GRADIENT REGULARIZATION OF NEWTON METHOD WITH BREGMAN DISTANCES

Nikita Doikov, Yurii Nesterov

REPRINT | 3253

UCLouvain

CORE | LIDAM
Louvain Institute of Data Analysis and Modeling in economics and statistics

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

Mathematical Programming
https://doi.org/10.1007/s10107-023-01943-7

FULL LENGTH PAPER

Series A

Gradient regularization of Newton method with Bregman distances

Nikita Doikov[1] · **Yurii Nesterov**[2]

Received: 10 March 2022 / Accepted: 22 February 2023

Abstract

In this paper, we propose a first second-order scheme based on arbitrary non-Euclidean norms, incorporated by Bregman distances. They are introduced directly in the Newton iterate with regularization parameter proportional to the square root of the norm of the current gradient. For the basic scheme, as applied to the composite convex optimization problem, we establish the global convergence rate of the order $O(k^{-2})$ both in terms of the functional residual and in the norm of subgradients. Our main assumption on the smooth part of the objective is Lipschitz continuity of its Hessian. For uniformly convex functions of degree three, we justify global linear rate, and for strongly convex function we prove the local superlinear rate of convergence. Our approach can be seen as a relaxation of the Cubic Regularization of the Newton method (Nesterov and Polyak in Math Program 108(1):177–205, 2006) for convex minimization problems. This relaxation preserves the convergence properties and global complexities of the Cubic Newton in convex case, while the auxiliary subproblem at each iteration is simpler. We equip our method with adaptive search procedure for choosing the regularization parameter. We propose also an accelerated scheme with convergence rate $O(k^{-3})$, where k is the iteration counter.

Keywords Newton method · Regularization · Convex optimization · Global complexity bounds · Large-scale optimization

This paper has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 788368). It was also supported by Multidisciplinary Institute in Artificial intelligence MIAI@Grenoble Alpes (ANR-19-P3IA-0003).

✉ Nikita Doikov
Nikita.Doikov@uclouvain.be

Yurii Nesterov
Yurii.Nesterov@uclouvain.be

[1] Institute of Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM), Catholic University of Louvain (UCLouvain), Louvain-la-Neuve, Belgium

[2] Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCLouvain), Louvain-la-Neuve, Belgium

Published online: 24 March 2023

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

1 Introduction

The classical Newton's method is a powerful tool for solving various optimization problems and for dealing with ill-conditioning. The practical implementation of this method for solving unconstrained minimization problem $\min_x f(x)$ can be written as follows:

$$x_{k+1} = x_k - \alpha_k \big(\nabla^2 f(x_k)\big)^{-1} \nabla f(x_k), \qquad k \geq 0,$$

where $0 < \alpha_k \leq 1$ is a damping parameter. However, this approach has two serious drawbacks. Firstly, the next point is not well-defined when the Hessian is a degenerate matrix. And secondly, while the method has a very fast local quadratic convergence, it is difficult to establish any *global* properties for this process. Indeed, for $\alpha_k = 1$ (the classical pure Newton method), there are known examples of problems for which the method does not converge globally [5]. The pure Newton step might not work even if the objective is strongly convex (see, e.g., Example 1.4.3 in [6]). For the damped Newton method with line search, it is possible to prove some global convergence rates. But, typically, they are worse than the rates of the classical Gradient Method [18].

A breakthrough in the second-order optimization theory was made after [19], where the Cubic Regularization of the Newton method was presented together with its global convergence properties. The main standard assumption is that the Hessian of the objective is Lipschitz continuous with some parameter $L_2 \geq 0$:

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L_2 \|x - y\|, \qquad \forall x, y,$$

ensuring the *global upper approximation* of our function formed by the second-order Taylor polynomial augmented by the third power of the norm. The next point is then defined as the minimum of the upper model:

$$x_{k+1} = \operatorname*{argmin}_y \Big[\langle \nabla f(x_k), y - x_k \rangle + \frac{1}{2} \langle \nabla^2 f(x_k)(y - x_k), y - x_k \rangle + \frac{L_2}{6} \|y - x_k\|^3 \Big]. \tag{1}$$

Initially, this idea had a full theoretical justification only for the Euclidean norm $\|\cdot\|$. In this case, the solution to the auxiliary minimization problem (1) does not have a closed form expression, but it can be found by solving a one-dimensional nonlinear equation and by using the standard factorization tools of Linear Algebra. The use of general and even variable norms with cubic regularization in second-order methods was considered recently in [11, 15], which can be useful for solving optimization problems with non-Euclidean geometry.

However, even in the Euclidean case, the presence of the cubic term in the objective makes it more difficult to use the classical gradient-type methods with their developed

complexity theory. While it is possible to apply the gradient descent [2], the cubic subproblem prevents the usage of the standard accelerated and conjugate gradients methods. This drawback restricts the application of method (1) to large-scale problems.

In this paper, we show how to avoid these restrictions. Namely, we will show that it is possible to use a *quadratic regularization* of the Taylor polynomial with a properly chosen coefficient that depends only on the current iterate. In the simplest form, one iteration of our method is as follows:

$$x_{k+1} = x_k - \left(\nabla^2 f(x_k) + A_k I\right)^{-1} \nabla f(x_k), \tag{2}$$

where

$$A_k = \sqrt{\frac{L_2}{3} \|\nabla f(x_k)\|}. \tag{3}$$

We see that it is very easy for implementation, since it requires only *one* matrix inversion, the very standard operation of Linear Algebra. At the same time, this subproblem is now suitable for the classical Congugate Gradient method as well.[1]

For the class of Trust Region methods as applied to unconstrained minimization problems, the trust region radius proportional to the gradient norm was proposed in [9]. The use of the gradient norm as a regularizer for the Newton method was considered in the work [20]. Then, the method has a local quadratic convergence. However, to ensure some global rate for such regularization, one need to use damping steps, which makes the rate slower.

It appears that for the optimization process (2), (3), we can establish the global convergence guarantees of the same type as for the Cubic Newton method (1). Namely, we prove the global rate of the order $O(1/k^2)$ in terms of the functional residual and in terms of the subgradient norm for the general convex functions. This is much faster than the standard $O(1/k)$-rate of the Gradient Method. Moreover, for the uniformly convex functions of degree three, we prove the global linear rate. For the strongly convex functions we establish a local superlinear convergence.

In this paper, we consider convex optimization problems in a general composite form. Recently, globally convergent Newton methods for nonsmooth optimization were proposed in [13]. They are based on the damping steps and regularization by the gradient norm, which is different from the rule (3).

We also work with arbitrary (possibly non-Euclidean) norms by employing the technique of Bregman distances. An alternative approach of using general norms in the cubically regularized Newton scheme was proposed in [11], that uses the adaptive regularization framework of [3].

