

MINIMIZING UNIFORMLY CONVEX FUNCTIONS BY CUBIC REGULARIZATION OF NEWTON METHOD

Nikita Doikov, Yurii Nesterov

REPRINT | 3235

CORE

Voie du Roman Pays 34, L1.03.01

B-1348 Louvain-la-Neuve

Tel (32 10) 47 43 04

Email: lidam-library@uclouvain.be

<https://uclouvain.be/en/research-institutes/lidam/core/core-reprints.html>



Minimizing Uniformly Convex Functions by Cubic Regularization of Newton Method

Nikita Doikov¹ · Yuri Nesterov²

Received: 16 July 2019 / Accepted: 15 February 2021 / Published online: 10 March 2021
© The Author(s) 2021

Abstract

In this paper, we study the iteration complexity of cubic regularization of Newton method for solving composite minimization problems with uniformly convex objective. We introduce the notion of second-order condition number of a certain degree and justify the linear rate of convergence in a nondegenerate case for the method with an adaptive estimate of the regularization parameter. The algorithm automatically achieves the best possible global complexity bound among different problem classes of uniformly convex objective functions with Hölder continuous Hessian of the smooth part of the objective. As a byproduct of our developments, we justify an intuitively plausible result that the global iteration complexity of the Newton method is always better than that of the gradient method on the class of strongly convex functions with uniformly bounded second derivative.

Keywords Newton method · Cubic regularization · Global complexity bounds · Strong convexity · Uniform convexity

Mathematics Subject Classification 49M15 · 49M37 · 58C15 · 90C25 · 90C30

1 Introduction

A big step in a second-order optimization theory is related to the global complexity guarantees which were justified in [17] for the cubic regularization of the Newton

Communicated by Lionel Thibault.

✉ Nikita Doikov
nikita.doikov@uclouvain.be
Yurii Nesterov
yurii.nesterov@uclouvain.be

¹ ICTEAM (Catholic University of Louvain), Louvain-la-Neuve, Belgium

² CORE (Catholic University of Louvain), Louvain-la-Neuve, Belgium

method. The following results provide a good perspective for the development of this approach, discovering accelerated [14], adaptive [4,5] and universal [10] schemes. The latter methods can automatically adjust to a smoothness properties of the particular objective function. In the same vein, the second-order algorithms for solving a system of nonlinear equations were discovered in [13], and randomized variants for solving large-scale optimization problems were proposed in [7–9,12,18].

Despite to a number of nice properties, global complexity bounds of the cubically regularized Newton method for the cases of strongly convex and uniformly convex objective are not still fully investigated, as well as the notion of second-order non-degeneracy (see discussion in Sect. 5 in [14]). We are going to address this issue in the current paper.

The rest of the paper is organized as follows. Section 2 contains all necessary definitions and main properties of the classes of uniformly convex functions and twice-differentiable functions with Hölder continuous Hessian. We introduce the notion of the *condition number* $\gamma_f(\nu)$ of a certain degree $\nu \in [0, 1]$ and present some basic examples.

In Sect. 3, we describe a general regularized Newton scheme and show the linear rate of convergence for this method on the class of uniformly convex functions with a known degree $\nu \in [0, 1]$ of nondegeneracy. Then, we introduce the adaptive cubically regularized Newton method and collect useful inequalities and properties, which are related to this algorithm.

In Sect. 4, we study global iteration complexity of the cubically regularized Newton method on the classes of uniformly convex functions with Hölder continuous Hessian. We show that for nondegeneracy of *any* degree $\nu \in [0, 1]$, which is formalized by the condition $\gamma_f(\nu) > 0$, the algorithm automatically achieves the linear rate of convergence with the value $\gamma_f(\nu)$ being the main complexity factor.

Finally, in Sect. 5 we compare our complexity bounds with the known bounds for other methods and discuss the results. In particular, we justify an intuitively plausible (but quite a delayed) result that the global complexity of the cubically regularized Newton method is always better than that of the gradient method on the class of strongly convex functions with uniformly bounded second derivative.

2 Uniformly Convex Functions with Hölder Continuous Hessian

Let us start from some notation. In what follows, we denote by \mathbb{E} a finite-dimensional real vector space and by \mathbb{E}^* its dual space, which is a space of linear functions on \mathbb{E} . The value of function $s \in \mathbb{E}^*$ at point $x \in \mathbb{E}$ is denoted by $\langle s, x \rangle$. Let us fix some linear self-adjoint positive-definite operator $B : \mathbb{E} \rightarrow \mathbb{E}^*$ and introduce the following Euclidean norms in the primal and dual spaces:

$$\|x\| := \langle Bx, x \rangle^{1/2}, \quad x \in \mathbb{E}, \quad \|s\|_* := \langle s, B^{-1}s \rangle^{1/2}, \quad s \in \mathbb{E}^*.$$

For any linear operator $A : \mathbb{E} \rightarrow \mathbb{E}^*$, its norm is induced in a standard way:

$$\|A\| := \max_{x \in \mathbb{E}} \{ \|Ax\|_* \mid \|x\| \leq 1 \}.$$

Our goal is to solve the convex optimization problem in the composite form:

$$\min_{x \in \text{dom } F} F(x) := f(x) + h(x), \tag{1}$$

where f is a twice differentiable on its open domain uniformly convex function, and h is a *simple* closed convex function with $\text{dom } h \subseteq \text{dom } f$. *Simple* means that all auxiliary subproblems with an explicit presence of h are easily solvable.

For a smooth function f , its gradient at point x is denoted by $\nabla f(x) \in \mathbb{E}^*$, and its Hessian is denoted by $\nabla^2 f(x) : \mathbb{E} \rightarrow \mathbb{E}^*$. For convex but not necessary differentiable function h , we denote by $\partial h(x) \subset \mathbb{E}^*$ its subdifferential at the point $x \in \text{dom } h$.

We say that differentiable function f is *uniformly convex* of degree $p \geq 2$ on a convex set $C \subseteq \text{dom } f$ if for some constant $\sigma > 0$ it satisfies inequality

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\sigma}{p} \|y - x\|^p, \quad x, y \in C. \tag{2}$$

Uniformly convex functions of degree $p = 2$ are known as *strongly convex*. If inequality (2) holds with $\sigma = 0$, the function f is called just *convex*. The following convenient condition is sufficient for function f to be uniformly convex on a convex set $C \subseteq \text{dom } f$:

Lemma 2.1 Lemma 1 in [14]) *Let for some $\sigma > 0$ and $p \geq 2$ the following inequality holds:*

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \sigma \|x - y\|^p, \quad x, y \in C. \tag{3}$$

Then, function f is uniformly convex of degree p on set C with parameter σ .

From now on, we assume $C := \text{dom } F \subseteq \text{dom } f$. By the composite representation (1), we have for every $x \in \text{dom } F$ and for all $F'(x) \in \partial F(x)$:

$$F(y) \geq F(x) + \langle F'(x), y - x \rangle + \frac{\sigma}{p} \|x - y\|^p, \quad y \in \text{dom } F. \tag{4}$$

Therefore, if $\sigma > 0$, then we can have only one point $x^* \in \text{dom } F$ with $F(x^*) = F^*$, which always exists for F being uniformly convex and closed. A useful consequence of uniform convexity is the following upper bound for the residual.

Lemma 2.2 *Let f be uniformly convex of degree $p \geq 2$ with constant $\sigma > 0$ on set $\text{dom } F$. Then, for every $x \in \text{dom } F$ and for all $F'(x) \in \partial F(x)$ we have*

$$F(x) - F^* \leq \frac{p-1}{p} \left(\frac{1}{\sigma}\right)^{\frac{1}{p-1}} \|F'(x)\|_*^{\frac{p}{p-1}}. \tag{5}$$

Proof In view of (4), bound (5) follows as in the proof of Lemma 3 in [14]. □

It is reasonable to define the best possible constant σ in inequality (3) for a certain degree p . This leads us to a system of constants:

$$\sigma_f(p) := \inf_{\substack{x, y \in \text{dom } F \\ x \neq y}} \frac{\langle \nabla f(x) - \nabla f(y), x - y \rangle}{\|x - y\|^p}, \quad p \geq 2. \tag{6}$$

We prefer to use inequality (3) for the definition of $\sigma_f(p)$, instead of (2), because of its symmetry in x and y . Note that the value $\sigma_f(p)$ also depends on the domain of F . However, we omit this dependence in our notation since it is always clear from the context.

