \|y\|^p - \langle Bx, y \rangle (\|x\|^{p-2} + \|y\|^{p-2}). \end{aligned}$$

For proving (2.6), we need to show that the right-hand side of the latter equality is greater or equal than

$$\left(\tfrac{1}{2}\right)^{p-2} \|x - y\|^p = \left(\tfrac{1}{2}\right)^{p-2} \left[\|x\|^2 + \|y\|^2 - 2\langle Bx, y \rangle \right]^{p/2}.$$

Without loss of generality we can assume that $x \neq 0$ and $y \neq 0$. Then, denoting by

$$\tau \;=\; \tfrac{\|y\|}{\|x\|}, \quad \alpha = \tfrac{\langle Bx, y \rangle}{\|x\| \cdot \|y\|} \in [-1, 1],$$

we get the statement required to be proved:

$$1+\tau^p \;\geq\; \alpha\tau(1+\tau^{p-2}) + \left(\tfrac{1}{2}\right)^{p-2}[1+\tau^2-2\alpha\tau]^{p/2}, \quad \tau\geq 0, \quad |\alpha|\leq 1. \tag{2.8}$$

Since the right-hand side of this inequality is convex in α, we need to justify two marginal inequalities:

$$\begin{aligned} \alpha = 1: \qquad 1+\tau^p &\;\geq\; \tau(1+\tau^{p-2}) + \left(\tfrac{1}{2}\right)^{p-2}|1-\tau|^p, \\ \alpha = -1: \qquad 1+\tau^p &\;\geq\; -\tau(1+\tau^{p-2}) + \left(\tfrac{1}{2}\right)^{p-2}(1+\tau)^p \end{aligned} \tag{2.9}$$

for all $\tau\geq 0$.

The second inequality in (2.9) can be derived from the lower bound for the ratio

$$\frac{1+\tau^p+\tau(1+\tau^{p-2})}{(1+\tau)^p} \;=\; \frac{1+\tau^{p-1}}{(1+\tau)^{p-1}}, \quad \tau\geq 0.$$

Indeed, its minimum is attained at $\tau=1$, and that proves the second line in (2.9). For proving the first line, note that it is valid for $\tau=1$. If $\tau\geq 0$ and $\tau\neq 1$, then we need to estimate from below the ratio

$$\frac{1+\tau^p-\tau(1+\tau^{p-2})}{|1-\tau|^p} \;=\; \frac{(1-\tau)(1-\tau^{p-1})}{|1-\tau|^p} \;=\; \frac{1+\tau+\ldots+\tau^{p-2}}{|1-\tau|^{p-2}}$$

Since the absolute value of any coefficient of the polynomial $(1-\tau)^{p-2}$ does not exceed 2^{p-2}, the first line in inequality (2.9) is also justified. This proves (2.6), and, for proving (2.7), we can use now Lemma 1. □

Another type of regularity that we are interested concerns the smoothness conditions (see, for example, [7]). Usually they are stated in terms of Lipschitz conditions for derivatives of a certain order:

$$\|\nabla^k f(x) - \nabla^k f(y)\| \;\leq\; L_{k+1}(f)\|x-y\|, \quad x,y\in E, \quad k\geq 0.$$

In this paper we mainly consider functions with Lipschitz-continuous Hessian:

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \;\leq\; L_3\|x-y\|, \quad x,y\in E, \tag{2.10}$$

where $L_3 \stackrel{\rm def}{=} L_3(f)$. Consequently, for all x and y from E we have

$$\|\nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y-x)\|_* \;\leq\; \tfrac{1}{2}L_3\|y-x\|^2. \tag{2.11}$$

Moreover, for the quadratic model

$$f_2(x;y) \stackrel{\rm def}{=} f(x) + \langle \nabla f(x), y-x\rangle + \tfrac{1}{2}\langle \nabla^2 f(x)(y-x), y-x\rangle$$

we can bound the residual:

$$|f(y) - f_2(x;y)| \;\leq\; \tfrac{L_3}{6}\|y-x\|^3, \quad x,y\in E. \tag{2.12}$$

The simplest functions satisfying condition (2.10) are, of course, the quadratic functions. However, sometimes another function $d_3(x)$ is quite useful.

Lemma 5 *For any $x, y \in E$ we have*

$$\|\nabla^2 d_3(x) - \nabla^2 d_3(y)\| \;\leq\; 2\,\|x-y\|. \tag{2.13}$$

Proof:
Without loss of generality, assume $x_0 = 0$. Then $d_3(x) = \frac{1}{3}\|x\|^3$, and for any $x \in E$ we have

$$\nabla^2 d_3(x) \;=\; \|x\| B + \tfrac{1}{\|x\|} B x x^* B.$$

Let us fix two points $x, y \in E$ and arbitrary direction $h \in E$. Define $x(\tau) = x + \tau(y-x)$ and

$$\phi(\tau) \;=\; \langle \nabla^2 d_3(x(\tau))h, h\rangle \;=\; \|x(\tau)\| \cdot \|h\|^2 + \tfrac{1}{\|x(\tau)\|}\langle Bx(\tau), h\rangle^2, \quad \tau \in [0,1].$$

Assume first that $0 \notin [x, y]$. Then $\phi(\tau)$ is continuously differentiable on $[0,1]$ and

$$\begin{aligned}
\phi'(\tau) \;&=\; \tfrac{\langle Bx(\tau), y-x\rangle}{\|x(\tau)\|} \cdot \|h\|^2 + \tfrac{2\langle Bx(\tau),h\rangle}{\|x(\tau)\|}\langle Bh, y-x\rangle - \tfrac{\langle Bx(\tau),h\rangle^2}{\|x(\tau)\|^3}\langle Bx(\tau), y-x\rangle \\
&=\; \tfrac{\langle Bx(\tau), y-x\rangle}{\|x(\tau)\|} \cdot \underbrace{\left(\|h\|^2 - \frac{\langle Bx(\tau), h\rangle^2}{\|x(\tau)\|^2} \right)}_{\geq 0} + \tfrac{2\langle Bx(\tau),h\rangle}{\|x(\tau)\|}\langle Bh, y-x\rangle.
\end{aligned}$$

Denote $\alpha = \frac{\langle Bx(\tau),h\rangle}{\|x(\tau)\|\cdot\|h\|} \in [-1,1]$. Then

$$|\phi'(\tau)| \;\leq\; \|y-x\| \cdot \|h\|^2 \cdot (1 - \alpha^2 + 2|\alpha|) \;\leq\; 2\,\|y-x\| \cdot \|h\|^2.$$

Hence,

$$|\langle (\nabla^2 d_3(y) - \nabla^2 d_3(x))h, h\rangle| \;=\; |\phi(1) - \phi(0)| \;\leq\; 2\,\|y-x\| \cdot \|h\|^2,$$

and we get (2.13) from (1.1).