Contents. The rest of the paper is organized as follows. In Sect. 2, we present the main properties of one iteration of the scheme.

[1] When this paper was already finished, we discovered that this idea was recently proposed by K. Mishchenko [16] for solving unconstrained minimization problem with smooth objective. As compared to his work, our main advances consist in the usage of Bregman distances, composite form of optimization problem, linear rate of convergence for uniformly convex functions, and developments of accelerated variant of the method.

We study the convergence rate of the basic process in Sect. 3. In Sect. 4, we establish convergence for the norm of the gradient. An adaptive search procedure for our method is discussed in Sect. 5.

In Sect. 6, we consider an accelerated scheme based on the iterations of the basic method and justify its global complexity of the order $\tilde{O}(\epsilon^{-1/3})$ assuming Lipschitz continuity of the Hessian of the smooth part of the objective function. Section 7 contains numerical experiments. Some concluding remarks are in Sect. 8.

Notation. Let us fix a finite-dimensional real vector space $\mathbb{E}$. Our goal is to solve the following *Composite Minimization Problem*

$$F^* \;=\; \min_{x \in \mathrm{dom}\, \psi} \big[F(x) \;\stackrel{\mathrm{def}}{=}\; f(x) + \psi(x) \big], \tag{4}$$

where $\psi(\cdot)$ is a *simple* closed convex function with $\mathrm{dom}\, \psi \subseteq \mathbb{E}$, and $f(\cdot)$ is a convex and two times continuously differentiable function.

We measure distances in $\mathbb{E}$ by a general norm $\|\cdot\|$. Its dual space is denoted by $\mathbb{E}^*$. It is a space of all linear functions on $\mathbb{E}$, for which we define the norm in the standard way:

$$\|g\|_* = \max_{x \in \mathbb{E}} \{\, \langle g, x \rangle :\; \|x\| \le 1 \,\}, \qquad g \in \mathbb{E}^*.$$

Using this norm, we can define an induced norm for a self-adjoint linear operator $B : \mathbb{E} \to \mathbb{E}^*$ as follows:

$$\|B\| = \max_{x \in \mathbb{E}} \{ |\langle Bx, x \rangle| :\; \|x\| \le 1 \}.$$

We can also define the bounds of its spectrum as the best values $\lambda_{\min}(B)$ and $\lambda_{\max}(B)$ satisfying conditions

$$\lambda_{\min}(B) \|x\|^2 \le \langle Bx, x \rangle \;\le\; \lambda_{\max}(B) \|x\|^2, \quad \forall x \in \mathbb{E}.$$

Our optimization schemes will be based on some scaling function $d(\cdot)$, which we assume to be a strongly convex function with Lipschitz-continuous gradients:

$$d(y) \ge d(x) + \langle \nabla d(x), y - x \rangle + \frac{\sigma}{2} \|y - x\|^2, \tag{5}$$

$$\|\nabla d(x) - \nabla d(y)\|_* \le \|x - y\|, \tag{6}$$

where $\sigma \in (0, 1]$ and the points $x, y \in \mathrm{dom}\, \psi$ are arbitrary. For twice-differentiable scaling functions, this condition can be characterized by the following bounds on the Hessian:

$$\sigma \|h\|^2 \le \langle \nabla^2 d(x) h, h \rangle \;\le\; \|h\|^2, \qquad \forall x \in \mathrm{dom}\, \psi,\; h \in \mathbb{E}.$$

Using this function, we define the following *Bregman distance*:

$$\rho(x, y) = \beta_d(x, y) \stackrel{\text{def}}{=} d(y) - d(x) - \langle \nabla d(x), y - x \rangle, \quad x, y \in \operatorname{dom} \psi. \tag{7}$$

We will employ this object to regularize the second-order model of the objective.

The standard condition for the smooth part of the objective function in problem (4) is Lipschitz continuity of the Hessians:

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L_2 \|x - y\|, \qquad \forall x, y \in \operatorname{dom} \psi, \tag{8}$$

that we always assume to be satisfied. This inequality has the following consequences, which are valid for all $x, y \in \operatorname{dom} \psi$:

$$\|\nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y - x)\|_* \leq \frac{1}{2} L_2 \|y - x\|^2, \tag{9}$$

and

$$\begin{aligned} &|f(y) - f(x) - \langle \nabla f(x), y - x \rangle + \frac{1}{2} \langle \nabla^2 f(x)(y - x), y - x \rangle| \\ &\quad \leq \frac{1}{6} L_2 \|y - x\|^3. \end{aligned} \tag{10}$$

2 Gradient regularization

Our main iteration at some point $\bar{x} \in \operatorname{dom} \psi$ with a step-size $A > 0$ is defined as follows:

$$\begin{aligned} T_A(\bar{x}) \stackrel{\text{def}}{=} \operatorname*{argmin}_{y \in \operatorname{dom} \psi} \Big[M_A(\bar{x}, y) &\stackrel{\text{def}}{=} f(\bar{x}) + \langle \nabla f(\bar{x}), y - \bar{x} \rangle \\ &+ \frac{1}{2} \langle \nabla^2 f(\bar{x})(y - \bar{x}), y - \bar{x} \rangle + A \rho(\bar{x}, y) + \psi(y) \Big]. \end{aligned} \tag{11}$$

This is minimization of a convex quadratic function augmented by Bregman distance and the composite part. Our main structural assumption is that both $\rho(\bar{x}, \cdot)$ and $\psi(\cdot)$ are *simple*, meaning that problem (11) is efficiently solvable.

The use of the general scaling function $d(\cdot)$ can be beneficial in practice for solving problems with some specific non-Euclidean geometry.

Example 1 Let $\psi(x) \equiv 0$ and the scaling function is $d(x) := \frac{1}{2} \langle Bx, x \rangle$ for a fixed positive definite self-adjoint operator $B = B^* \succ 0$. Then,

$$\rho(\bar{x}, y) = \frac{1}{2} \langle B(y - \bar{x}), y - \bar{x} \rangle,$$

and one iteration (11) can be written in an explicit form, as follows:

$$T_A(\bar{x}) = \bar{x} - \left(\nabla^2 f(\bar{x}) + AB \right)^{-1} \nabla f(\bar{x}).$$

Example 2 Consider the unconstrained minimization problem $\min_{x \in \mathbb{R}^n} f(x)$, with

$$f(x) = g(Cx), \qquad C \in \mathbb{R}^{m \times n},$$

where $g : \mathbb{R}^m \to \mathbb{R}$ is a convex smooth function. Let us fix the standard Euclidean norm $\|\cdot\|_2$ in $\mathbb{R}^m$ and assume that the Hessian of g is Lipschitz continuous w.r.t. this norm with constant L_g. Then, if we use the standard Euclidean norm $\|\cdot\|_2$ for our primal space $\mathbb{R}^n$, the corresponding Lipschitz constant of $\nabla^2 f(\cdot)$ is

$$L_f = \|C\|^3 L_g.$$

At the same time, using the following scaled norm $\|x\| := \langle Bx, x\rangle^{1/2}, x \in \mathbb{R}^n$ with matrix $B = C^T C$ (assuming $B \succ 0$, so the rows of C have a full rank) and the scaling function from the previous example, we have

$$L_f = L_g,$$

which is much better.