It is easy to see that the univariate function $\sigma_f(\cdot)$ is log-concave. Thus, for all $p_2 > p_1 \geq 2$ we have:

$$\sigma_f(p) \geq (\sigma_f(p_1))^{\frac{p_2-p}{p_2-p_1}} \cdot (\sigma_f(p_2))^{\frac{p-p_1}{p_2-p_1}}, \quad p \in [p_1, p_2]. \quad (7)$$

For a twice-differentiable function f , we say that it has *Hölder continuous Hessian* of degree $\nu \in [0, 1]$ on a convex set $C \subseteq \text{dom } f$, if for some constant \mathcal{H} , it holds:

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq \mathcal{H}\|x - y\|^\nu, \quad x, y \in C. \quad (8)$$

Two simple consequences of (8) are as follows:

$$\|\nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y - x)\|_* \leq \frac{\mathcal{H}\|x - y\|^{1+\nu}}{1 + \nu}, \quad (9)$$

$$|f(y) - Q(x; y)| \leq \frac{\mathcal{H}\|x - y\|^{2+\nu}}{(1 + \nu)(2 + \nu)}, \quad (10)$$

where $Q(x; y)$ is the quadratic model of f at the point x :

$$Q(x; y) := f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \langle \nabla^2 f(x)(y - x), y - x \rangle.$$

In order to characterize the level of smoothness of function f on the set $C := \text{dom } F$, let us define the system of Hölder constants (see [10]):

$$\mathcal{H}_f(\nu) := \sup_{\substack{x, y \in \text{dom } F \\ x \neq y}} \frac{\|\nabla^2 f(x) - \nabla^2 f(y)\|}{\|x - y\|^\nu}, \quad \nu \in [0, 1]. \quad (11)$$

We allow $\mathcal{H}_f(\nu)$ to be equal to $+\infty$ for some ν . Note that function $\mathcal{H}_f(\cdot)$ is log-convex. Thus, any $0 \leq \nu_1 < \nu_2 \leq 1$ such that $\mathcal{H}_f(\nu_i) < +\infty$, $i = 1, 2$, provide us with the following upper bounds for the whole interval:

$$\mathcal{H}_f(\nu) \leq (\mathcal{H}_f(\nu_1))^{\frac{\nu_2 - \nu}{\nu_2 - \nu_1}} \cdot (\mathcal{H}_f(\nu_2))^{\frac{\nu - \nu_1}{\nu_2 - \nu_1}}, \quad \nu \in [\nu_1, \nu_2]. \quad (12)$$

If for some specific $\nu \in [0, 1]$ we have $\mathcal{H}_f(\nu) = 0$, this implies that $\nabla^2 f(x) = \nabla^2 f(y)$ for all $x, y \in \text{dom } F$. In this case restriction, $f|_{\text{dom } F}$ is a quadratic function and we conclude that $\mathcal{H}_f(\nu) = 0$ for all $\nu \in [0, 1]$. At the same time, having two points $x, y \in \text{dom } F$ with $0 < \|x - y\| \leq 1$, we get a simple uniform lower bound for all constants $\mathcal{H}_f(\nu)$:

$$\mathcal{H}_f(\nu) \geq \|\nabla^2 f(x) - \nabla^2 f(y)\|, \quad \nu \in [0, 1].$$

Let us give an example of function, which has Hölder continuous Hessian for all $v \in [0, 1]$.

Example 2.1 For a given $a_i \in \mathbb{E}^*$, $1 \leq i \leq m$, consider the following convex function:

$$f(x) = \ln \left(\sum_{i=1}^m e^{\langle a_i, x \rangle} \right), \quad x \in \mathbb{E}.$$

Let us fix Euclidean norm $\|x\| = \langle Bx, x \rangle^{1/2}$, $x \in \mathbb{E}$, with operator $B := \sum_{i=1}^m a_i a_i^*$. Without loss of generality, we assume that $B \succ 0$ (otherwise we can reduce dimension of the problem). Then,

$$\mathcal{H}_f(0) \leq 1, \quad \mathcal{H}_f(1) \leq 2.$$

Therefore, by (12) we get, for any $v \in [0, 1]$:

$$\mathcal{H}_f(v) \leq 2^v.$$

Proof Denote $\kappa(x) \equiv \sum_{i=1}^m e^{\langle a_i, x \rangle}$. Let us fix arbitrary $x, y \in \mathbb{E}$ and direction $h \in \mathbb{E}$. Then, straightforward computation gives:

$$\begin{aligned} \langle \nabla f(x), h \rangle &= \frac{1}{\kappa(x)} \sum_{i=1}^m e^{\langle a_i, x \rangle} \langle a_i, h \rangle, \\ \langle \nabla^2 f(x)h, h \rangle &= \frac{1}{\kappa(x)} \sum_{i=1}^m e^{\langle a_i, x \rangle} \langle a_i, h \rangle^2 - \left(\frac{1}{\kappa(x)} \sum_{i=1}^m e^{\langle a_i, x \rangle} \langle a_i, h \rangle \right)^2 \\ &= \frac{1}{\kappa(x)} \sum_{i=1}^m e^{\langle a_i, x \rangle} (\langle a_i, h \rangle - \langle \nabla f(x), h \rangle)^2 \geq 0. \end{aligned}$$

Hence, we get

$$\|\nabla^2 f(x)\| = \max_{\|h\| \leq 1} \langle \nabla^2 f(x)h, h \rangle \leq \max_{\|h\| \leq 1} \sum_{i=1}^m \langle a_i, h \rangle^2 = \max_{\|h\| \leq 1} \|h\|^2 = 1.$$

Since all Hessians of function f are positive definite, we conclude that $\mathcal{H}_f(0) \leq 1$. Inequality $\mathcal{H}_f(1) \leq 2$ can be easily obtained from the following representation of the third derivative:

$$\begin{aligned} f'''(x)[h, h, h] &= \frac{1}{\kappa(x)} \sum_{i=1}^m e^{\langle a_i, x \rangle} (\langle a_i, h \rangle - \langle \nabla f(x), h \rangle)^3 \\ &\leq \langle \nabla^2 f(x)h, h \rangle \max_{1 \leq i, j \leq m} \langle a_i - a_j, h \rangle \leq 2\|h\|^3. \end{aligned}$$

□

Let us imagine now that we want to describe the iteration complexity of some method, which solves the composite optimization problem (1) up to an absolute accuracy $\epsilon > 0$ in the function value. We assume that the smooth part f of its objective is uniformly convex and has Hölder continuous Hessians. Which degrees p and ν should be used in our analysis? Suppose that, for the number of *calls of the oracle*, we are interested in obtaining a polynomial-time bound of the form:

$$O \left((\mathcal{H}_f(\nu))^\alpha \cdot (\sigma_f(p))^\beta \cdot \log \frac{F(x_0) - F^*}{\epsilon} \right), \quad \alpha, \beta \neq 0.$$

Denote by $[x]$ the *physical dimension* of variable $x \in \mathbb{E}$, and by $[f]$ the *physical dimension* of the value $f(x)$. Then, we have $[\nabla f(x)] = [f]/[x]$ and $[\nabla^2 f(x)] = [f]/[x]^2$. This gives us

$$[\mathcal{H}_f(\nu)] = \frac{[f]}{[x]^{2+\nu}}, \quad [\sigma_f(p)] = \frac{[f]}{[x]^p}, \quad [(\mathcal{H}_f(\nu))^\alpha \cdot (\sigma_f(p))^\beta] = \frac{[f]^{\alpha+\beta}}{[x]^{\alpha(2+\nu)+\beta p}}.$$

While x and $f(x)$ can be measured in arbitrary physical quantities, the value “number of iterations” *cannot have* physical dimension. This leads to the following relations:

$$\alpha + \beta = 0 \quad \text{and} \quad \alpha(2 + \nu) + \beta p = 0.$$

Therefore, despite to the fact that our function can belong to several problem classes simultaneously, from the physical point of view only one option is available:

$$p = 2 + \nu$$

Hence, for a twice-differentiable convex function f with $\inf_{\nu \in [0,1]} \mathcal{H}_f(\nu) > 0$, we can define only one meaningful *condition number* of degree $\nu \in [0, 1]$:

$$\gamma_f(\nu) := \frac{\sigma_f(2+\nu)}{\mathcal{H}_f(\nu)}. \tag{13}$$

If for some particular ν we have $\mathcal{H}_f(\nu) = +\infty$, then by our definition: $\gamma_f(\nu) = 0$.

It will be shown that the condition number $\gamma_f(\nu)$ serves as a main factor in the global iteration complexity bounds for the regularized Newton method as applied to the problem (1). Let us prove that this number cannot be big.

Lemma 2.3 *Let $\inf_{\nu \in [0,1]} \mathcal{H}_f(\nu) > 0$ and therefore the condition number $\gamma_f(\cdot)$ be well defined. Then,*

$$\gamma_f(\nu) \leq \frac{1}{1+\nu} + \inf_{x,y \in \text{dom } F} \frac{\|\nabla^2 f(x)\|}{\|\nabla^2 f(y) - \nabla^2 f(x)\|}, \quad \nu \in [0, 1]. \tag{14}$$

In the case when $\text{dom } F$ is unbounded: $\sup_{x \in \text{dom } F} \|x\| = +\infty$, then

$$\gamma_f(\nu) \leq \frac{1}{1+\nu}, \quad \nu \in (0, 1]. \tag{15}$$

Proof Indeed, for any $x, y \in \text{dom } F, x \neq y$, we have:

$$\begin{aligned} \sigma_f(2 + \nu) &\stackrel{(6)}{\leq} \frac{\langle \nabla f(y) - \nabla f(x), y - x \rangle}{\|y - x\|^{2+\nu}} \\ &= \frac{\langle \nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y - x), y - x \rangle}{\|y - x\|^{2+\nu}} + \frac{\langle \nabla^2 f(x)(y - x), y - x \rangle}{\|y - x\|^{2+\nu}} \\ &\stackrel{(9)}{\leq} \frac{\mathcal{H}_f(\nu)}{1 + \nu} + \frac{\|\nabla^2 f(x)\|}{\|y - x\|^\nu}. \end{aligned}$$

Now, dividing both sides of this inequality by $\mathcal{H}_f(\nu)$, we get inequality (14) from the definition of $\mathcal{H}_f(\nu)$ (11). Inequality (15) can be obtained by taking the limit $\|y\| \rightarrow +\infty$. □

From inequalities (7) and (12), we can get the following lower bound:

$$\gamma_f(\nu) \geq (\gamma_f(\nu_1))^{\frac{\nu_2 - \nu}{\nu_2 - \nu_1}} \cdot (\gamma_f(\nu_2))^{\frac{\nu - \nu_1}{\nu_2 - \nu_1}}, \quad \nu \in [\nu_1, \nu_2],$$

where $0 \leq \nu_1 < \nu_2 \leq 1$. However, it turns out that in *unbounded case* we can have a nonzero condition number $\gamma_f(\nu)$ only for a *single degree*.