The remaining case $0 \in [x, y]$ is trivial since $\nabla^2 d_3(0) = 0$. □

In the sequel we often establish the complexity of different problem classes in terms of *condition numbers* of certain degree:

$$\gamma_p(f) \;\stackrel{\text{def}}{=}\; \tfrac{\sigma_p(f)}{L_p(f)}, \quad p \geq 2. \tag{2.14}$$

It is clear, for example, that $\gamma_2(d_2) = 1$. On the other hand, from (2.7) and (2.13) we conclude that $\gamma_3(d_3) = \frac{1}{4}$.

3 Cubic regularization of Newton iteration

In this section we present the most important properties of cubic regularization of the Newton's method. As compared with [6], here we take into account the convexity of the objective function.

The main problem of interest is as follows:

$$\min_{x \in E} \; f(x), \tag{3.1}$$

where f is a twice differentiable convex function with Lipschitz-continuous Hessian. As suggested in [6], we introduce the following mapping:

$$T_M(x) \stackrel{\text{def}}{=} \operatorname{Arg}\min_{y \in E} \left[\hat{f}_M(x;y) \stackrel{\text{def}}{=} f_2(x;y) + \tfrac{M}{6}\|y-x\|^3 \right]. \tag{3.2}$$

Note that $T = T_M(x)$ is the unique solution of the following system:

$$\nabla f(x) + \nabla^2 f(x)(T-x) + \tfrac{1}{2} M \cdot \|T-x\| \cdot B(T-x) = 0. \tag{3.3}$$

Denote $r_M(x) = \|x - T_M(x)\|$. Then,

$$\begin{aligned} \|\nabla f(T)\|_* &\stackrel{(3.3)}{=} \|\nabla f(T) - \nabla f(x) - \nabla^2 f(x)(T-x) - \tfrac{M}{2} r_M(x) B(T-x)\|_* \\ &\stackrel{(2.11)}{\leq} \tfrac{L_3+M}{2} r_M^2(x). \end{aligned} \tag{3.4}$$

Further, multiplying (3.3) by $T - x$, we obtain

$$\langle \nabla f(x), x-T \rangle \;=\; \langle \nabla^2 f(x)(T-x), T-x \rangle + \tfrac{1}{2} M r_M^3(x). \tag{3.5}$$

Let us assume $M \geq L_3$. Then, in view of (2.12), we have

$$\begin{aligned} f(x) - f(T) &\geq f(x) - \hat{f}_M(x;T) \\ &= \langle \nabla f(x), x-T \rangle - \tfrac{1}{2}\langle \nabla^2 f(x)(T-x), T-x \rangle - \tfrac{M}{6} r_M^3(x) \\ &= \tfrac{1}{2}\langle \nabla^2 f(x)(T-x), T-x \rangle + \tfrac{M}{3} r_M^3(x). \end{aligned} \tag{3.6}$$

In particular, since f is convex,

$$f(x) - f(T) \stackrel{(3.6)}{\geq} \tfrac{M}{3} r_M^3(x) \stackrel{(3.4)}{\geq} \tfrac{M}{3} \left(\tfrac{2}{L_3+M} \|\nabla f(T)\|_* \right)^{3/2}. \tag{3.7}$$

Sometimes we need to interpret this step from a global perspective:

$$f(T) \stackrel{(M \geq L_3)}{\leq} \min_y \left[f_2(x;y) + \tfrac{M}{6}\|y-x\|^3 \right] \stackrel{(2.12)}{\leq} \min_y \left[f(y) + \tfrac{L_3+M}{6}\|y-x\|^3 \right]. \tag{3.8}$$

Finally, let us prove the following result.

Lemma 6 *If $M \geq 2L_3$, then*

$$\langle \nabla f(T), x-T \rangle \;\geq\; \sqrt{\tfrac{2}{L_3+M}} \cdot \|\nabla f(T)\|_*^{3/2}. \tag{3.9}$$

Proof:
Denote $T = T_M(x)$ and $r = r_M(x)$. Then

$$\begin{aligned} \tfrac{1}{4} L_3^2 r^4 \;=\; \left(\tfrac{L_3}{2}\|T-x\|^2 \right)^2 &\stackrel{(2.11)}{\geq} \|\nabla f(T) - \nabla f(x) - \nabla^2 f(x)(T-x)\|_*^2 \\ &\stackrel{(3.3)}{=} \|\nabla f(T) + \tfrac{1}{2} M \cdot r \cdot B(T-x)\|_*^2 \\ &= \|\nabla f(T)\|_*^2 + M r \langle \nabla f(T), T-x \rangle + \tfrac{1}{4} M^2 r^4. \end{aligned}$$

Hence,

$$\langle \nabla f(T), x - T \rangle \;\; \geq \;\; \tfrac{1}{Mr} \|\nabla f(T)\|_*^2 + \tfrac{1}{4M}(M^2 - L_3^2) r^3. \tag{3.10}$$

In view of the conditions of the lemma, we can estimate the derivative in r of the right-hand side of the latter inequality:

$$-\tfrac{1}{Mr^2}\|\nabla f(T)\|_*^2 + \tfrac{3r^2}{4M}(M^2 - L_3^2) \;\geq\; -\tfrac{1}{Mr^2}\|\nabla f(T)\|_*^2 + \left(\tfrac{L_3+M}{2}\right)^2 \tfrac{r^2}{M} \;\overset{(3.4)}{\geq}\; 0.$$

Thus, its minimum is attained at the boundary point $r = \left[\frac{2}{L_3+M}\|\nabla f(T)\|_*\right]^{1/2}$ of the feasible ray (3.4). Substituting this value in (3.10), we obtain (3.9). $\square$

To conclude this section, let us estimate the rate of convergence of the CNM method as applied to our basic problem (3.1). We assume that a solution of this problem x^* exists, and the Lipschitz constant L_3 for the Hessian of the objective function is known. Thus, we just iterate

$$x_{k+1} \;=\; T_{L_3}(x_k), \quad k = 0, 1, \ldots. \tag{3.11}$$

Using the same arguments as in [6], we can prove the following statement.