Example 3 Let $\psi(\cdot)$ be $\{0, +\infty\}$-indicator of the standard simplex

$$\Delta_n \overset{\text{def}}{=} \left\{ x \in \mathbb{R}^n_+ \; : \; \sum_{i=1}^n x^{(i)} = 1 \right\}.$$

Thus, problem (4) is to minimize a smooth convex function over this set:

$$\min_{x \in \Delta_n} f(x).$$

One of the most suitable choices of the norm for this problem is ℓ_1-norm [1], defined as $\|x\|_1 \overset{\text{def}}{=} \sum_{i=1}^n |x^{(i)}|$ for $x \in \mathbb{R}^n$. The Lipschitz constant w.r.t. this norm is smaller than that one measured in ℓ_2-norm. Let us fix some $\delta > 0$, and use the following scaling function,

$$d(x) := \delta \sum_{i=1}^n (x^{(i)} + \delta) \ln(x^{(i)} + \delta).$$

We have, for any $h \in \mathbb{R}^n$ and $x \in \Delta_n$:

$$\langle \nabla^2 d(x) h, h \rangle = \delta \sum_{i=1}^n \frac{(h^{(i)})^2}{x^{(i)} + \delta} \quad \leq \quad \|h\|_2^2 \quad \leq \quad \|h\|_1^2.$$

And, by Cauchy-Schwarz inequality, it holds

$$\|h\|_1 = \sum_{i=1}^{n} \frac{|h^{(i)}|\sqrt{x^{(i)}+\delta}}{\sqrt{x^{(i)}+\delta}} \;\le\; \left(\sum_{i=1}^{n} \frac{(h^{(i)})^2}{x^{(i)}+\delta}\right)^{1/2} \left(1+n\delta\right)^{1/2}.$$

Hence,

$$\langle \nabla^2 d(x)h, h\rangle \ge \frac{\delta}{1+n\delta}\|h\|_1^2.$$

and conditions (5), (6) are satisfied with $\sigma = \frac{\delta}{1+n\delta} = \frac{1}{1/\delta+n}$.

In general, the solution to this problem $T = T_A(\bar{x})$ is characterized by the following variational principle (see, e.g. [18]):

$$\begin{aligned} &\langle \nabla f(\bar{x}) + \nabla^2 f(\bar{x})(T-\bar{x}) + A(\nabla d(T) - \nabla d(\bar{x})), y - T\rangle \\ &\quad + \psi(y) \;\ge\; \psi(T), \quad \forall y \in \operatorname{dom}\psi. \end{aligned} \tag{12}$$

Thus, defining $\psi'(T) = -\nabla f(\bar{x}) - \nabla^2 f(\bar{x})(T-\bar{x}) - A(\nabla d(T) - \nabla d(\bar{x}))$, we see that $\psi'(T) \in \partial\psi(T)$. Consequently,

$$\begin{aligned} F'(T) &= \nabla f(T) + \psi'(T) \\ &= \nabla f(T) - \nabla f(\bar{x}) - \nabla^2 f(\bar{x})(T-\bar{x}) \\ &\quad -A(\nabla d(T) - \nabla d(\bar{x})) \;\in\; \partial F(T). \end{aligned} \tag{13}$$

Note that this is a very special way of selecting subgradient of a possibly nonsmooth function $F(\cdot)$, which allows $\|F'(T)\|_*$ approach zero.

Denote $M_A(\bar{x}) = M_A(\bar{x}, T_A(\bar{x})) \le M_A(\bar{x}, \bar{x}) = F(\bar{x})$. Let us prove the following important fact, that uses convexity of the original problem (4).

Lemma 1 *For all $y \in \operatorname{dom}\psi$ and $T = T_A(\bar{x})$, we have*

$$M_A(\bar{x}, y) \ge M_A(\bar{x}) + \frac{1}{2}\langle \nabla^2 f(\bar{x})(y-T), y-T\rangle + \frac{1}{2}\sigma A\|y-T\|^2. \tag{14}$$

Moreover,

$$\|T_A(\bar{x}) - \bar{x}\| \le \frac{1}{\sigma A}\|F'(\bar{x})\|_*, \tag{15}$$

where $F'(\bar{x}) = \nabla f(\bar{x}) + \psi'(\bar{x})$ and $\psi'(\bar{x})$ is an arbitrary element of $\partial\psi(\bar{x})$.

Proof For optimization problem in (11), define the scaling function

$$\xi(x) = \frac{1}{2}\langle \nabla^2 f(\bar{x})x, x\rangle + Ad(x).$$

Note that the objective function in this problem is strongly convex relatively to $\xi(\cdot)$ with constant one. Therefore, for any $y \in \operatorname{dom} \psi$,

$$M_A(\bar{x}, y) - M_A(\bar{x}) \geq \beta_\xi(T, y) \;=\; \frac{1}{2}\langle \nabla^2 f(\bar{x})(y - T), y - T\rangle + A\beta_d(T, y)$$
$$\overset{(5)}{\geq} \frac{1}{2}\langle \nabla^2 f(\bar{x})(y - T), y - T\rangle + \frac{1}{2}\sigma A\|y - T\|^2.$$

In order to prove (15), note that

$$\begin{aligned} M_A(\bar{x}) &\geq F(\bar{x}) + \min_{y \in \operatorname{dom}\psi}\left[\langle F'(\bar{x}), y - \bar{x}\rangle + \frac{1}{2}\sigma A\|y - \bar{x}\|^2\right] \\ &\geq F(\bar{x}) + \min_{y \in \mathbb{E}}\left[\langle F'(\bar{x}), y - \bar{x}\rangle + \frac{1}{2}\sigma A\|y - \bar{x}\|^2\right] \\ &= F(\bar{x}) - \frac{1}{2\sigma A}\|F'(\bar{x})\|_*^2. \end{aligned}$$

Since $M_A(\bar{x}, \bar{x}) = F(\bar{x})$, we get (15) from (14) with $y = \bar{x}$. □

In what follows, the parameter A in the optimization problem (11) is chosen as

$$A \;=\; A_H(\bar{x}) \;=\; \frac{1}{\sigma}\sqrt{\frac{H}{3}\|F'(\bar{x})\|_*}, \tag{16}$$

where $H > 0$ is an estimate of the Lipschitz constant L_2 in (8). This choice is explained by the following result.