Lemma 2.4 *Let $\text{dom } F$ be unbounded: $\sup_{x \in \text{dom } F} \|x\| = +\infty$. Assume that for a fixed $\nu \in [0, 1]$ we have $\gamma_f(\nu) > 0$. Then,*

$$\gamma_f(\alpha) = 0 \text{ for all } \alpha \in [0, 1] \setminus \{\nu\}.$$

Proof Consider firstly the case: $\alpha > \nu$. From the condition $\gamma_f(\nu) > 0$, we conclude that $\mathcal{H}_f(\nu) < +\infty$. Then, for any $x, y \in \text{dom } F$ we have:

$$\begin{aligned} \frac{\sigma_f(2 + \alpha)\|y - x\|^{2+\alpha}}{2 + \alpha} &\stackrel{(2)}{\leq} f(y) - f(x) - \langle \nabla f(x), y - x \rangle \\ &\stackrel{(10)}{\leq} \frac{1}{2} \langle \nabla^2 f(x)(y - x), (y - x) \rangle + \frac{\mathcal{H}_f(\nu)\|y - x\|^{2+\nu}}{(1 + \nu)(2 + \nu)}. \end{aligned}$$

Dividing both sides of this inequality by $\|y - x\|^{2+\alpha}$ and letting $\|x\| \rightarrow +\infty$, we get $\sigma_f(2 + \alpha) = 0$. Therefore, $\gamma_f(\alpha) = 0$. For the second case, $\alpha < \nu$, we cannot have $\gamma_f(\alpha) > 0$, since the previous reasoning results in $\gamma_f(\nu) = 0$. □

Let us look now at an important example of a uniformly convex function with Hölder continuous Hessian. It is convenient to start with some properties of powers of Euclidean norm.

Lemma 2.5 *For fixed real $p \geq 1$, consider the following function:*

$$f_p(x) = \frac{1}{p} \|x\|^p, \quad x \in \mathbb{E}.$$

1. For $p \geq 2$, function $f_p(\cdot)$ is uniformly convex of degree p :¹⁾

$$\langle \nabla f_p(x) - \nabla f_p(y), x - y \rangle \geq 2^{2-p} \|x - y\|^p, \quad x, y \in \mathbb{E}. \tag{16}$$

2. If $1 \leq p \leq 2$, then function $f_p(\cdot)$ has v -Hölder continuous gradient with $v = p - 1$:

$$\|\nabla f_p(x) - \nabla f_p(y)\|_* \leq 2^{1-v} \|x - y\|^v, \quad x, y \in \mathbb{E}. \tag{17}$$

Proof Firstly, recall two useful inequalities, which are valid for all $a, b \geq 0$:

$$|a^\alpha - b^\alpha| \leq |a - b|^\alpha, \quad \text{when } 0 \leq \alpha \leq 1, \tag{18}$$

$$|a^\alpha - b^\alpha| \geq |a - b|^\alpha, \quad \text{when } \alpha \geq 1. \tag{19}$$

Let us fix arbitrary $x, y \in \mathbb{E}$. The left-hand side of inequality (16) equals

$$\langle \|x\|^{p-2} Bx - \|y\|^{p-2} By, x - y \rangle = \|x\|^p + \|y\|^p - \langle Bx, y \rangle (\|x\|^{p-2} + \|y\|^{p-2}),$$

and we need to verify that it is bigger than $2^{2-p} [\|x\|^2 + \|y\|^2 - 2\langle Bx, y \rangle]^{\frac{p}{2}}$. The case $x = 0$ or $y = 0$ is trivial. Therefore, assume $x \neq 0$ and $y \neq 0$. Denoting $\tau := \frac{\|y\|}{\|x\|}$, $r := \frac{\langle Bx, y \rangle}{\|x\| \cdot \|y\|}$, we have the following statement to prove:

$$1 + \tau^p \geq r\tau(1 + \tau^{p-2}) + 2^{2-p} [1 + \tau^2 - 2r\tau]^{\frac{p}{2}}, \quad \tau > 0, \quad |r| \leq 1.$$

Since the function in the right-hand side is convex in r , we need to check only two marginal cases:

1. $r = 1$: $1 + \tau^p \geq \tau(1 + \tau^{p-2}) + 2^{2-p} |1 - \tau|^p$, which is equivalent to $(1 - \tau)(1 - \tau^{p-1}) \geq 2^{2-p} |1 - \tau|^p$. This is true by (19).
2. $r = -1$: $1 + \tau^p \geq -\tau(1 + \tau^{p-2}) + 2^{2-p} (1 + \tau)^p$, which is equivalent to $(1 + \tau^{p-1}) \geq 2^{2-p} (1 + \tau)^{p-1}$. This is true in view of convexity of function τ^{p-1} for $\tau \geq 0$.

Thus, we have proved (16). Let us prove the second statement. Consider the function $\hat{f}_q(s) = \frac{1}{q} \|s\|_*^q, s \in \mathbb{E}^*$, with $q = \frac{p}{p-1} \geq 2$. In view of our first statement, we have:

$$\langle s_1 - s_2, \nabla \hat{f}_q(s_1) - \nabla \hat{f}_q(s_2) \rangle \geq \left(\frac{1}{2}\right)^{q-2} \|s_1 - s_2\|_*^q, \quad s_1, s_2 \in \mathbb{E}^*. \tag{20}$$

For arbitrary $x_1, x_2 \in \mathbb{E}$, define $s_i = \nabla f_p(x_i) = \frac{Bx_i}{\|x_i\|^{2-p}}, i = 1, 2$. Then $\|s_i\|_* = \|x_i\|^{p-1}$, and consequently,

$$x_i = \|x_i\|^{2-p} B^{-1} s_i = \|s_i\|_*^{\frac{2-p}{p-1}} B^{-1} s_i = \nabla \hat{f}_q(s_i).$$

¹⁾ For the integer values of p , this inequality was proved in [14].

Therefore, substituting these vectors in (20), we get

$$\left(\frac{1}{2}\right)^{q-2} \|\nabla f_p(x_1) - \nabla f_p(x_2)\|_*^q \leq \langle \nabla f_p(x_1) - \nabla f_p(x_2), x_1 - x_2 \rangle.$$

Thus, $\|\nabla f_p(x_1) - \nabla f_p(x_2)\|_* \leq 2^{\frac{q-2}{q-1}} \|x_1 - x_2\|^{\frac{1}{q-1}}$. It remains to note that $\frac{1}{q-1} = p - 1 = v$. \square

Example 2.2 For real $p \geq 2$ and arbitrary $x_0 \in \mathbb{E}$, consider the following function:

$$f(x) = \frac{1}{p} \|x - x_0\|^p = f_p(x - x_0), \quad x \in \mathbb{E}.$$

Then, $\sigma_f(p) = \left(\frac{1}{2}\right)^{p-2}$. Moreover, if $p = 2 + v$ for some $v \in (0, 1]$, then it holds

$$\mathcal{H}_f(v) \leq (1 + v)2^{1-v},$$

and $\mathcal{H}_f(\alpha) = +\infty$, for all $\alpha \in [0, 1] \setminus \{v\}$. Therefore, in this case we have $\gamma_f(v) \geq \frac{1}{2(1+v)}$, and $\gamma_f(\alpha) = 0$ for all $\alpha \in [0, 1] \setminus \{v\}$.

Proof Let us take an arbitrary $x \neq 0$ and set $y := -x$. Then,

$$\langle \nabla f(x) - \nabla f(y), y - x \rangle = \langle \|x\|^{p-2} Bx + \|x\|^{p-2} Bx, 2x \rangle = 4\|x\|^p.$$

On the other hand, $\|y - x\|^p = 2^p \|x\|^p$. Therefore, $\sigma_f(p) \stackrel{(6)}{\leq} 2^{2-p}$, and (16) tells us that this inequality is satisfied as equality.