Theorem 1 *Assume that the level sets of the problem (3.1) are bounded:*

$$\|x - x^*\| \leq D \quad \forall x : \; f(x) \leq f(x_0). \tag{3.12}$$

If the sequence $\{x_k\}_{k=1}^{\infty}$ is generated by (3.11), then

$$f(x_k) - f(x^*) \;\leq\; \tfrac{9L_3D^3}{(k+4)^2}, \quad k \geq 1. \tag{3.13}$$

Proof:
In view of (3.6), $f(x_{k+1}) \leq f(x_k)$, $k \geq 0$. Thus, $\|x_k - x^*\| \leq D$ for all $k \geq 0$. Further, in view of (3.8), we have

$$f(x_1) \;\leq\; f(x^*) + \tfrac{L_3}{3} D^3. \tag{3.14}$$

Consider now an arbitrary $k \geq 1$. Denote $x_k(\tau) = x^* + (1-\tau)(x_k - x^*)$. In view of (3.8), for any $\tau \in [0,1]$ we have

$$f(x_{k+1}) \;\;\leq\;\; f(x_k(\tau)) + \tau^3 \tfrac{L_3}{3}\|x_k - x^*\|^3 \;\leq\; f(x_k) - \tau(f(x_k) - f(x^*)) + \tau^3 \tfrac{L_3 D^3}{3}$$

The minimum of the right-hand side is attained for

$$\tau \;\;=\;\; \sqrt{\tfrac{f(x_k)-f(x^*)}{L_3D^3}} \;\leq\; \sqrt{\tfrac{f(x_1)-f(x^*)}{L_3D^3}} \;\overset{(3.14)}{<}\; 1.$$

Thus, for any $k \geq 1$ we have

$$f(x_{k+1}) \;\;\leq\;\; f(x_k(\tau)) - \tfrac{2}{3} \cdot \tfrac{(f(x_k)-f(x^*))^{3/2}}{\sqrt{L_3D^3}}. \tag{3.15}$$

Denote $\delta_k = f(x_k) - f(x^*)$. Then

$$\tfrac{1}{\sqrt{\delta_{k+1}}} - \tfrac{1}{\sqrt{\delta_k}} \;\;=\;\; \tfrac{\delta_k - \delta_{k+1}}{\sqrt{\delta_k\delta_{k+1}}(\sqrt{\delta_k}+\sqrt{\delta_{k+1}})} \;\overset{(3.15)}{\geq}\; \tfrac{2}{3\sqrt{L_3D^3}} \cdot \tfrac{\delta_k}{\sqrt{\delta_{k+1}}(\sqrt{\delta_k}+\sqrt{\delta_{k+1}})} \;\geq\; \tfrac{1}{3\sqrt{L_3D^3}}.$$

Thus, for any $k \geq 1$, we have

$$\frac{1}{\sqrt{\delta_k}} \;\geq\; \frac{1}{\sqrt{\delta_1}} + \frac{k-1}{3\sqrt{L_3 D^3}} \overset{(3.14)}{\geq} \frac{1}{\sqrt{L_3 D^3}} \cdot \left(\sqrt{3} + \frac{k-1}{3}\right) \;\geq\; \frac{k+4}{3\sqrt{L_3 D^3}}.$$

□

4 Accelerated scheme

In order to accelerate method (3.11), we apply a variant of the technique of *estimate sequences*, which was described in Section 2.2.1 [5] as a tool for accelerating the usual gradient method. In our situation, this idea can be applied to CNM in different ways. We mention only two of them.

1. Linear estimate functions. For solving the optimization problem (3.1), we recursively update the following sequences.

- Sequence of estimate functions

$$\psi_k(x) \;=\; l_k(x) + \tfrac{N}{6}\|x - x_0\|^3, \quad k = 1, 2, \ldots,$$

  where $l_k(x)$ are linear functions in $x \in E$, and N is a positive real parameter.
- A minimizing sequence $\{x_k\}_{k=1}^{\infty}$.
- A sequence of scaling parameters $\{A_k\}_{k=1}^{\infty}$:

$$A_{k+1} \overset{\text{def}}{=} A_k + a_k, \quad k = 1, 2, \ldots.$$

For these objects, we are going to maintain the following relations:

$$\left.\begin{array}{lrcl} \mathcal{R}_k^1: & A_k f(x_k) & \leq & \psi_k^* \;\equiv\; \min\limits_x \psi_k(x), \\ \mathcal{R}_k^2: & \psi_k(x) & \leq & A_k f(x) + \frac{2L_3+N}{6}\|x - x_0\|^3, \; \forall x \in E. \end{array}\right\}, \quad k \geq 1. \tag{4.1}$$

Let us ensure that relations (4.1) hold for $k = 1$. We choose

$$x_1 \;=\; T_{L_3}(x_0), \quad l_1(x) \;\equiv\; f(x_1), \; x \in E, \quad A_1 \;=\; 1. \tag{4.2}$$

Then $\psi_1^* = f(x_1)$, so $\mathcal{R}_1^1$ holds. On the other hand, in view of , we get

$$\begin{array}{rcl} \psi_1(x) & = & f(x_1) + \frac{N}{6}\|x - x_0\|^3 \\ & \overset{(3.8)}{\leq} & \min\limits_y \left[f(y) + \frac{2L_3}{6}\|y - x_0\|^3 \right] + \frac{N}{6}\|x - x_0\|^3, \end{array}$$

and $\mathcal{R}_1^2$ follows.