Corollary 1 *For $A = A_H(\bar{x})$, we have*

$$H\|T_A(\bar{x}) - \bar{x}\| \leq 3\sigma A. \tag{17}$$

Proof Indeed, this is a simple consequence of inequality (15) and definition (11). □

Let us relate the optimal value of the auxiliary problem (11) with the cubic over-approximation (10).

Lemma 2 *Let $A = A_H(\bar{x})$ and $T = T_A(\bar{x})$. Assume that for some $H > 0$ the following condition is satisfied:*

$$\begin{aligned} f(T) \leq\; & f(\bar{x}) + \langle \nabla f(\bar{x}), T - \bar{x}\rangle \\ & + \frac{1}{2}\langle \nabla^2 f(\bar{x})(T - \bar{x}), T - \bar{x}\rangle + \frac{H\|T - \bar{x}\|^3}{6} \end{aligned} \tag{18}$$

(clearly, it holds for $H \geq L_2$, where L_2 is the Lipschitz constant of the Hessian). Then

$$F(\bar{x}) - F(T) \geq \frac{1}{2}\langle \nabla^2 f(\bar{x})(T - \bar{x}), T - \bar{x}\rangle + \frac{1}{2}\sigma A\|T - \bar{x}\|^2. \tag{19}$$

Proof Indeed,

$$\begin{aligned} f(T) &\overset{(18)}{\leq} M_A(\bar{x}) - A\rho(\bar{x}, T) - \psi(T) + \frac{H}{6}\|T - \bar{x}\|^3 \\ &\overset{(5)}{\leq} M_A(\bar{x}) - \psi(T) + \frac{H}{6}\|T - \bar{x}\|^3 - \frac{1}{2}\sigma A\|T - \bar{x}\|^2 \\ &\overset{(17)}{\leq} M_A(\bar{x}) - \psi(T). \end{aligned}$$

Thus, $F(T) \leq M_A(\bar{x})$ and (19) follows from (14) with $y = \bar{x}$. □

Finally, we need to estimate the norm of subgradient at the new point.

Lemma 3 *Let the Hessian be Lipschitz continuous with constant L_2. Fix arbitrary $H > 0$. Let $A = A_H(\bar{x})$ and $T = T_A(\bar{x})$. Then*

$$\|F'(T)\|_* \leq \sigma A\left(\sigma^{-1} + \frac{3L_2}{2H}\right)\|T - \bar{x}\| \;\leq\; c\|F'(\bar{x})\|_*, \tag{20}$$

where

$$c \overset{\text{def}}{=} \sigma^{-1} + \frac{3L_2}{2H}.$$

Proof Indeed,

$$\begin{aligned} \|F'(T)\|_* &\overset{(13)}{=} \|\nabla f(T) - \nabla f(\bar{x}) - \nabla^2 f(\bar{x})(T - \bar{x}) - A(\nabla d(T) - \nabla d(\bar{x}))\|_* \\ &\overset{(9)}{\leq} \frac{1}{2}L_2\|T - \bar{x}\|^2 + A\|\nabla d(T) - \nabla d(\bar{x})\|_* \overset{(6)}{\leq} \frac{1}{2}L_2\|T - \bar{x}\|^2 + A\|T - \bar{x}\|_* \\ &\overset{(17)}{\leq} A\left(1 + \frac{3\sigma L_2}{2H}\right)\|T - \bar{x}\|. \end{aligned}$$

This is the first inequality in (20). For the second one, we can continue as follows:

$$\|F'(T)\|_* \overset{(17)}{\leq} \left(1 + \frac{3\sigma L_2}{2H}\right) \cdot \frac{3\sigma A^2}{H} \overset{(17)}{=} c\|F'(\bar{x})\|_*.$$

□

Now we can prove the main theorem of this section.

Theorem 1 *Let the Hessian be Lipschitz continuous with constant L_2. Fix arbitrary $H > 0$. Let $A = A_H(\bar{x})$ and $T = T_A(\bar{x})$. If for this point relation (18) is valid, then*

$$F(\bar{x}) - F(T) \geq \frac{1}{2c^2}\sqrt{\frac{3}{H}} \cdot \frac{\|F'(T)\|_*^2}{\|F'(\bar{x})\|_*^{1/2}}. \tag{21}$$

Proof We only need to insert in (19) the first inequality of (20) and definition (16). □

3 Properties of the minimization process

In this section, we propose an iterative scheme based on the gradient regularization of the Newton steps. Note that the choice of the regularization parameter (16) depends solely on the *current gradient norm* and it can be easily computed at each iteration. Then, we do one regularized Newton step defined by (11). According to Theorem 1, repeating this process would result in monotone decrease of the objective.

First, we prove global convergence for the function value. In the next section, we also prove the convergence in terms of the gradient norm. Thus, small gradient norm can serve as a stopping criteria for our scheme.

Let us analyze the following algorithm with a fixed value of parameter H.

$$\boxed{\begin{array}{l} \textbf{Initialization.} \text{ Choose } H \geq L_2, x_0 \in \operatorname{dom}\psi, \text{ and } F_0' \in \partial F(x_0). \\ \\ k\textbf{th iteration } (k \geq 0)\textbf{.} \ 1). \text{ Set } g_k = \|F_k'\|_* \text{ and } A_k = \frac{1}{\sigma}\sqrt{\frac{H}{3}g_k}. \\ \\ 2). \text{ Compute } x_{k+1} = T_{A_k}(x_k) \text{ and define} \\ \\ \quad F_{k+1}' \ = \ \nabla f(x_{k+1}) - \nabla f(x_k) - \nabla^2 f(x_k)(x_{k+1} - x_k) \\ \\ \qquad\quad - \ A_k(\nabla d(x_{k+1}) - \nabla d(x_k)). \end{array}} \tag{22}$$

Let us introduce the distance to the initial level set:

$$D = \sup_{x \in \operatorname{dom}\psi} \{\|x - x^*\| : \ F(x) \leq F(x_0)\},$$

which we assume to be bounded: $D < +\infty$. We can prove the following convergence rate for method (22).