Let us prove now that $\mathcal{H}_f(v) \leq (1 + v)2^{1-v}$ for $p = 2 + v$ with some $v \in (0, 1]$. This is

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq (1 + v)2^{1-v} \|x - y\|^v, \quad x, y \in \mathbb{E}. \tag{21}$$

The corresponding Hessians can be represented as follows:

$$\nabla^2 f(x) = \|x\|^v B + \frac{v B x x^* B}{\|x\|^{2-v}}, \quad x \in \mathbb{E} \setminus \{0\}, \quad \nabla^2 f(0) = 0.$$

For the case $x = y = 0$, inequality (21) is trivial. Assume now that $x \neq 0$. If $0 \in [x, y]$, then $y = -\beta x$ for some $\beta \geq 0$ and we have:

$$\begin{aligned} \|\nabla^2 f(x) - \nabla^2 f(-\beta x)\| &\leq |1 - \beta^v| (1 + v) \|x\|^v \leq (1 + \beta)^v (1 + v) 2^{1-v} \|x\|^v \\ &= (1 + v) 2^{1-v} \|x - y\|^v, \end{aligned}$$

which is (21). Let $0 \notin [x, y]$. For an arbitrary fixed direction $h \in \mathbb{E}$, we get:

$$\left| \langle (\nabla^2 f(x) - \nabla^2 f(y))h, h \rangle \right| = \left| (\|x\|^v - \|y\|^v) \cdot \|h\|^2 + v \cdot \left(\frac{\langle Bx, h \rangle^2}{\|x\|^{2-v}} - \frac{\langle By, h \rangle^2}{\|y\|^{2-v}} \right) \right|.$$

Consider the points $u = \frac{Bx}{\|x\|^{1-\nu}} = \nabla f_q(x)$ and $v = \frac{By}{\|y\|^{1-\nu}} = \nabla f_q(y)$ with $q = 1 + \nu$. Then,

$$\|x\|^\nu = \|u\|_*, \quad \frac{\langle Bx, h \rangle^2}{\|x\|^{2-\nu}} = \frac{\langle u, h \rangle^2}{\|u\|_*^2} \quad \text{and} \quad \|y\|^\nu = \|v\|_*, \quad \frac{\langle By, h \rangle^2}{\|y\|^{2-\nu}} = \frac{\langle v, h \rangle^2}{\|v\|_*^2}.$$

Therefore,

$$\begin{aligned} & | \langle (\nabla^2 f(x) - \nabla^2 f(y))h, h \rangle | \\ &= \left| (\|u\|_* - \|v\|_*) \cdot \|h\|^2 + \nu \cdot \left(\frac{\langle u, h \rangle^2}{\|u\|_*^2} - \frac{\langle v, h \rangle^2}{\|v\|_*^2} \right) \right|. \end{aligned} \quad (22)$$

Let us estimate the right-hand side of (22) from above. Consider a continuously differentiable univariate function:

$$\begin{aligned} \phi(\tau) &:= \|u(\tau)\|_* \cdot \|h\|^2 + \nu \cdot \frac{\langle u(\tau), h \rangle^2}{\|u(\tau)\|_*^2}, \\ u(\tau) &:= u + \tau(v - u), \quad \tau \in [0, 1]. \end{aligned}$$

Note that

$$\begin{aligned} \phi'(\tau) &= \frac{\langle u(\tau), B^{-1}(v - u) \rangle}{\|u(\tau)\|_*} \cdot \|h\|^2 + \frac{2\nu \langle u(\tau), h \rangle \langle v - u, h \rangle}{\|u(\tau)\|_*^3} \\ &\quad - \frac{\nu \langle u(\tau), h \rangle^2 \langle u(\tau), B^{-1}(v - u) \rangle}{\|u(\tau)\|_*^4} \\ &= \frac{\langle u(\tau), B^{-1}(v - u) \rangle}{\|u(\tau)\|_*} \cdot \underbrace{\left(\|h\|^2 - \frac{\nu \langle u(\tau), h \rangle^2}{\|u(\tau)\|_*^2} \right)}_{\geq 0} + \frac{2\nu \langle u(\tau), h \rangle \langle v - u, h \rangle}{\|u(\tau)\|_*^3}. \end{aligned}$$

Denote $\gamma := \frac{\langle u(\tau), h \rangle}{\|u(\tau)\|_* \cdot \|h\|} \in [-1, 1]$. Then,

$$|\phi'(\tau)| \leq \|v - u\|_* \cdot \|h\|^2 \cdot (1 - \nu\gamma^2 + 2\nu|\gamma|) \leq (1 + \nu) \cdot \|v - u\|_* \cdot \|h\|^2.$$

Thus, we have:

$$| \langle (\nabla^2 f(x) - \nabla^2 f(y))h, h \rangle | = |\phi(1) - \phi(0)| \leq (1 + \nu) \cdot \|v - u\|_* \cdot \|h\|^2. \quad (23)$$

It remains to use the definition of u and v and apply inequality (17) with $p = q$. Thus, we have proved, that for $p = 2 + \nu$ the Hessian of f is Hölder continuous of degree ν . At the same time, taking $y = 0$, we get $\|\nabla^2 f(x) - \nabla^2 f(y)\| = \|\nabla^2 f(x)\| = (1 + \nu)\|x\|^\nu$. These values cannot be uniformly bounded in $x \in \mathbb{E}$ by any multiple of $\|x\|^\alpha$ with $\alpha \neq \nu$. So, the Hessian of f is *not* Hölder continuous for any degree different from $2 + \nu$. \square

Remark 2.1 Inequalities (16) and (17) have the following symmetric consequences:

$$\begin{aligned}
 p \geq 2 &\Rightarrow \|\nabla f_p(x) - \nabla f_p(y)\|_* \geq 2^{2-p} \|x - y\|^{p-1}, \\
 p \leq 2 &\Rightarrow \|\nabla f_p(x) - \nabla f_p(y)\|_* \leq 2^{2-p} \|x - y\|^{p-1},
 \end{aligned}$$

which are valid for all $x, y \in \mathbb{E}$.

3 Regularized Newton Method

Let us start from the case when we know that for a specific $\nu \in [0, 1]$ function f has Hölder continuous Hessian: $\mathcal{H}_f(\nu) < +\infty$. Then, from (10), we have the global upper bound for the objective function:

$$F(y) \leq M_{\nu,H}(x; y) := Q(x; y) + \frac{H\|x-y\|^{2+\nu}}{(1+\nu)(2+\nu)} + h(y), \quad x, y \in \text{dom } F,$$

where $H > 0$ is large enough: $H \geq \mathcal{H}_f(\nu)$. Thus, it is natural to employ the minimum of a regularized quadratic model:

$$T_{\nu,H}(x) := \operatorname{argmin}_{y \in \text{dom } F} M_{\nu,H}(x; y), \quad M_{\nu,H}^*(x) := \min_{y \in \text{dom } F} M_{\nu,H}(x; y),$$

and define the following general iteration process [10]:

$$\boxed{x_{k+1} := T_{\nu,H_k}(x_k), \quad k \geq 0} \tag{24}$$

where the value H_k is chosen either to be a constant from the interval $[0, 2\mathcal{H}_f(\nu)]$ or by some adaptive procedure.

For the class of uniformly convex functions of degree $p = 2 + \nu$, we can justify the following global convergence result for this process.

Theorem 3.1 *Assume that for some $\nu \in [0, 1]$ we have $0 < \mathcal{H}_f(\nu) < +\infty$ and $\sigma_f(2 + \nu) > 0$. Let the coefficients $\{H_k\}_{k \geq 0}$ in the process (24) satisfy the following conditions:*

$$0 \leq H_k \leq \beta \mathcal{H}_f(\nu), \quad F(x_{k+1}) \leq M_{\nu,H_k}^*(x_k), \quad k \geq 0, \tag{25}$$

with some constant $\beta \geq 0$. Then, for the sequence $\{x_k\}_{k \geq 0}$ generated by the process we have:

$$F(x_{k+1}) - F^* \leq \left(1 - \frac{1+\nu}{2+\nu} \cdot \min\left\{\frac{\gamma_f(\nu)(1+\nu)}{(1+\beta)(2+\nu)}, 1\right\}^{\frac{1}{1+\nu}}\right) (F(x_k) - F^*). \tag{26}$$

Thus, the rate of convergence is linear and for reaching the gap $F(x_K) - F^* \leq \varepsilon$ it is enough to perform $K = \left\lceil \frac{2+\nu}{1+\nu} \cdot \max\left\{\frac{(1+\beta)(2+\nu)}{\gamma_f(\nu)(1+\nu)}, 1\right\}^{\frac{1}{1+\nu}} \log \frac{F(x_0) - F^*}{\varepsilon} \right\rceil$ iterations.

Proof As in the proof of Theorem 3.1 in [10], from (25) one can see that

$$F(x_{k+1}) \leq F(x_k) - \alpha (F(x_k) - F^*) + \alpha^{2+v} \frac{(1+\beta)\mathcal{H}_f(v)\|x_k - x^*\|^{2+v}}{(1+v)(2+v)},$$

for any $\alpha \in [0, 1]$. Then, taking into account the uniform convexity (4), we get

$$F(x_{k+1}) \leq F(x_k) - \left(\alpha - \alpha^{2+v} \frac{(1+\beta)\mathcal{H}_f(v)}{(1+v)\sigma_f(2+v)} \right) (F(x_k) - F^*).$$

The minimum of the right-hand side is attained at $\alpha^* = \min\left\{ \frac{\gamma_f(v)(1+v)}{(2+v)(1+\beta)}, 1 \right\}^{\frac{1}{1+v}}$. Plugging this value into the bound above, we get inequality (26). \square

Unfortunately, in practice it is difficult to decide on an appropriate value of $v \in [0, 1]$ with $\mathcal{H}_f(v) < +\infty$. Therefore, it is interesting to develop the *universal methods* which are not based on some particular parameters. Recently, it was shown [10] that one good choice for such universal scheme is the cubic regularization of the Newton Method [17]. This is actually the process (24) with the fixed parameter $v = 1$. For this choice, in the rest part of the paper we omit the corresponding index in the definitions of all necessary objects: $M_H(x; y) := M_{1,H}(x; y)$, $T_H(x) := T_{1,H}(x)$, and $M_H^*(x) := M_{1,H}^*(x) = M_H(x; T_H(x))$. The adaptive scheme of our method with dynamic estimation of the constant H is as follows.