Assume now that relations (4.1) hold for some $k \geq 1$. Denote

$$v_k \;=\; \arg\min_x \; \psi_k(x).$$

Let us choose some $a_k > 0$ and $M \geq 2L_3$. Define

$$\begin{array}{rcl} \alpha_k & = & \frac{a_k}{A_k + a_k}, \\ y_k & = & (1-\alpha_k)x_k + \alpha_k v_k, \\ x_{k+1} & = & T_M(y_k), \\ \psi_{k+1}(x) & = & \psi_k(x) + a_k[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle]. \end{array} \tag{4.3}$$

In view of $\mathcal{R}_k^2$, for any $x \in E$ we have

$$\begin{array}{rcl} \psi_{k+1}(x) & \leq & A_k f(x) + \frac{2L_3+N}{6}\|x - x_0\|^3 + a_k[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle] \\ & \leq & (A_k + a_k) f(x) + \frac{2L_3+N}{6}\|x - x_0\|^3, \end{array}$$

and this is $\mathcal{R}_{k+1}^2$. Let us show now that, for an appropriate choice of a_k, N and M, relation $\mathcal{R}_{k+1}^1$ is also valid.

Indeed, in view of $\mathcal{R}_k^1$ and by Lemma 4 with $p = 3$, for any $x \in E$, we have

$$\begin{array}{rcl} \psi_k(x) & \equiv & l_k(x) + \frac{N}{2} d_3(x) \;\geq\; \psi_k^* + \frac{N}{2} \cdot \frac{1}{6}\|x - v_k\|^3 \\ & \geq & A_k f(x_k) + \frac{N}{2} \cdot \frac{1}{6}\|x - v_k\|^3. \end{array} \tag{4.4}$$

Therefore

$$\begin{array}{rcl} \psi_{k+1}^* & = & \min\limits_x \;\; \{\psi_k(x) + a_k[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle]\} \\ & \overset{(4.4)}{\geq} & \min\limits_x \;\; \Big\{A_k f(x_k) + \frac{N}{12}\|x - v_k\|^3 + a_k[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle]\Big\} \\ & \geq & \min\limits_x \;\; \{(A_k + a_k) f(x_{k+1}) + A_k\langle \nabla f(x_{k+1}), x_k - x_{k+1}\rangle \\ & & \quad + a_k\langle \nabla f(x_{k+1}), x - x_{k+1}\rangle + \frac{N}{12}\|x - v_k\|^3]\} \\ & \overset{(4.3)}{=} & \min\limits_x \;\; \{A_{k+1} f(x_{k+1}) + \langle \nabla f(x_{k+1}), A_{k+1} y_k - a_k v_k - A_k x_{k+1}\rangle \\ & & \quad + a_k\langle \nabla f(x_{k+1}), x - x_{k+1}\rangle + \frac{N}{12}\|x - v_k\|^3]\} \\ & = & \min\limits_x \;\; \{A_{k+1} f(x_{k+1}) + A_{k+1}\langle \nabla f(x_{k+1}), y_k - x_{k+1}\rangle \\ & & \quad + a_k\langle \nabla f(x_{k+1}), x - v_k\rangle + \frac{N}{12}\|x - v_k\|^3]\}. \end{array}$$

Further, if we choose $M \geq 2L_3$, then by (3.9) we have

$$\langle \nabla f(x_{k+1}), y_k - x_{k+1}\rangle \;\geq\; \sqrt{\frac{2}{L_3 + M}} \cdot \|\nabla f(x_{k+1})\|_*^{3/2}.$$

Hence, our choice of parameters must ensure the following inequality:

$$A_{k+1}\sqrt{\tfrac{2}{L_3+M}}\cdot\|\nabla f(x_{k+1})\|_*^{3/2}+a_k\langle\nabla f(x_{k+1}),x-v_k\rangle+\tfrac{N}{12}\|x-v_k\|^3\geq 0,$$

for all $x\in E$. Using inequality (2.3) with $p=3$, $s=a_k\nabla f(x_{k+1})$, and $\sigma=\frac{1}{4}N$, we come to the following condition:

$$A_{k+1}\sqrt{\tfrac{2}{L_3+M}} \;\geq\; \tfrac{2}{3}\sqrt{\tfrac{4}{N}}\,a_k^{3/2}. \tag{4.5}$$

For $k\geq 1$, let us choose

$$\begin{aligned} A_k &= \tfrac{k(k+1)(k+2)}{6}, \\ a_k &= A_{k+1}-A_k \;=\; \tfrac{(k+1)(k+2)(k+3)}{6}-\tfrac{k(k+1)(k+2)}{6} \\ &= \tfrac{(k+1)(k+2)}{2}. \end{aligned} \tag{4.6}$$

Since

$$a_k^{-3/2}A_{k+1} \;=\; \tfrac{2^{3/2}(k+1)(k+2)(k+3)}{6[(k+1)(k+2)]^{3/2}} \;=\; \tfrac{2^{1/2}(k+3)}{3[(k+1)(k+2)]^{1/2}} \;\geq\; \tfrac{1}{3}\sqrt{2},$$

inequality (4.5) leads to the following condition on the parameters:

$$\tfrac{1}{\sqrt{L_3+M}} \;\geq\; \tfrac{2}{\sqrt{N}}.$$

Hence, we can choose

$$M \;=\; 2L_3, \quad N \;=\; 4(L_3+M) \;=\; 12L_3. \tag{4.7}$$

In this case $2L_3+N=14L_3$.

Now we are ready to put all the pieces together.

Accelerated cubic regularization of Newton's method (4.8)

Initialization: Choose $x_0\in E$. Set $M=2L_3$ and $N=12L_3$.