Theorem 2 *Let the Hessian be Lipschitz continuous with constant L_2. Let $H \geq L_2$ and $F(x_k) - F^* \geq \epsilon$ for some $k \geq 0$. Then,*

$$\frac{1}{[F(x_k) - F^*]^{1/2}} \geq \frac{1}{[F(x_0) - F^*]^{1/2}} + \frac{1}{4c^2}\sqrt{\frac{3}{HD^3}}\left(k - \ln\frac{(F(x_0) - F^*)\|F'(x_0)\|_*^{1/2} D^{1/2}}{\epsilon^{3/2}}\right). \tag{23}$$

Proof Denote $F_k = F(x_k) - F(x^*)$ and $g_k = \|F'(x_k)\|_*$. Thus, $F_k \leq D g_k$. Note that

$$\frac{1}{F_{k+1}^{1/2}} - \frac{1}{F_k^{1/2}} = \frac{F_k^{1/2} - F_{k+1}^{1/2}}{F_k^{1/2} F_{k+1}^{1/2}} \ = \ \frac{F_k - F_{k+1}}{F_k^{1/2} F_{k+1}^{1/2}(F_k^{1/2} + F_{k+1}^{1/2})} \ \geq \ \frac{F_k - F_{k+1}}{2F_k F_{k+1}^{1/2}}.$$

Since for all $k \geq 1$, the subgradients of $\psi(\cdot)$ are defined by the rule (13), we can use the results of Sect. 2. We can continue as follows:

$$\frac{1}{F_{k+1}^{1/2}} - \frac{1}{F_k^{1/2}} \overset{(21)}{\geq} \frac{\sqrt{3} g_{k+1}^2}{4\sqrt{H} c^2 g_k^{1/2} F_k F_{k+1}^{1/2}} \geq \frac{\sqrt{3} g_{k+1}^{1/2} F_{k+1}}{4\sqrt{H} c^2 g_k^{1/2} F_k D^{3/2}} = \frac{g_{k+1}^{1/2} F_{k+1}}{4c^2 g_k^{1/2} F_k} \sqrt{\frac{3}{H D^3}}.$$

Summing up these bounds and using the inequality of arithmetic and geometric means, we get

$$\begin{aligned} \frac{1}{F_k^{1/2}} - \frac{1}{F_0^{1/2}} &\geq \frac{1}{4c^2} \sqrt{\frac{3}{H D^3}} \sum_{i=0}^{k-1} \frac{F_{i+1} g_{i+1}^{1/2}}{F_i g_i^{1/2}} \geq \frac{k}{4c^2} \sqrt{\frac{3}{H D^3}} \left(\frac{F_k g_k^{1/2}}{F_0 g_0^{1/2}} \right)^{1/k} \\ &\geq \frac{k}{4c^2} \sqrt{\frac{3}{H D^3}} \left(\frac{\epsilon^{3/2}}{F_0 g_0^{1/2} D^{1/2}} \right)^{1/k}. \end{aligned} \tag{24}$$

Since

$$\left(\frac{\epsilon^{3/2}}{F_0 g_0^{1/2} D^{1/2}} \right)^{1/k} = \exp\Big(-\frac{1}{k} \ln \frac{F_0 g_0^{1/2} D^{1/2}}{\epsilon^{3/2}} \Big) \geq 1 - \frac{1}{k} \ln \frac{F_0 g_0^{1/2} D^{1/2}}{\epsilon^{3/2}},$$

we obtain inequality (23). □

Corollary 2 *The second condition of Theorem 2 can be valid only for*

$$k \leq 4c^2 \sqrt{\frac{H D^3}{3\epsilon}} + \ln \frac{(F(x_0) - F^*) \|F'(x_0)\|_*^{1/2} D^{1/2}}{\epsilon^{3/2}}. \tag{25}$$

Remark 1 Note that up to the additive logarithmic term, the iteration complexity (25) corresponds to that one of the Cubically regularized Newton method as applied to convex functions [19] in the Euclidean case. However, iterations of our method (22) are easier to implement, and it is also possible to use an arbitrary scaling function $d(\cdot)$.

Remark 2 The right-hand side of inequality (25) can be used for defining the optimal value of parameter H. Indeed, it can be chosen as a minimizer of the following function:

$$2 \ln(2H\sigma^{-1} + 3L_2) - \frac{3}{2} \ln H.$$

This gives us

$$H_* = \frac{9}{2} L_2 \sigma. \tag{26}$$

In this case,

$$4c^2\sqrt{\frac{H_* D^3}{3\epsilon}} = \frac{64}{9\sigma}\sqrt{\frac{3L_2 D^3}{2\epsilon\sigma}} \;<\; 8.71\sqrt{\frac{L_2 D^3}{\epsilon\sigma^3}}. \tag{27}$$

Let us estimate now the performance of method (22) on uniformly convex functions. Consider the case when function $F(\cdot)$ is uniformly convex of degree three:

$$F(y) \geq F(x) + \langle F'(x), y - x\rangle + \frac{\sigma_3}{3}\|y - x\|^3, \quad x, y \in \operatorname{dom}\psi. \tag{28}$$

For the composite $F(\cdot)$, this property can be ensured either by its smooth component $f(\cdot)$, or by the general component $\psi(\cdot)$. In the latter case, it is not necessary to coordinate this assumption with the smoothness condition (8).

In our analysis, we need the following straightforward consequence of definition (28):

$$F(x) - F^* \leq \frac{2}{3\sqrt{\sigma_3}}\|F'(x)\|_*^{3/2}, \quad x \in \operatorname{dom}\psi. \tag{29}$$

Theorem 3 *Let the Hessian be Lipschitz continuous with constant L_2. Let $F(\cdot)$ satisfies condition (28). If $H \geq L_2$, then for all $k \geq 0$ we have*

$$F(x_k) - F^* \leq D\|F'(x_0)\|_* \cdot \exp\left(-\frac{k\ln(1+S)}{c^{1/2} + \frac{1}{2}\ln(1+S)}\right), \tag{30}$$

where $S = \frac{3\sqrt{3}}{4c^{3/2}}\sqrt{\frac{\sigma_3}{H}}$.