Algorithm 1: Adaptive Cubic Regularization of Newton Method

Initialization. Choose $x_0 \in \text{dom } F$, $H_0 > 0$.

Iteration $k \geq 0$.

1: Find the minimal integer $i_k \geq 0$ such that $F(T_{H_k 2^{i_k}}(x_k)) \leq M_{H_k 2^{i_k}}^*(x_k)$.

2: Perform the Cubic Step: $x_{k+1} = T_{H_k 2^{i_k}}(x_k)$.

3: Set $H_{k+1} := 2^{i_k - 1} H_k$.

Let us present the main properties of the composite Cubic Newton step $x \mapsto T_H(x)$. Denote

$$r_H(x) := \|T_H(x) - x\|.$$

Since point $T_H(x)$ is a minimum of strictly convex function $M_H(x; \cdot)$, it satisfies the following first-order optimality condition:

$$\begin{aligned} \langle \nabla f(x) + \nabla^2 f(x)(T_H(x) - x) + \frac{Hr_H(x)}{2} B(T_H(x) - x), y - T_H(x) \rangle \\ + h(y) \geq h(T_H(x)), \quad y \in \text{dom } F. \end{aligned} \tag{27}$$

In other words, the vector

$$h'(T_H(x)) := -\nabla f(x) - \nabla^2 f(x)(T_H(x) - x) - \frac{Hr_H(x)}{2} B(T_H(x) - x)$$

belongs to the subdifferential of h :

$$h'(T_H(x)) \in \partial h(T_H(x)). \tag{28}$$

Computation of a point $T = T_H(x)$, satisfying condition (28), requires some standard techniques of Convex Optimization and Linear Algebra (see [1,3,16,17]). Arithmetical complexity of such a procedure is usually similar to that of the standard Newton step.

Plugging into (27) $y := x \in \text{dom } F$, we get:

$$\begin{aligned} &\langle \nabla f(x), x - T_H(x) \rangle \\ &\geq \langle \nabla^2 f(x)(T_H(x) - x), T_H(x) - x \rangle + \frac{Hr_H^3(x)}{2} + h(T_H(x)) - h(x). \end{aligned} \tag{29}$$

Thus, we obtain the following bound for the minimal value $M_H^*(x)$ of the cubic model:

$$\begin{aligned} M_H^*(x) &\stackrel{(29)}{\leq} f(x) - \frac{1}{2} \langle \nabla^2 f(x)(T_H(x) - x), T_H(x) - x \rangle - \frac{Hr_H^3(x)}{3} + h(x) \\ &= F(x) - \frac{1}{2} \langle \nabla^2 f(x)(T_H(x) - x), T_H(x) - x \rangle - \frac{Hr_H^3(x)}{3}. \end{aligned} \tag{30}$$

If for some value $\nu \in [0, 1]$ the Hessian is Hölder continuous: $\mathcal{H}_f(\nu) < +\infty$, then by (9) and (28) we get the following bound for the subgradient:

$$F'(T_H(x)) := \nabla f(T_H(x)) + h'(T_H(x))$$

at the new point:

$$\begin{aligned} &\|F'(T_H(x))\|_* \\ &\leq \|\nabla f(T_H(x)) - \nabla f(x) - \nabla^2 f(x)(T_H(x) - x)\|_* + \frac{Hr_H^2(x)}{2} \\ &\stackrel{(9)}{\leq} \frac{\mathcal{H}_f(\nu)r_H^{1+\nu}(x)}{1+\nu} + \frac{Hr_H^2(x)}{2} = r_H^{1+\nu}(x) \cdot \left(\frac{\mathcal{H}_f(\nu)}{1+\nu} + \frac{Hr_H^{1-\nu}(x)}{2} \right). \end{aligned} \tag{31}$$

One of the main strong point of the classical Newton’s is its local *quadratic convergence* for the class of strongly convex functions with Lipschitz continuous Hessian: $\sigma_f(2) > 0$ and $0 < \mathcal{H}_f(1) < +\infty$ (see, for example, [15]). This property holds for the cubically regularized Newton as well [14,17]. Indeed, ensuring $F(T_H(x)) \leq M_H^*(x)$ as in Algorithm 1, and having $H \leq \beta \mathcal{H}_f(1)$ with some $\beta \geq 0$, we get:

$$\begin{aligned} F(T_H(x)) - F^* &\stackrel{(5)}{\leq} \frac{1}{2\sigma_f(2)} \|F'(T_H(x))\|_*^2 \stackrel{(31)}{\leq} \frac{(1 + \beta)^2 \mathcal{H}_f^2(1)}{8\sigma_f(2)} r_H^4(x) \\ &\leq \frac{(1 + \beta)^2 \mathcal{H}_f^2(1)}{8\sigma_f^3(2)} \langle \nabla^2 f(x)(T_H(x) - x), T_H(x) - x \rangle^2 \\ &\stackrel{(30)}{\leq} \frac{(1 + \beta)^2 \mathcal{H}_f^2(1)}{2\sigma_f^3(2)} (F(x) - F^*)^2. \end{aligned}$$

And the region of quadratic convergence is as follows:

$$\mathcal{Q} = \left\{ x \in \text{dom } F : F(x) - F^* \leq \frac{2\sigma_f^3(2)}{(1 + \beta)^2 \mathcal{H}_f^2(1)} \right\}.$$

After reaching it, the method starts to double the right digits of the answer at every step, and this cannot last for a long time. Therefore, from now on we are mainly interested in the *global complexity bounds* of Algorithm 1, which work for an arbitrary starting point x_0 .

For noncomposite case, as it was shown in [10], if for some $\nu \in [0, 1]$ we have $0 < \mathcal{H}_f(\nu) < +\infty$ and the objective is just *convex*, then Algorithm 1 with small initial parameter H_0 generates a solution \hat{x} with $f(\hat{x}) - f^* \leq \varepsilon$ in $\mathcal{O}\left(\left(\frac{\mathcal{H}_f(\nu)D_0^{2+\nu}}{\varepsilon}\right)^{\frac{1}{1+\nu}}\right)$ iterations, where $D_0 := \max_x \{\|x - x^*\| : f(x) \leq f(x_0)\}$. Thus, the method in [10] has a sublinear rate of convergence on the class of convex functions with Hölder continuous Hessian. It can *automatically adapt* to the actual level of smoothness. In what follows we show that the same algorithm achieves linear rate of convergence for the class of *uniformly convex* functions of degree $p = 2 + \nu$, namely for functions with strictly positive condition number: $\sup_{\nu \in [0, 1]} \gamma_f(\nu) > 0$.

In the remaining part of the paper, we usually assume that the smooth part of our objective is not *purely quadratic*. This is equivalent to the condition $\inf_{\nu \in [0, 1]} \mathcal{H}_f(\nu) > 0$. However, to conclude this section, let us briefly discuss the case $\min_{\nu \in [0, 1]} \mathcal{H}_f(\nu) = 0$. If we would know in advance that f is a convex quadratic function, then no regularization is needed since a single step $x \mapsto T_H(x)$ with $H := 0$ solves the problem. However, if our function is given by a black-box oracle and we do not know a priori that its smooth part is quadratic, then we can still use Algorithm 1. For this case, we prove the following simple result.

Proposition 3.1 *Let $A : \mathbb{E} \rightarrow \mathbb{E}^*$ be a self-adjoint positive semidefinite linear operator and $b \in \mathbb{E}^*$. Assume that $f(x) := \frac{1}{2}\langle Ax, x \rangle - \langle b, x \rangle$, and the minimum $x^* \in \text{Argmin}\{F(x) := f(x) + h(x)\}$ does exist. Then, in order to get $F(x_K) - F^* \leq \varepsilon$ with arbitrary $\varepsilon > 0$, it is enough to perform*

$$K = \left\lceil \log_2 \frac{H_0 \|x_0 - x^*\|^3}{6\varepsilon} + 1 \right\rceil \tag{32}$$

iterations of Algorithm 1.

Proof In our case, the quadratic model coincides with the smooth part of the objective: $Q(x; y) \equiv f(y)$, $x, y \in \mathbb{E}$. Therefore, at every iteration $k \geq 0$ of Algorithm 1 we have $i_k = 0$ and $H_k = 2^{-k}H_0$. Note that $x_{k+1} = T_{2^{-k}H_0}(x_k) = \text{argmin}_{y \in \text{dom } F} \{F(y) +$

$$\frac{2^{-k}H_0}{6} \|y - x_k\|^3\}, \text{ and}$$

$$F(x_{k+1}) \leq F(y) + \frac{2^{-k}H_0}{6} \|y - x_k\|^3, \quad y \in \text{dom } F. \tag{33}$$

Let us prove that $\|x_{k+1} - x^*\| \leq \|x_k - x^*\|$ for all $k \geq 0$. If this is true, then plugging $y \equiv x^*$ into (33), we get: $F(x_{k+1}) - F^* \leq 2^{-k} \frac{H_0}{6} \|x_0 - x^*\|^3$ which results in the estimate (32). Indeed,

$$\begin{aligned} \|x_k - x^*\|^2 &= \|(x_k - x_{k+1}) + (x_{k+1} - x^*)\|^2 \\ &= \|x_{k+1} - x^*\|^2 + \|x_k - x_{k+1}\|^2 + 2\langle B(x_k - x_{k+1}), x_{k+1} - x^* \rangle, \end{aligned}$$

and it is enough to show that $\langle B(x_k - x_{k+1}), x^* - x_{k+1} \rangle \leq 0$. Since x_{k+1} satisfies the first-order optimality condition:

$$-2^{-(k+1)} H_0 \|x_{k+1} - x_k\| B(x_{k+1} - x_k) := F'(x_{k+1}) \in \partial F(x_{k+1}), \tag{34}$$

we have:

$$\begin{aligned} &\langle B(x_k - x_{k+1}), x^* - x_{k+1} \rangle \\ &\stackrel{(34)}{=} \frac{2^{k+1}}{H_0 \|x_k - x_{k+1}\|} \langle F'(x_{k+1}), x^* - x_{k+1} \rangle \leq 0, \end{aligned}$$

where the last inequality follows from the convexity of the objective. □

4 Complexity Results for Uniformly Convex Functions

In this section, we are going to justify the global linear rate of convergence of Algorithm 1 for a class of twice differentiable uniformly convex functions with Hölder continuous Hessian. Universality of this method is ensured by the adaptive estimation of the parameter H over the whole sequence of iterations. It is important to distinguish two cases: $H_{k+1} < H_k$ and $H_{k+1} \geq H_k$.