Compute $x_1=T_{L_3}(x_0)$ and define $\psi_1(x)=f(x_1)+\frac{N}{6}\|x-x_0\|^3$.

Iteration k, ($k\geq 1$):

1. Compute $v_k=\arg\min\limits_{x\in E}\psi_k(x)$ and choose $y_k=\frac{k}{k+3}x_k+\frac{3}{k+3}v_k$.

2. Compute $x_{k+1}=T_M(y_k)$ and update

$$\psi_{k+1}(x)=\psi_k(x)+\tfrac{(k+1)(k+2)}{2}\cdot[f(x_{k+1})+\langle\nabla f(x_{k+1}),x-x_{k+1}\rangle].$$

The above discussion proves the following theorem.

Theorem 2 *If sequence $\{x_k\}_{k=1}^{\infty}$ is generated by method (4.8) as applied to the problem (3.1), then for any $k \geq 1$ we have:*

$$f(x_k) - f(x^*) \;\leq\; \frac{14\, L_3\, \|x_0 - x^*\|^3}{k(k+1)(k+2)}, \tag{4.9}$$

where x^* is an optimal solution to the problem.

Proof:
Indeed, we have shown that

$$A_k f(x_k) \;\overset{\mathcal{R}_k^1}{\leq}\; \psi_k^* \;\overset{\mathcal{R}_k^2}{\leq}\; A_k f(x^*) + \tfrac{2L_3+N}{6}\|x_0 - x^*\|^3.$$

Thus, (4.9) follows from (4.6) and (4.7). □

Note that the point v_k can be found in (4.8) by an explicit formula. Consider

$$s_k = \nabla l_k(x).$$

This vector does not depend on x since the function $l_k(x)$ is linear. Then

$$v_k \;=\; x_0 - \sqrt{\tfrac{2}{N}} \cdot \tfrac{B^{-1} s_k}{\|s_k\|_*^{1/2}}.$$

2. Quadratic estimate functions. Similar analysis can be applied to a variant of scheme (4.8) based on quadratic estimate functions. Indeed, let us define

$$\psi_k(x) \;=\; q_k(x) + \tfrac{N}{6}\|x - x_0\|^3, \quad k = 1, 2, \ldots,$$

where $q_k(x)$ are some *quadratic* functions. If we choose

$$x_1 \;=\; T_{L_3}(x_0), \quad q_1(x) \;=\; f_2(x_0; x), \quad A_1 \;=\; 1, \tag{4.10}$$

then, for any $N \geq L_3$ we can guarantee that the conditions

$$A_k f(x_k) \;\overset{(3.8)}{\leq}\; \psi_k^*,$$

$$\psi_k(x) \;\overset{(2.12)}{\leq}\; f(x) + \tfrac{L_3+N}{6}\|x - x_0\|^3, \quad \forall x \in E,$$

hold for $k = 1$. Then, by the same arguments as above, we can see that the rules (4.3) maintain (4.10) for all $k \geq 1$ provided that $N \geq 4(L_3 + M)$. Thus, the new version of (4.8) has a slightly better constant in the estimate of the rate of convergence:

$$f(x_k) - f(x^*) \;\leq\; \frac{13\, L_3\, \|x_0 - x^*\|^3}{k(k+1)(k+2)}.$$

However, the rules for computing v_k become more complicated.

5 Global non-degeneracy for second order schemes

Traditionally, in numerical analysis the term *non-degenerate* is applied to certain classes of efficiently solvable problems. For unconstrained optimization, non-degeneracy of the objective function is usually characterized by a lower bound on the angle between the gradient at point x and the direction pointing to the optimal solution:

$$\alpha(x) \stackrel{\text{def}}{=} \frac{\langle \nabla f(x), x-x^* \rangle}{\|\nabla f(x)\|_* \cdot \|x-x^*\|} \geq \mu(f) > 0, \quad x \in E. \tag{5.1}$$

This condition has a nice geometric interpretation. Moreover, there exists a large class of smooth convex functions, possessing the required property. This is the class of strongly convex functions with Lipschitz-continuous gradient.

Lemma 7 $\mu(f) \geq \frac{2\sqrt{\gamma_2(f)}}{1+\gamma_2(f)} > \sqrt{\gamma_2(f)}$.

Indeed, in view of inequality (2.1.24) in [5], we have

$$\begin{aligned} \langle \nabla f(x), x-x^* \rangle &\geq \tfrac{1}{\sigma_2+L_2}\|\nabla f(x)\|_*^2 + \tfrac{\sigma_2 L_2}{\sigma_2+L_2}\|x-x^*\|^2 \\ &\geq \tfrac{2\sqrt{\sigma_2 L_2}}{\sigma_2+L_2} \cdot \|\nabla f(x)\|_* \cdot \|x-x^*\|, \end{aligned}$$

and this proves the required inequality. □

Note that the complexity of the first-order schemes for the class of smooth strongly convex function can be completely characterized in terms of the condition number γ_2. Indeed, on the one hand, the lower complexity bound for finding an ϵ-solution of any problem from this problem class is proven to be

$$O\left(\frac{1}{\sqrt{\gamma_2}} \ln \frac{L_2 D^2}{\epsilon}\right) \tag{5.2}$$

calls of oracle, where the constant D bounds the distance between the initial point and the optimal solution. On the other hand, there exist simple numerical schemes, which exhibit the required rate of convergence (see Chapter 2 in [5] for details).

What can be said about the complexity of the above problem class for the second order schemes? Surprisingly enough, in this situation it is quite difficult to find any advantages of the condition (5.1). We will discuss the complexity of this class in detail later in Section 6. Now let us present a new non-degeneracy condition, which replaces (5.1) for the second-order methods.