Proof As in the proof of Theorem 2, denote $F_k = F(x_k) - F^*$ and $g_k = \|F'(x_k)\|_*$. Then, we have

$$\ln\frac{1}{F_{k+1}} - \ln\frac{1}{F_k} = \ln\left(1 + \frac{F_k - F_{k+1}}{F_{k+1}}\right) \overset{(21)}{\geq} \ln\left(1 + \frac{\sqrt{3}g_{k+1}^2}{2\sqrt{H}c^2 g_k^{1/2} F_{k+1}}\right)$$
$$\overset{(29)}{\geq} \ln\left(1 + \frac{3}{4c^2}\sqrt{\frac{3\sigma_3}{H}}\cdot\frac{g_{k+1}^{1/2}}{g_k^{1/2}}\right) = \ln\left(1 + S\cdot\sqrt{\frac{g_{k+1}}{c g_k}}\right),$$

where $S = \frac{3}{4c^{3/2}}\sqrt{\frac{3\sigma_3}{H}}$. Denote $\tau_k = \sqrt{\frac{g_{k+1}}{c g_k}} \overset{(20)}{\leq} 1$. Since $\ln(\cdot)$ is a concave function, we have $\ln(1 + S\tau_k) \geq \tau_k \ln(1+S)$. Hence,

$$\xi_k \overset{\text{def}}{=} \ln\frac{g_0 D}{F_k} \;\geq\; \ln\frac{F_0}{F_k} \geq \ln(1+S)\sum_{i=0}^{k-1}\tau_i \;\geq\; \frac{k}{c^{1/2}}\ln(1+S)\left(\prod_{i=0}^{k-1}\frac{g_{i+1}^{1/2}}{g_i^{1/2}}\right)^{1/k}$$
$$= \frac{k}{c^{1/2}}\ln(1+S)\left(\frac{g_k}{g_0}\right)^{1/(2k)}.$$

Note that $\left(\frac{g_k}{g_0}\right)^{1/(2k)} = \exp\left(-\frac{1}{2k}\ln\frac{g_0}{g_k}\right) \geq 1 + \frac{1}{2k}\ln\frac{g_k}{g_0} \geq 1 + \frac{1}{2k}\ln\frac{F_k}{g_0 D} = 1 - \frac{1}{2k}\xi_k$. Thus,

$$\xi_k \geq \frac{k\ln(1+S)}{c^{1/2} + \frac{1}{2}\ln(1+S)},$$

and this is inequality (30). □

Remark 3 in accordance to the estimate (30), the highest rate of convergence corresponds to the maximal value of S. This means that we need to minimize the factor $c^{3/2}H^{1/2}$ in H. The optimal value is given by $H_\# = 3\sigma L_2$. In this case,

$$S = \sigma\sqrt{\frac{\sigma_3}{6L_2}} \;>\; 0.4\sigma\sqrt{\frac{\sigma_3}{L_2}}. \tag{31}$$

Note that this condition number also corresponds to the global convergence of the Cubically regularized Newton method [8].

Finally, let us prove a superlinear rate of local convergence for scheme (22).

Theorem 4 *Let the Hessian be Lipschitz continuous with constant L_2. Let function $f(\cdot)$ be strongly convex on* dom ψ *with parameter $\mu > 0$. If $H \geq L_2$, then for all $k \geq 0$ we have*

$$\|F'(x_{k+1})\|_* \leq \frac{2c}{\mu}\sqrt{\frac{H}{3}}\|F'(x_k)\|_*^{3/2}. \tag{32}$$

Proof Indeed, for any $k \geq 0$ we have

$$\begin{aligned}\frac{\mu}{2}\|x_{k+1} - x_k\|^2 &\leq \frac{1}{2}\langle\nabla^2 f(x_k)(x_{k+1} - x_k), x_{k+1} - x_k\rangle \\ &\overset{(19)}{\leq} F(x_k) - F(x_{k+1}) \;\leq\; \|F'(x_k)\|_*\|x_k - x_{k+1}\|.\end{aligned}$$

Therefore,

$$\begin{aligned}\|F'(x_{k+1})\|_* &\overset{(20)}{\leq} \sigma c A_k\|x_{k+1} - x_k\| \;\leq\; \frac{2\sigma c}{\mu}A_k\|F'(x_k)\|_* \\ &\overset{(16)}{=} \frac{2c}{\mu}\sqrt{\frac{H}{3}}\|F'(x_k)\|_*^{3/2}.\end{aligned}$$

□

Thus, the region of superlinear convergence of method (22) is as follows:

$$\mathcal{R}_Q \overset{\text{def}}{=} \left\{x \in \text{dom}\,\psi : \|F'(x)\|_* \leq \frac{3\mu^2}{4Hc^2}\right\}. \tag{33}$$

Note that outside this region, the constant of strong convexity of the objective function in problem (11) with $A = A_H(x)$ satisfies the following lower bound:

$$\sigma A_H(x) \geq \frac{\mu}{2c}, \quad x \notin \mathcal{R}_Q. \tag{34}$$

4 Estimating the norm of the gradient

Let us estimate the efficiency of method (22) in decreasing the norm of gradients. For that, we are going to derive an upper bound for the number of steps N of method (22), for which we still have

$$\|F'(x_k)\|_* \geq \delta > 0, \quad 0 \leq k \leq N. \tag{35}$$

We will see that global complexities of our method for minimizing the gradient norm in convex case are the same as that one of the basic Cubic Newton [10].

In this section, we use notation of Sect. 3:

$$F_k = F(x_k) - F^*, \quad g_k = \|F'(x_k)\|_*.$$

Firstly, consider the case when the smooth component $f(\cdot)$ in the objective function of problem (4) satisfies condition (8). Then

$$F_k - F_{k+1} \overset{(21)}{\geq} \kappa \frac{g_{k+1}^2}{g_k^{1/2}}, \quad \kappa \overset{\text{def}}{=} \frac{1}{2c^2}\sqrt{\frac{3}{H}}. \tag{36}$$

It is convenient to assume that the number of iteration N of the method is a multiple of three:

$$N = 3m, \quad m \geq 1. \tag{37}$$

Then for the last m iterations of the scheme we have

$$\begin{aligned} F_{2m} \geq F_{2m} - F_{3m} &\geq \kappa \sum_{i=0}^{m-1} \frac{g_{2m+i+1}^2}{g_{2m+i}^{1/2}} \overset{(35)}{\geq} \kappa\delta^{3/2} \sum_{i=0}^{m-1} \frac{g_{2m+i+1}^{1/2}}{g_{2m+i}^{1/2}} \\ &\geq \kappa m \delta^{3/2} \left(\frac{g_{3m}^{1/2}}{g_{2m}^{1/2}}\right)^{1/m} \overset{(35)}{\geq} \kappa m \delta^{3/2} \left(\frac{\delta^{1/2}}{g_{2m}^{1/2}}\right)^{1/m}. \end{aligned} \tag{38}$$

At the same time, for the first $2m$ iterations we obtain

$$\frac{1}{F_{2m}^{1/2}} - \frac{1}{F_0^{1/2}} \overset{(24)}{\geq} \frac{2m}{4c^2}\sqrt{\frac{3}{HD^3}}\left(\frac{F_{2m}g_{2m}^{1/2}}{F_0 g_0^{1/2}}\right)^{1/(2m)} = \kappa m D^{-3/2}\left(\frac{F_{2m}g_{2m}^{1/2}}{F_0 g_0^{1/2}}\right)^{1/(2m)}. \tag{39}$$

Therefore,

$$\left(\frac{1}{F_{2m}^{1/2}} - \frac{1}{F_0^{1/2}}\right)^2 \overset{(39)}{\geq} (km)^2 D^{-3}\left(\frac{F_{2m}g_{2m}^{1/2}}{F_0 g_0^{1/2}}\right)^{1/m}. \tag{40}$$

Note that the power of g_{2m} in the last term is equal to that one of $\frac{1}{g_{2m}}$ in (38). This explains our choice $2m$ for the length of the first stage.