First, we need to estimate the progress in the objective function after minimizing the cubic model. There are two different situations here:

$$\text{either } Hr_H^{1-\nu}(x) \leq \frac{2\mathcal{H}_f(\nu)}{1+\nu}, \text{ or } Hr_H^{1-\nu}(x) > \frac{2\mathcal{H}_f(\nu)}{1+\nu}.$$

Lemma 4.1 *Let $0 < \mathcal{H}_f(\nu) < +\infty$ and $\sigma_f(2 + \nu) > 0$ for some $\nu \in [0, 1]$. Then, for arbitrary $x \in \text{dom } F$ and $H > 0$ we have:*

$$\begin{aligned} &F(x) - M_H^*(x) \\ &\geq \min \left[(F(x) - F^*) \cdot \frac{(1 + \nu)}{(2 + \nu)} \cdot \min \left\{ \left(\frac{(1 + \nu)\gamma_f(\nu)}{2(2 + \nu)} \right)^{\frac{1}{1+\nu}}, 1 \right\}, \right. \\ &\quad \left. (F(T_H(x)) - F^*)^{\frac{3(1+\nu)}{2(2+\nu)}} \cdot \left(\frac{2 + \nu}{1 + \nu} \right)^{\frac{3(1+\nu)}{2(2+\nu)}} \cdot \frac{(\sigma_f(2 + \nu))^{\frac{3}{2(2+\nu)}}}{3\sqrt{H}} \right]. \tag{35} \end{aligned}$$

Proof Let us consider two cases. 1) $Hr_H^{1-\nu}(x) \leq \frac{2\mathcal{H}_f(\nu)}{1+\nu}$. Then, for arbitrary $y \in \text{dom } F$, we have:

$$\begin{aligned} M_H^*(x) &:= Q(x; T_H(x)) + \frac{H}{6} \|T_H(x) - x\|^3 + h(T_H(x)) \\ &\leq Q(x; y) + \frac{Hr_H^{1-\nu}(x)\|y - x\|^{2+\nu}}{2(2 + \nu)} + h(y) \\ &\stackrel{(10)}{\leq} F(y) + \frac{\mathcal{H}_f(\nu)\|y - x\|^{2+\nu}}{(1 + \nu)(2 + \nu)} + \frac{Hr_H^{1-\nu}(x)\|y - x\|^{2+\nu}}{2(2 + \nu)} \\ &\leq F(y) + \frac{2\mathcal{H}_f(\nu)\|y - x\|^{2+\nu}}{(1 + \nu)(2 + \nu)}, \end{aligned}$$

where the first inequality follows from the fact, that

$$T_H(x) = \underset{y \in \text{dom } F}{\text{argmin}} \left\{ Q(x; y) + \frac{Hr_H^{1-\nu}(x)\|y - x\|^{2+\nu}}{2(2 + \nu)} + h(y) \right\}.$$

Let us restrict y to the segment: $y = \alpha x^* + (1 - \alpha)x$, with $\alpha \in [0, 1]$. Taking into account the uniform convexity, we get:

$$\begin{aligned} M_H^*(x) &\leq F(x) - \alpha (F(x) - F^*) + \alpha^{2+\nu} \frac{2\mathcal{H}_f(\nu)\|x^* - x\|^{2+\nu}}{(1 + \nu)(2 + \nu)} \\ &\stackrel{(4)}{\leq} F(x) - \left(\alpha - \alpha^{2+\nu} \frac{2\mathcal{H}_f(\nu)}{(1 + \nu)\sigma_f(2 + \nu)} \right) (F(x) - F^*). \end{aligned}$$

The minimum of the right-hand side is attained at $\alpha^* = \min\left\{ \frac{(1+\nu)\gamma_f(\nu)}{2(2+\nu)}, 1 \right\}^{\frac{1}{1+\nu}}$. Plugging this value into the bound, we have:

$$\begin{aligned} M_H^*(x) &\leq F(x) - \min\left\{ \left(\frac{(1 + \nu)\gamma_f(\nu)}{2(2 + \nu)} \right)^{1/(1+\nu)}, 1 \right\} \\ &\quad \cdot \frac{(1 + \nu)}{(2 + \nu)} \cdot (F(x) - F^*), \end{aligned}$$

and this is the first argument of the minimum in (35).

2) $Hr_H^{1-\nu}(x) > \frac{2\mathcal{H}_f(\nu)}{1+\nu}$. By (31), we have the bound:

$$\|F'(T_H(x))\|_* < Hr_H^2(x). \tag{36}$$

Using the fact that $\nabla^2 f(x) \geq 0$, we get the second argument of the minimum:

$$F(x) - M_H^*(x) \stackrel{(30)}{\geq} \frac{Hr_H^3(x)}{3} \stackrel{(36)}{\geq} \frac{\|F'(T_H(x))\|_*^{\frac{3}{2}}}{3\sqrt{H}}$$

$$\stackrel{(5)}{\geq} \left(\frac{2 + \nu}{1 + \nu}\right)^{\frac{3(1+\nu)}{2(2+\nu)}} \cdot \frac{(\sigma_f(2 + \nu))^{2(2+\nu)}}{3\sqrt{H}} \cdot (F(T_H(x)) - F^*)^{\frac{3(1+\nu)}{2(2+\nu)}}.$$

□

Denote by $\kappa_f(\nu)$ the following auxiliary value:

$$\begin{aligned} \kappa_f(\nu) &:= \frac{\mathcal{H}_f(\nu)^{\frac{2}{1+\nu}}}{(\sigma_f(2 + \nu))^{\frac{1-\nu}{(1+\nu)(2+\nu)}}} \\ &\cdot \frac{6 \cdot (8 + \nu)^{\frac{1-\nu}{1+\nu}}}{((1 + \nu)(2 + \nu))^{\frac{2}{1+\nu}}} \cdot \left(\frac{1 + \nu}{2 + \nu}\right)^{\frac{1-\nu}{2+\nu}}, \quad \nu \in [0, 1]. \end{aligned} \tag{37}$$

The next lemma shows what happens when parameter H is increasing during the iterations.

Lemma 4.2 *Assume that for a fixed $x \in \text{dom } F$ the parameter $H > 0$ is such that:*

$$F(T_H(x)) > M_H^*(x). \tag{38}$$

If for some $\nu \in [0, 1]$, we have $\sigma_f(2 + \nu) > 0$, then it holds:

$$H (F(T_{2H}(x)) - F^*)^{\frac{1-\nu}{2+\nu}} < \kappa_f(\nu). \tag{39}$$

Proof Firstly, let us prove that from (38) we have:

$$Hr_H^{1-\nu}(x) < \frac{6\mathcal{H}_f(\nu)}{(1+\nu)(2+\nu)}. \tag{40}$$

Assuming by contradiction, $Hr_H^{1-\nu}(x) \geq \frac{6\mathcal{H}_f(\nu)}{(1+\nu)(2+\nu)}$, we get:

$$\begin{aligned} M_H^*(x) &:= \frac{H\|T_H(x) - x\|^3}{6} + Q(x; T_H(x)) + h(T_H(x)) \\ &\geq \frac{\mathcal{H}_f(\nu)\|T_H(x) - x\|^{2+\nu}}{(1 + \nu)(2 + \nu)} + Q(x; T_H(x)) + h(T_H(x)) \\ &\stackrel{(10)}{\geq} F(T_H(x)), \end{aligned}$$

which contradicts (38). Secondly, by its definition, $M_H^*(x)$ is a concave function of H . Therefore, its derivative $\frac{d}{dH} M_H^*(x) = \frac{1}{6}r_H^3(x)$ is non-increasing. Hence, it holds:

$$r_{2H}(x) \leq r_H(x) \stackrel{(40)}{<} \left(\frac{6\mathcal{H}_f(\nu)}{(1+\nu)(2+\nu)H}\right)^{\frac{1}{1-\nu}}. \tag{41}$$