Assume that $\gamma_3(f) = \frac{\sigma_3(f)}{L_3(f)} > 0$. In this case,

$$f(x) - f(x^*) \stackrel{(2.4)}{\leq} \frac{2}{3\sqrt{\sigma_3}} \cdot \|\nabla f(x)\|_*^{3/2} \tag{5.3}$$

Therefore, for method (3.11) we have

$$f(x_k) - f(x_{k+1}) \stackrel{(3.7)}{\geq} \frac{1}{3\sqrt{L_3}}\|\nabla f(x_{k+1})\|_*^{3/2} \stackrel{(5.3)}{\geq} \frac{1}{2}\sqrt{\gamma_3(f)} \cdot (f(x_{k+1}) - f(x^*)). \tag{5.4}$$

Hence, for any $k \geq 1$ we have

$$f(x_k) - f(x^*) \overset{(5.4)}{\leq} \frac{f(x_1)-f^*}{\left(1+\frac{1}{2}\sqrt{\gamma_3(f)}\right)^{k-1}} \overset{(3.8)}{\leq} e^{-\frac{\sqrt{\gamma_3(f)}\cdot(k-1)}{2+\sqrt{\gamma_3(f)}}} \cdot \frac{L_3}{3}\|x_0 - x^*\|^3 \tag{5.5}$$

Thus, the complexity of minimizing a function with *positive* third condition number $\gamma_3(f)$ by method (3.11) is of the order

$$O\left(\frac{1}{\sqrt{\gamma_3(f)}} \ln \frac{L_3 D^3}{\epsilon}\right) \tag{5.6}$$

calls of oracle. The structure of this estimate is similar to that of (5.2). Therefore, we will say that such functions possessing the *global second-order non-degeneracy* property.

Let us demonstrate that the accelerated variant of the Newton's method (4.8) can be used to improve the complexity estimate (5.6). Denote by $\mathcal{A}_k(x_0)$, $k \geq 1$, the point x_k generated by method (4.8) with starting point x_0. Consider the following process.

$$\begin{array}{ll} \textbf{1.} & \text{Define } m = 5\left\lfloor \frac{1}{\gamma_3(f)} \right\rfloor^{1/3}, \text{ and set } y_0 = x_0. \\ \textbf{2.} & \text{For } k \geq 0, \text{ iterate } y_{k+1} = \mathcal{A}_m(y_k). \end{array} \tag{5.7}$$

The performance of this scheme can be derived from the following lemma.

Lemma 8 *For any $k \geq 0$ we have $\|y_{k+1} - x^*\|^3 \leq \frac{1}{e}\|y_k - x^*\|^3$.*

Proof:
Indeed, since $m \geq \left(\frac{42e}{\gamma_3(f)}\right)^{1/3}$, we have

$$\frac{1}{3}\sigma_3\|y_{k+1} - x^*\|^3 \overset{(2.1)}{\leq} f(y_{k+1}) - f(x^*) \overset{(4.9)}{\leq} \frac{14L_3\|y_k - x^*\|^3}{m(m+1)(m+2)} \leq \frac{1}{3e}\sigma_3\|y_k - x^*\|^3.$$

□

Thus,

$$f(T_{L_3}(y_k)) - f(x^*) \overset{(3.8)}{\leq} \frac{L_3}{3}\|y_k - x^*\|^3 \overset{(3.8)}{\leq} \frac{L_3}{3}\|y_0 - x^*\|^3 \cdot e^{-k},$$

and we conclude that an ϵ-solution to our problem can be found by (5.7) in

$$O\left(\frac{1}{[\gamma_3(f)]^{1/3}} \ln\left[\frac{L_3}{\epsilon}\|x_0 - x^*\|^3\right]\right) \tag{5.8}$$

iterations. Unfortunately, since the complexity theory for this problem class is not developed yet, we cannot say how far our results are from the best possible ones.

6 Minimizing strongly convex functions

Let us look now at the complexity of problem (3.1) with

$$\sigma_2(f) \;>\; 0, \quad L_3(f) \;<\; \infty. \tag{6.1}$$

The main advantage of such functions consists in possibility for the Newton's method (3.11) to converge quadratically in a certain neighborhood of the solution. Indeed, for $T = T_{L_3}(x)$ we have

$$\begin{array}{rcl} f(x) - f(T) & \overset{(3.6)}{\geq} & \frac{1}{2}\langle \nabla^2 f(T)(T-x), T-x\rangle \;\overset{(2.5)}{\geq}\; \frac{\sigma_2}{2}\cdot r^2_{L_3}(x) \\ & \overset{(3.4)}{\geq} & \frac{\sigma_2}{2L_3}\cdot \|\nabla f(T)\|_* \;\overset{(2.4)}{\geq}\; \frac{\sigma_2}{2L_3}\cdot [2\sigma_2(f(T)-f(x^*))]^{1/2}. \end{array} \tag{6.2}$$

Hence,

$$f(T) - f(x^*) \;\overset{(6.2)}{\leq}\; \frac{2L_3^2}{\sigma_2^3}(f(x)-f(T))^2 \;\leq\; \frac{2L_3^2}{\sigma_2^3}(f(x)-f(x^*))^2, \tag{6.3}$$

and the region of quadratic convergence of method (3.11) can be defined as

$$\mathcal{Q}_f \;=\; \left\{ x \in E : \; f(x) - f(x^*) \leq \frac{\sigma_2^3}{2L_3^2} \right\}. \tag{6.4}$$

Alternatively, the region of quadratic convergence can be described by the norm of the gradients. Indeed,

$$\frac{\sigma_2}{2}\cdot r^2_{L_3}(x) \;\overset{(2.5)}{\leq}\; \frac{1}{2}\langle \nabla^2 f(T)(T-x), T-x\rangle \;\overset{(3.6)}{\leq}\; f(x) - f(T) \;\leq\; \|\nabla f(x)\|_* \cdot r_{L_3}(x).$$

Thus,

$$\|\nabla f(x)\|_* \;\geq\; \frac{\sigma_2}{2}\cdot r_{L_3}(x) \;\overset{(3.4)}{\geq}\; \frac{\sigma_2}{2}\left[\frac{1}{L_3}\|\nabla f(T)\|_*\right]^{1/2}.$$

Consequently,

$$\|\nabla f(T)\|_* \;\leq\; \frac{4L_3}{\sigma_2^2}\|\nabla f(x)\|_*^2, \tag{6.5}$$

and the required region of quadratic convergence can be defined as

$$\mathcal{Q}_g \;=\; \left\{ x \in E : \; \|\nabla f(x)\|_* \leq \frac{\sigma_2^2}{4L_3} \right\}. \tag{6.6}$$

Thus, the global complexity of problem (3.1), (6.1) is mainly related to the number of iterations, that is required to get from x_0 to the region $\mathcal{Q}_f$ (or, to $\mathcal{Q}_g$). For method (3.11), this value can be estimated from above by

$$O\left(\sqrt{\frac{L_3(f)D}{\sigma_2(f)}}\right), \tag{6.7}$$

where D is defined by (3.12) (see Section 6 in [6]). Let us show that, using the accelerated scheme (4.8), it is possible to improve this complexity bound.