Hence, using both inequalities (38) and (40), we obtain the following:

$$1 \geq \left(1 - \sqrt{\frac{F_{2m}}{F_0}}\right)^2 = \left(\frac{1}{F_{2m}^{1/2}} - \frac{1}{F_0^{1/2}}\right)^2 \cdot F_{2m} \geq \left(\frac{\kappa m \delta^{1/2}}{D}\right)^3 \left(\frac{F_{2m}\delta^{1/2}}{F_0 g_0^{1/2}}\right)^{1/m}$$

Note that $g_{2m} \overset{(20)}{\leq} c^{2m} g_0$. Therefore,

$$F_{2m} \overset{(38)}{\geq} \kappa m \delta^{3/2}\left(\frac{\delta^{1/2}}{c^m g_0^{1/2}}\right)^{1/m},$$

and we obtain

$$1 \geq \left(\frac{\kappa m \delta^{1/2}}{D}\right)^3 \left(\frac{\kappa m \delta^2}{c F_0 g_0^{1/2}} \cdot \left(\frac{\delta^{1/2}}{g_0^{1/2}}\right)^{1/m}\right)^{1/m}$$
$$\geq \left(\frac{\kappa m \delta^{1/2}}{D}\right)^{3+\frac{1}{m}} \left(\frac{\delta^{1/2}}{g_0^{1/2}}\right)^{(3+\frac{1}{m})\frac{1}{m}} \left(c\right)^{-\frac{1}{m}}.$$

Thus, we can prove the following theorem.

Theorem 5 *Let the Hessian be Lipschitz continuous with constant L_2. Fix $H \geq L_2$ and some $\delta > 0$. Then, the number of iterations of method (22) to reach small norm of the gradient $\|F'(x_N)\|_* \leq \delta$ satisfies the following bound:*

$$N \leq 2c^2\sqrt{\frac{3HD^2}{\delta}} + \frac{3}{2}\ln\frac{g_0}{\delta} + \ln c. \tag{41}$$

Proof Indeed,

$$1 \geq \frac{\kappa m \delta^{1/2}}{D} \left(\frac{\delta}{g_0}\right)^{\frac{1}{2m}} \left(c\right)^{-\frac{1}{3m+1}} = \frac{\kappa m \delta^{1/2}}{D} \exp\left(-\frac{1}{2m} \ln\left[\frac{g_0}{\delta}\left(c\right)^{\frac{2m}{3m+1}}\right]\right)$$
$$\geq \frac{\kappa \delta^{1/2}}{D}\left(m - \frac{1}{2}\ln\frac{g_0}{\delta} - \frac{m}{3m+1}\ln c\right) \geq \frac{\kappa \delta^{1/2}}{D}\left(m - \frac{1}{2}\ln\frac{g_0}{\delta} - \frac{1}{3}\ln c\right),$$

and this is inequality (41). □

Finally, let us estimate the efficiency of method (22) under additional assumption of uniform convexity (28). From the proof of Theorem 3, we know that

$$\ln\frac{F_0}{F_{2m}} \geq \frac{2m}{c^{1/2}}\ln(1+S)\left(\frac{g_{2m}}{g_0}\right)^{1/(2m)} \geq \frac{2m}{c^{1/2}}\ln(1+S)\exp\left(-\frac{1}{2m}\ln\frac{g_0}{g_{2m}}\right)$$
$$\geq \frac{1}{c^{1/2}}\ln(1+S)\left(2m - \ln\frac{g_0}{g_{2m}}\right) \overset{(35)}{\geq} \frac{1}{c^{1/2}}\ln(1+S)\left(2m - \ln\frac{g_0}{\delta}\right).$$

On the other hand,

$$\ln F_{2m} \overset{(38)}{\geq} \ln(\kappa m \delta^{3/2}) + \frac{1}{2m}\ln\frac{\delta}{g_{2m}} \overset{(20)}{\geq} \ln(\kappa m \delta^{3/2}) + \frac{1}{2m}\ln\frac{\delta}{g_0} - \ln c.$$

Thus,

$$\ln(cF_0) \geq \frac{2m}{c^{1/2}}\ln(1+S) - \frac{1}{c^{1/2}}\ln(1+S)\ln\frac{g_0}{\delta} + \ln(\kappa m \delta^{3/2}) + \frac{1}{2m}\ln\frac{\delta}{g_0}.$$

In other words,

$$\ln\frac{cF_0}{\kappa g_0^{3/2}} \geq \frac{2m}{c^{1/2}}\ln(1+S) - \frac{1}{c^{1/2}}\ln(1+S)\ln\frac{g_0}{\delta} + \frac{3}{2}\ln\frac{\delta}{g_0} - \ln\frac{1}{m} + \frac{1}{2m}\ln\frac{\delta}{g_0}$$
$$= \frac{2m}{c^{1/2}}\ln(1+S) - \left[\frac{1}{2m} + \frac{1}{c^{1/2}}\ln(1+S) + \frac{3}{2}\right]\ln\frac{g_0}{\delta} - \ln\frac{1}{m}.$$

Thus, we have proved the following theorem.

Theorem 6 *Let the Hessian be Lipschitz continuous with constant L_2. Let $F(\cdot)$ satisfies condition (28). Fix $H \geq L_2$ and some $\delta > 0$. Then, the number of iterations of method (22) to reach small norm of the gradient $\|F'(x_N)\|_* \leq \delta$ satisfies the following bound:*

$$N \leq \frac{3c^{1/2}}{2\ln(1+S)}\left\{\ln\frac{cF_0}{\kappa g_0^{3/2}} + \left[\frac{1}{2m} + \frac{1}{c^{1/2}}\ln(1+S) + \frac{3}{2}\right]\ln\frac{g_0}{\delta}\right\}$$
$$\overset{(29)}{\leq} \frac{3c^{1/2}}{2\ln(1+S)}\ln\frac{2c}{3\kappa\sqrt{\sigma_3}} + 3\left[\frac{1}{2} + \frac{c^{1/2}}{\ln(1+S)}\right]\ln\frac{g_0}{\delta}. \tag{42}$$

5 Adaptive search procedure

The main advantage of the method (22) consists in its easy implementation. Indeed, in the case $\psi(\cdot) \equiv 0$ with $\mathrm{dom}\,\psi = \mathbb{E}$, the iteration (11) is reduced mainly to matrix inversion, the very standard operation of Linear Algebra, which is available in the majority of software packages. However, for the better performance of this scheme, it is necessary to apply a dynamic strategy for updating the step-size coefficient H. Let us show how this can be done.