Finally, by the smoothness and the uniform convexity, we obtain:

$$\begin{aligned}
 H(F(T_{2H}(x)) - F^*) &\stackrel{(5)}{\leq} H\left(\frac{1+\nu}{2+\nu}\left(\frac{1}{\sigma_f(2+\nu)}\right)^{\frac{1}{1+\nu}}\right)^{\frac{1-\nu}{2+\nu}} \|F'(T_{2H}(x))\|_*^{\frac{1-\nu}{1+\nu}} \\
 &\stackrel{(31)}{\leq} H\left(\frac{1+\nu}{2+\nu}\left(\frac{1}{\sigma_f(2+\nu)}\right)^{\frac{1}{1+\nu}}\right)^{\frac{1-\nu}{2+\nu}} \left(r_{2H}^{1+\nu}(x) \cdot \left(\frac{\mathcal{H}_f(\nu)}{1+\nu} + Hr_{2H}^{1-\nu}(x)\right)\right)^{\frac{1-\nu}{1+\nu}} \\
 &\stackrel{(41)}{<} H\left(\frac{1+\nu}{2+\nu}\left(\frac{1}{\sigma_f(2+\nu)}\right)^{\frac{1}{1+\nu}}\right)^{\frac{1-\nu}{2+\nu}} \left(r_{2H}^{1+\nu}(x) \cdot \frac{(8+\nu)\mathcal{H}_f(\nu)}{(1+\nu)(2+\nu)}\right)^{\frac{1-\nu}{1+\nu}} \\
 &\stackrel{(41)}{<} \left(\frac{1+\nu}{2+\nu}\left(\frac{1}{\sigma_f(2+\nu)}\right)^{\frac{1}{1+\nu}}\right)^{\frac{1-\nu}{2+\nu}} \left(\frac{\mathcal{H}_f(\nu)}{(1+\nu)(2+\nu)}\right)^{\frac{2}{1+\nu}} 6(8+\nu)^{\frac{1-\nu}{1+\nu}} =: \kappa_f(\nu).
 \end{aligned}$$

□

We are ready to prove the main result of this paper.

Theorem 4.1 *Assume that for a fixed $\nu \in [0, 1]$ we have $0 < \mathcal{H}_f(\nu) < +\infty$ and $\sigma_f(2+\nu) > 0$. Let parameter H_0 in Algorithm 1 be small enough:*

$$H_0 \leq \frac{\kappa_f(\nu)}{(F(x_0) - F^*)^{(1-\nu)/(2+\nu)}}, \tag{42}$$

where $\kappa_f(\nu)$ is defined by (37). Let the sequence $\{x_k\}_{k=0}^K$ generated by the method satisfy condition:

$$F(T_{H_k 2^j}(x_k)) - F^* \geq \varepsilon > 0, \quad 0 \leq j \leq i_k, \quad 0 \leq k \leq K - 1. \tag{43}$$

Then, for every $0 \leq k \leq K - 1$, we have:

$$\begin{aligned}
 &F(x_{k+1}) - F^* \\
 &\leq \left(1 - \min\left\{\frac{(2+\nu)((1+\nu)(2+\nu))^{1/(1+\nu)}(\gamma_f(\nu))^{\frac{1}{1+\nu}}}{(1+\nu)6^{3/2} \cdot 2^{1/2} \cdot (8+\nu)^{(1-\nu)/(2+2\nu)}}, \frac{1}{2}\right\}\right) \cdot (F(x_k) - F^*).
 \end{aligned} \tag{44}$$

Therefore, the rate of convergence is linear, and

$$K \leq \max\left\{(\gamma_f(\nu))^{\frac{-1}{1+\nu}} \cdot \frac{1+\nu}{2+\nu} \cdot \frac{6^{3/2} \cdot 2^{1/2} \cdot (8+\nu)^{(1-\nu)/(2+2\nu)}}{((1+\nu)(2+\nu))^{1/(1+\nu)}}, 1\right\} \cdot \log \frac{F(x_0) - F^*}{\varepsilon}.$$

Moreover, we have the following bound for the total number of oracle calls N_K during the first K iterations:

$$N_K \leq 2K + \log_2 \frac{\kappa_f(\nu)}{\varepsilon^{(1-\nu)/(2+\nu)}} - \log_2 H_0. \tag{45}$$

Proof The proof is based on Lemmas 4.1 and 4.2, and monotonicity of the sequence $\{F(x_k)\}_{k \geq 0}$. Firstly, we need to show that every iteration of the method is well-defined. Namely, we are going to verify that for a fixed $0 \leq k \leq K - 1$, there exists a finite integer $\ell \geq 0$ such that either $F(T_{H_k 2^\ell}(x_k)) \leq M_{H_k 2^\ell}^*(x_k)$ or $F(T_{H_k 2^{\ell+1}}(x_k)) - F^* < \varepsilon$. Indeed, let us set

$$\ell := \max \left\{ 0, \log_2 \left[\frac{\kappa_f(v)}{H_k \varepsilon^{(1-v)/(2+v)}} \right] \right\}, \quad \text{and } H := H_k 2^\ell \geq \frac{\kappa_f(v)}{\varepsilon^{(1-v)/(2+v)}}. \tag{46}$$

Then, if we have both $F(T_H(x_k)) > M_H^*(x_k)$ and $F(T_{2H}(x_k)) - F^* \geq \varepsilon$, we get by Lemma 4.2:

$$H \stackrel{(39)}{<} \frac{\kappa_f(v)}{(F(T_{2H}(x_k)) - F^*)^{(1-v)/(2+v)}} \leq \frac{\kappa_f(v)}{\varepsilon^{(1-v)/(2+v)}},$$

which contradicts (46). Therefore, if we are unable to find the value $0 \leq i_k \leq \ell$ (see line 1 of Algorithm) in a finite number of steps, that only means we have already solved the problem up to accuracy ε .

Now, let us show that for every $0 \leq k \leq K$ it holds:

$$H_k (F(x_k) - F^*)^{\frac{1-v}{2+v}} \leq \max \left\{ \kappa_f(v), H_0 (F(x_0) - F^*)^{\frac{1-v}{2+v}} \right\}. \tag{47}$$

This inequality is obviously valid for $k = 0$. Assume it is also valid for some $k \geq 0$. Then, by definition of H_{k+1} (see line 3 of Algorithm), we have $H_{k+1} = H_k 2^{i_k - 1}$. There are two cases. 1) $i_k = 0$. Then, $H_{k+1} < H_k$. By monotonicity of $\{F(x_k)\}_{k \geq 0}$ and by induction, we get:

$$\begin{aligned} H_{k+1} (F(x_{k+1}) - F^*)^{\frac{1-v}{2+v}} &< H_k (F(x_k) - F^*)^{\frac{1-v}{2+v}} \\ &\leq \max \left\{ \kappa_f(v), H_0 (F(x_0) - F^*)^{\frac{1-v}{2+v}} \right\}. \end{aligned}$$

2) $i_k > 0$. Then, applying Lemma 4.2 with $H := H_k 2^{i_k - 1} = H_{k+1}$ and $x := x_k$, we have:

$$H_{k+1} (F(x_{k+1}) - F^*)^{\frac{1-v}{2+v}} = H (F(T_{2H}(x)) - F^*)^{\frac{1-v}{2+v}} \stackrel{(39)}{\leq} \kappa_f(v).$$

Thus, (47) is true by induction. Choosing H_0 small enough (42), we have:

$$2H_k (F(x_k) - F^*)^{\frac{1-v}{2+v}} \leq 2\kappa_f(v), \quad 0 \leq k \leq K. \tag{48}$$

From Lemma 4.1 we know, that one of the two following estimates is true (denote $\delta_k := F(x_k) - F^*$):

- 1) $F(x_k) - F(x_{k+1}) \geq \alpha \cdot \delta_k \Leftrightarrow \delta_{k+1} \leq (1 - \alpha) \cdot \delta_k$, or
- 2) $F(x_k) - F(x_{k+1}) \geq \beta \cdot \delta_{k+1} \Leftrightarrow \delta_{k+1} \leq (1 + \beta)^{-1} \delta_k \leq (1 - \min\{\beta, 1\}/2) \cdot \delta_k$,

where $\alpha := \frac{1+\nu}{2+\nu} \cdot \min\left\{\left(\frac{(1+\nu)\gamma_f(\nu)}{2(2+\nu)}\right)^{\frac{1}{1+\nu}}, 1\right\}$, and

$$\beta := \left(\frac{2+\nu}{1+\nu}\right)^{\frac{3(1+\nu)}{2(2+\nu)}} \cdot \frac{(\alpha_f(2+\nu))^{\frac{3}{2(2+\nu)}}}{3(2\kappa_f(\nu))^{1/2}} \stackrel{(37)}{=} \frac{2+\nu}{1+\nu} \cdot \frac{2^{1/2} \cdot ((1+\nu)(2+\nu))^{\frac{1}{1+\nu}}}{6^{3/2} \cdot (8+\nu)^{(1-\nu)/(2+2\nu)}} \cdot \gamma_f(\nu)^{\frac{1}{1+\nu}}.$$

It remains to notice that $\alpha \geq \min\{\beta, 1\}/2$. Thus, we obtain (44).