Assume that we know an upper bound for the distance to the solution:

$$\|x_0 - x^*\| \;\leq\; R \quad (\leq D).$$

Consider the following process.

$$
\begin{array}{ll}
\textbf{1.} & \text{Set } y_0 = T_{L_3}(x_0), \text{ and define } m_0 = 5\left[\frac{L_3(f)R}{\sigma_2(f)}\right]^{1/3}. \\
\textbf{2.} & \textbf{while } \|\nabla f(T_{L_3}(y_k))\|_* \geq \frac{\sigma_2^2}{4L_3} \textbf{ do } \{y_{k+1} = \mathcal{A}_{m_k}(y_k),\ m_{k+1} = \frac{1}{2^{1/3}}\, m_k\}.
\end{array}
\tag{6.8}
$$

Theorem 3 *The process (6.8) terminates at most after*

$$
\frac{1}{\ln 4} \ln \left[\frac{8}{3} \cdot \left(\frac{L_3(f)R}{\sigma_2(f)} \right)^3 \right] \tag{6.9}
$$

stages. The total number of Newton steps in all stages does not exceed $4m_0$.

Proof:
Denote $R_k = R \cdot \left(\frac{1}{2}\right)^k$. It is clear that

$$
m_k \;\geq\; 5\left[\frac{L_3(f)R_k}{\sigma_2(f)}\right]^{1/3}, \quad k \geq 0. \tag{6.10}
$$

For $k \geq 0$, let us prove by induction that

$$
\|y_k - x^*\| \;\leq\; R_k. \tag{6.11}
$$

Assume that for some $k \geq 0$ this statement is valid (it is true for $k = 0$). Then,

$$
\begin{aligned}
\frac{\sigma_2}{2}\|y_{k+1} - x^*\|^2 \;&\overset{(2.1)}{\leq}\; f(y_{k+1}) - f(x^*) \;\overset{(4.9)}{\leq}\; \frac{14L_3R_k^3}{m_k(m_k+1)(m_k+2)} \\
&\overset{(6.10)}{\leq}\; \frac{14}{125}\sigma_2 R_k^2 \;\leq\; \frac{1}{8}\sigma_2 R_k^2 \;=\; \frac{1}{2}\sigma_2 R_{k+1}^2.
\end{aligned}
$$

Thus, (6.11) is valid for all $k \geq 0$. On the other hand,

$$
\begin{aligned}
f(y_{k+1}) - f(x^*) \;&\overset{(4.9)}{\leq}\; \frac{14L_3\|y_k - x^*\|^3}{m_k(m_k+1)(m_k+2)} \;\overset{(6.11)}{\leq}\; \frac{14L_3\|y_k - x^*\|^2 R_k}{m_k(m_k+1)(m_k+2)} \\
&\overset{(6.10)}{\leq}\; \frac{1}{8}\sigma_2\|y_k - x^*\|^2 \;\overset{(2.1)}{\leq}\; \frac{1}{4}(f(y_k) - f(x^*)).
\end{aligned}
$$

Hence

$$
\begin{aligned}
\frac{\sigma_2}{2L_3}\|\nabla f(T_{L_3}(y_k))\|_* \;&\overset{(6.2)}{\leq}\; f(y_k) - f(T_{L_3}(y_k)) \;\leq\; f(y_k) - f(x^*) \\
&\leq\; \left(\frac{1}{4}\right)^k (f(y_0) - f(x^*)) \;\overset{(3.8)}{\leq}\; \left(\frac{1}{4}\right)^k \frac{L_3}{3} R^3,
\end{aligned}
$$

and (6.9) follows form (6.6). Finally, the total number of Newton steps does not exceed

$$
\sum_{k=0}^{\infty} m_k \;=\; m_0 \sum_{k=0}^{\infty} \frac{1}{2^{k/3}} \;=\; \frac{m_0}{2^{1/3}-1} \;<\; 4m_0.
$$

□

7 Fake acceleration

Note that the properties of the class of smooth strongly convex functions (6.1) leave some space for erroneous conclusions related to the rate of convergence of the optimization methods in the first stage of the process, aiming to enter the region of quadratic convergence of the Newton's method. Let us demonstrate a possible mistake on a particular example.

Consider a modified version $\mathcal{M}'$ of the method (4.8). The only modification is introduced in Step 2. Now it looks as follows:

2'. Compute $\hat{y}_k = T_M(y_k)$ and update

$$\psi_{k+1}(x) = \psi_k(x) + \tfrac{(k+1)(k+2)}{2} \cdot [f(\hat{y}_k) + \langle \nabla f(\hat{y}_k), x - \hat{y}_k \rangle]. \tag{7.1}$$

Choose $\hat{x}_k : \; f(\hat{x}_k) = \min\{f(x_k), f(\hat{y}_k)\}$. Set $x_{k+1} = T_M(\hat{x}_k)$.

Note that for $\mathcal{M}'$ the statement of Theorem 2 is valid. Moreover, the process now becomes monotone, and, using the same reasoning as in (6.2) and $M = 2L_3$, we obtain

$$f(x_k) - f(x_{k+1}) \;\geq\; f(\hat{x}_k) - f(x_{k+1}) \;\geq\; \tfrac{\sqrt{2}\,\sigma_2^{3/2}}{3L_3} \cdot [f(x_{k+1}) - f(x^*)]^{1/2}. \tag{7.2}$$

Further, let us fix the number of steps N. Define $\hat{k} = \frac{2}{3}N$. Then, in view of (4.9), we can guarantee that

$$f(x_{\hat{k}}) - f(x^*) \;\leq\; \tfrac{3^3 \cdot 7 \cdot L_3 R^3}{2^2 N^3}. \tag{7.3}$$

On the other hand

$$f(x_{\hat{k}}) - f(x^*) \;\geq\; f(x_{\hat{k}}) - f(x_{N+1}) \overset{(7.2)}{\geq} \tfrac{1}{3}N \cdot \tfrac{\sqrt{2}\,\sigma_2^{3/2}}{3L_3} \cdot [f(x_{N+1}) - f(x^*)]^{1/2}. \tag{7.4}$$

Combining (7.3) and (7.4) we obtain

$$f(x_{N+1}) - f(x^*) \;\leq\; \tfrac{3^{10} \cdot 7^2 \cdot L_3^4 \cdot R^6}{2^5 \cdot \sigma_2^3} \cdot N^{-8}. \tag{7.5}$$

As compared with (4.9), the proposed modification looks amazingly efficient. However, that is just an illusion. Indeed, in view of (6.4), in order to enter the region of quadratic convergence of the Newton's method, we need to make the right-hand-side of inequality (7.5) smaller than $\frac{\sigma_2^3}{2L_3^2}$. For that we need

$$O\left(\left[\tfrac{L_3 R}{\sigma_2}\right]^{3/4}\right) \tag{7.6}$$

iterations of $\mathcal{M}'$. This is much worse than the complexity estimate (6.7) of the basic scheme (3.11) even without acceleration (4.8).

Another test could be an estimate for the number of steps, which is necessary for $\mathcal{M}'$ to halve the distance to the minimum. From (7.5) we see that it needs $O\left(\left[\frac{L_3 R}{\sigma_2}\right]^{1/2}\right)$ iterations, which is worse than the corresponding estimate for the method (4.8).

8 Discussion

1. From the complexity results presented in the previous sections we can derive a class of problems, which are *easy* for the second order schemes:

$$\sigma_2(f) \;>\; 0, \quad \sigma_3(f) \;>\; 0, \quad L_3(f) \;<\; \infty. \tag{8.1}$$

For such functions, the second-order methods exhibit a global linear rate of convergence and a local quadratic convergence. In accordance with (5.8) and (6.4), we need

$$O\left(\left[\tfrac{L_3(f)}{\sigma_3(f)}\right]^{1/3} \ln\left[\tfrac{L_3(f)}{\sigma_2(f)}\|x_0 - x^*\|\right]\right) \tag{8.2}$$

iterations of (4.8) to enter the region of quadratic convergence.

Note that the class (8.1) is non-trivial. It contains, for example, all functions

$$\xi_{\alpha,\beta}(x) = \alpha d_2(x) + \beta d_3(x), \quad \alpha, \beta > 0,$$

with parameters

$$\sigma_2(\xi_{\alpha,\beta}) = \alpha, \quad \sigma_3(\xi_{\alpha,\beta}) = \tfrac{1}{2}\beta, \quad L_3(\xi_{\alpha,\beta}) = 2\beta.$$

Moreover, any convex function with Lipschitz-continuous Hessian can be *regularized* by adding an auxiliary function $\xi_{\alpha,\beta}$.

2. For one important class of convex problems, that is

$$\sigma_2(f) \;>\; 0, \quad L_2(f) \;<\; \infty, \quad L_3(f) \;<\; \infty, \tag{8.3}$$

we have actually failed to clarify the situation. The standard theory of the optimal *first-order* methods (see, for example, Section 2.2 in [5])) can bound the number of iterations, that are required to enter the region of quadratic convergence (6.4), as follows:

$$O\left(\left[\tfrac{L_2(f)}{\sigma_2(f)}\right]^{1/2} \ln\left[\tfrac{L_2(f)L_3^2(f)}{\sigma_2^3(f)}\|x_0 - x^*\|^2\right]\right). \tag{8.4}$$

Note that in this estimate the role of the second-order scheme is quite weak: it is used only to establish the bounds of the termination stage. Of course, as it is shown in Section 6, we could use it on the first stage also. However, in this case the size of the optimal solution x^* enters *polynomially* the estimate for the number of iterations. Thus, the following question is still open:

> *For the problem class (8.3), can we get any advantage from the second order schemes being used at the initial stage of minimization process?*

3. From the computational point of view, the results presented in this paper are rather preliminary. For example, we always assume that all necessary constants are known. In practical algorithms, the adaptive strategies for updating these parameters are of crucial importance. However, the author believes that the theory presented here could serve as a useful guideline for such developments.

References

[1] Bennet A.A., Newton's method in general analysis, *Proc. Nat. Ac. Sci. USA*, 1916, **2**, No 10, 592–598.

[2] Conn A.B., Gould N.I.M., Toint Ph.L., *Trust Region Methods*, SIAM, Philadelphia, 2000.

[3] Dennis J.E., Jr., Schnabel R.B., *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, SIAM, Philadelphia, 1996.

[4] Kantorovich L.V., Functional analysis and applied mathematics, *Uspehi Matem. Nauk*, 1948, **3**, No 1, 89–185, (in Russian), translated as N.B.S. Report 1509, Washington D.C., 1952.

[5] Nesterov Yu., *Introductory lectures on convex programming. A basic course*, Kluwer, Boston, 2004.

[6] Yu.Nesterov, B.Polyak. Cubic regularization of Newton method and its global performance. CORE Discussion paper # 2003/41, (2003). Submitted to *Mathematical Programming*.

[7] Ortega J.M., Rheinboldt W.C., *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, NY, 1970.

[8] A. Vladimirov, Yu. Nesterov, Yu. Chekanov. Uniformly convex functionals, (In Russian) *Vestnik Moskovskogo universiteta*, ser. *Vychislit. Matem. i Kibern.*, 1978, No.4, 18-27.