Finally, let us estimate the total number of the oracle calls N_K during the first K iterations. At each iteration, the oracle is called $i_k + 1$ times, and we have $H_{k+1} = H_k 2^{i_k - 1}$. Therefore,

$$\begin{aligned} N_K &= \sum_{k=0}^{K-1} (i_k + 1) = \sum_{k=0}^{K-1} \left(\log_2 \frac{H_{k+1}}{H_k} + 2 \right) \\ &= 2K + \log_2 H_K - \log_2 H_0 \stackrel{(48),(43)}{\leq} 2K + \log_2 \frac{\kappa_f(\nu)}{\varepsilon^{(1-\nu)/(2+\nu)}} - \log_2 H_0. \end{aligned}$$

□

Note that condition (42) for the initial choice of H_0 can be seen as a definition of the moment, after which we can guarantee the linear rate of convergence (44). In practice, we can launch Algorithm 1 with *arbitrary* $H_0 > 0$. There are two possible options: either the method halves H_k at every step in the beginning, so H_k becomes small very quickly, or this value is increased at least once, and the required bound is guaranteed by Lemma 4.2. It can be easily proved, that this initial phase requires no more than $K_0 = \lceil \log_2 \frac{H_0 \varepsilon^{(1-\nu)/(1+\nu)}}{\kappa_f(\nu)} \rceil$ oracle calls.

5 Discussion

Let us discuss the global complexity results, provided by Theorem 4.1 for the Cubic Regularization of the Newton Method with the adaptive adjustment of the regularization parameter.

For the class of twice continuously differentiable strongly convex functions with Lipschitz continuous gradients $f \in \mathcal{S}_{\mu,L}^{2,1}(\text{dom } F)$, it is well known that the classical gradient descent method needs

$$O\left(\frac{L}{\mu} \log \frac{F(x_0) - F^*}{\varepsilon}\right) \tag{49}$$

iterations for computing ε -solution of the problem (e.g., [15]). As it was shown in [6], this result is shared by a variant of Cubic Regularization of the Newton method. This is much better than the bound $O\left(\left(\frac{L}{\mu}\right)^2 \log \frac{F(x_0) - F^*}{\varepsilon}\right)$, known for the damped Newton method (e.g., [2]).

For the class of uniformly convex functions of degree $p = 2 + \nu$ having Hölder continuous Hessian of degree $\nu \in [0, 1]$, we have proved the following parametric estimates: $O\left(\max\left\{\left(\gamma_f(\nu)\right)^{\frac{1}{1+\nu}}, 1\right\} \cdot \log \frac{F(x_0) - F^*}{\varepsilon}\right)$, where $\gamma_f(\nu) := \frac{\sigma_f(2+\nu)}{\mathcal{H}_f(\nu)}$ is the

condition number of degree ν . However, in practice we may not know exactly an appropriate value of the parameter ν . It is important that our algorithm automatically adjusts to the best possible complexity bound:

$$O\left(\max\left\{\inf_{\nu \in [0,1]}(\gamma_f(\nu))^{\frac{-1}{1+\nu}}, 1\right\} \cdot \log \frac{F(x_0) - F^*}{\varepsilon}\right). \tag{50}$$

Note that for $f \in S_{\mu,L}^{2,1}(\text{dom } F)$ we have:

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L - \mu, \quad x, y \in \text{dom } F.$$

Thus, $\mathcal{H}_f(0) \leq L - \mu$ and $\gamma_f(0) \geq \frac{\mu}{L - \mu}$. So we can conclude that the estimate (50) is better than (49). Moreover, addition to our objective arbitrary convex quadratic function does not change any of $\mathcal{H}_f(\nu)$, $\nu \in [0, 1]$. Thus, it can only improve the condition number $\gamma_f(\nu)$, while the ratio L/μ may become arbitrarily bad. It confirms an intuition that a natural Newton-type minimization scheme should not be affected by any quadratic parts of the objective, and the notion of *well-conditioned* and *ill-conditioned* problems for second-order methods should be different from that of for first-order ones.

Note that in the recent paper [11], a linear rate of convergence was also proven for the accelerated second-order scheme, with the complexity bound:

$$O\left(\max\{(\gamma_f(\nu))^{\frac{-1}{2+\nu}}, 1\} \cdot \log \frac{\mathcal{H}_f(\nu)D_0^{2+\nu}}{\varepsilon}\right). \tag{51}$$

This is the better rate than (50). However, the method requires to know the parameter ν , and the constant of uniform convexity. Thus, one theoretical question remains open: is it possible to construct *universal* second-order scheme, matching (51) in the uniformly convex case.

Looking at the definitions of $\mathcal{H}_f(\nu)$ and $\sigma_f(2 + \nu)$, we can see that, for all $x, y \in \text{dom } F$, $x \neq y$,

$$\sigma_f(2 + \nu) \leq \frac{\langle \nabla f(x) - \nabla f(y), x - y \rangle}{\|x - y\|^{2+\nu}}, \quad \frac{1}{\mathcal{H}_f(\nu)} \leq \frac{\|x - y\|^\nu}{\|\nabla^2 f(x) - \nabla^2 f(y)\|},$$

and

$$\gamma_f(\nu) := \frac{\sigma_f(2+\nu)}{\mathcal{H}_f(\nu)} \leq \frac{\langle \nabla f(x) - \nabla f(y), x - y \rangle}{\|\nabla^2 f(x) - \nabla^2 f(y)\| \cdot \|x - y\|^2}.$$

The last fraction does not depend on any particular ν . So, for any twice-differentiable convex function, we can define the following number:

$$\gamma_f := \inf_{\substack{x, y \in \text{dom } F \\ x \neq y}} \frac{\langle \nabla f(x) - \nabla f(y), x - y \rangle}{\|\nabla^2 f(x) - \nabla^2 f(y)\| \cdot \|x - y\|^2}.$$

If it is positive, then it could serve as an indicator of the *second-order non-degeneracy*, for which we have a lower bound: $\gamma_f \geq \gamma_f(\nu)$, $\nu \in [0, 1]$.

6 Conclusions

In this work, we have introduced the second-order condition number of a certain degree, which plays as the main complexity factor for solving uniformly convex minimization problems with Hölder-continuous Hessian of the objective by second-order optimization schemes.

We have proved that cubically regularized Newton method with an adaptive estimation of the regularization parameter achieves global linear rate of convergence on this class of functions. The algorithm does not require to know any parameters of the problem class and automatically fits to the best possible degree of nondegeneracy.

Using this technique, we have justified that global iteration complexity of cubic Newton is always better than corresponding one of gradient method for the standard class of strongly convex functions with uniformly bounded second derivative.

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Acknowledgements The research results of this paper were obtained with support of ERC Advanced Grant 788368.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Agarwal, N., Allen-Zhu, Z., Bullins, B., Hazan, E., Ma, T.: Finding approximate local minima faster than gradient descent. In: Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, pp. 1195–1199. ACM (2017)
2. Boyd, S., Vandenberghe, L.: *Convex Optimization*. Cambridge University Press, Cambridge (2004)
3. Carmon, Y., Duchi, J.C.: Gradient descent efficiently finds the cubic-regularized non-convex Newton step. [arXiv:1612.00547](https://arxiv.org/abs/1612.00547) (2016)
4. Cartis, C., Gould, N.I., Toint, P.L.: Adaptive cubic regularisation methods for unconstrained optimization. Part II: worst-case function-and derivative-evaluation complexity. *Math. Program.* **130**(2), 295–319 (2011)
5. Cartis, C., Gould, N.I., Toint, P.L.: Adaptive cubic regularisation methods for unconstrained optimization. Part I: motivation, convergence and numerical results. *Math. Program.* **127**(2), 245–295 (2011)
6. Cartis, C., Gould, N.I., Toint, P.L.: Evaluation complexity of adaptive cubic regularization methods for convex unconstrained optimization. *Optim. Methods Softw.* **27**(2), 197–219 (2012)
7. Cartis, C., Scheinberg, K.: Global convergence rate analysis of unconstrained optimization methods based on probabilistic models. *Math. Program.* **169**(2), 337–375 (2018)
8. Doikov, N., Richtárik, P.: Randomized block cubic Newton method. In: International Conference on Machine Learning, pp. 1289–1297 (2018)
9. Ghadimi, S., Liu, H., Zhang, T.: Second-order methods with cubic regularization under inexact information. [arXiv:1710.05782](https://arxiv.org/abs/1710.05782) (2017)
10. Grapiglia, G.N., Nesterov, Y.: Regularized Newton methods for minimizing functions with Hölder continuous Hessians. *SIAM J. Optim.* **27**(1), 478–506 (2017)

11. Grapiglia, G.N., Nesterov, Y.: Accelerated regularized Newton methods for minimizing composite convex functions. *SIAM J. Optim.* **29**(1), 77–99 (2019)
12. Kohler, J.M., Lucchi, A.: Sub-sampled cubic regularization for non-convex optimization. In: *International Conference on Machine Learning*, pp. 1895–1904 (2017)
13. Nesterov, Y.: Modified Gauss–Newton scheme with worst case guarantees for global performance. *Optim. Methods Softw.* **22**(3), 469–483 (2007)
14. Nesterov, Y.: Accelerating the cubic regularization of Newton’s method on convex problems. *Math. Program.* **112**(1), 159–181 (2008)
15. Nesterov, Y.: *Lectures on Convex Optimization*, vol. 137. Springer, Berlin (2018)
16. Nesterov, Y.: Implementable tensor methods in unconstrained convex optimization. In: *Mathematical Programming* pp. 1–27 (2019)
17. Nesterov, Y., Polyak, B.T.: Cubic regularization of Newton’s method and its global performance. *Math. Program.* **108**(1), 177–205 (2006)
18. Tripuraneni, N., Stern, M., Jin, C., Regier, J., Jordan, M.I.: Stochastic cubic regularization for fast nonconvex optimization. In: *Advances in Neural Information Processing Systems*, pp. 2899–2908 (2018)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.