Gradient Regularizaion of Newton Method with Adaptive Search

Initialization. Choose $H_0 \le L_2, x_0 \in \mathrm{dom}\,\psi$, and $F_0' \in \partial F(x_0)$.

k**th iteration** ($k \ge 0$)**.** 1). Set $g_k = \|F_k'\|_*$.

2). **For** $i_k = 0, 1, \ldots$ **do:**

Set $H = 2^{i_k} H_k$ and $T = T_{A_H(x_k)}(x_k)$

Until $f(T) \le f(x_k) + \langle \nabla f(x_k), T - x_k \rangle + \frac{1}{2}\nabla^2 f(x_k)[T - x_k]^2 + \frac{H}{6}\|T - x_k\|^3$.

3). Set $x_{k+1} = T$, $H_{k+1} = \max\{H_0, 2^{i_k - 1} H_k\}$, and

$$F'_{k+1} = \nabla f(x_{k+1}) - \nabla f(x_k) - \nabla^2 f(x_k)(x_{k+1} - x_k) - A_k(\nabla d(x_{k+1}) - \nabla d(x_k))$$

(43)

For the initialization, we need an initial guess H_0 for the regularization parameter, which can be an *arbitrary* sufficiently small number.

Note that this scheme does not depend on any particular value of the Lipschitz constant. By definitions of the updates and from inequality (10), we conclude that inequalities $H_0 \le H_k \le L_2$ and $2^{i_k} H_k \le 2L_2$ imply $H_{k+1} \le L_2$. Thus,

$$H_0 \le H_k \le L_2, \qquad 2^{i_k} H_k \le 2L_2, \qquad k \ge 0. \tag{44}$$

Hence, from Theorem 1, we have the following progress established for each iteration $k \ge 0$:

$$F(x_k) - F(x_{k+1}) \ge \frac{1}{2c_0^2}\sqrt{\frac{3}{2L_2}} \cdot \frac{\|F'(x_{k+1}\|_*^2}{\|F'(x_k)\|_*^{1/2}},$$

where

$$c_0 \stackrel{\mathrm{def}}{=} \sigma^{-1} + \frac{3L_2}{2H_0}.$$

Repeating the reasoning of Theorem 2, we obtain the following complexity result.

Theorem 7 *Let the Hessian be Lipschitz continuous with constant L_2. Let $F(x_k) - F^* \geq \epsilon$ for some iteration $k \geq 0$ of method (43). Then,*

$$k \leq 4c_0^2\sqrt{\frac{2L_2D^3}{3\epsilon}} + \ln\frac{(F(x_0) - F^*)\|F'(x_0)\|_*^{1/2}D^{1/2}}{\epsilon^{3/2}}.$$

Note that some scaling of the domain or the target objective may affect the fixed choice of regularization parameter in the basic scheme (22). At the same time, we expect the adaptive method (43) to be robust with respect to these changes.

6 Acceleration

Let us present a conceptual acceleration of our method, that is based on the contracting proximal iterations [7].

First, we fix an auxiliary prox-function $\phi(\cdot)$ that we assume to be uniformly convex of degree three with respect to the initial norm:

$$\beta_\phi(x, y) = \phi(y) - \phi(x) - \langle\nabla\phi(x), y - x\rangle \;\geq\; \frac{1}{3}\|y - x\|^3, \quad \forall x, y \in \mathrm{dom}\,\psi. \tag{45}$$

At each iteration $k \geq 0$ of the accelerated scheme, we form the following functions:

$$g_{k+1}(x) \stackrel{\text{def}}{=} B_{k+1}f\Big(\frac{b_{k+1}x + B_kx_k}{B_{k+1}}\Big),$$
$$h_{k+1}(x) \stackrel{\text{def}}{=} g_{k+1}(x) + b_{k+1}\psi(x) + \beta_\phi(v_k; x),$$

where $\{b_k\}_{k\geq 1}$ is a sequence of positive numbers, $B_k \stackrel{\text{def}}{=} \sum_{i=1}^k b_i$, $B_0 \stackrel{\text{def}}{=} 0$, and

$$\{x_k\}_{k\geq 0}, \quad \{v_k\}_{k\geq 0}, \quad x_0 = v_0,$$

are sequences of trial points that belong to $\mathrm{dom}\,\psi$.

Note that the derivatives of $g_{k+1}(\cdot)$ and $f(\cdot)$ are related as follows:

$$D^3g_{k+1}(x) \equiv \frac{b_{k+1}^3}{B_{k+1}^2}D^3f\Big(\frac{b_{k+1}x + B_kx_k}{B_{k+1}}\Big).$$

For simplicity of the presentation, we assume that f is three times differentiable on the open set containing $\mathrm{dom}\,\psi$. Let us choose

$$b_k := \frac{k^2}{9L_2(f)}.$$

Then, $B_k = \frac{1}{9L_2(f)} \sum_{i=1}^{k} i^2 \geq \frac{k^3}{27L_2(f)}$. Therefore, for any $h \in \mathbb{E}$:

$$|D^3 g_{k+1}(x)[h]^3| \leq \frac{1}{L_2(f)} \left| D^3 f\left(\frac{b_{k+1}x + B_k x_k}{B_{k+1}}\right) \right| \quad \leq \quad \|h\|^3,$$

thus $L_2(g_{k+1}) = 1$, and we can minimize objective h_{k+1} very efficiently by using our method (22). Namely, in order to find a point v with a small norm of a subgradient:

$$\|g\|_* \leq \delta, \qquad g \in \partial h_{k+1}(v),$$

the method needs to do no more than

$$N \overset{(41)}{\leq} \tilde{O}\left(\ln \frac{1}{\delta}\right)$$

steps, where $\tilde{O}(\cdot)$ hides absolute constants and logarithmic factors that depends on the initial residual and subgradient norm.

Let us write down the accelerated algorithm.

Acceleration of Newton Method with Grad. Regularization

Initialization. Choose $x_0 \in \operatorname{dom} \psi$ and $\delta > 0$. Set $v_0 = x_0$, $B_0 = 0$.

k**th iteration** ($k \geq 0$)**.** 1). Set $b_{k+1} = \frac{(k+1)^2}{9L_2(f)}$ and $B_{k+1} = B_k + b_{k+1}$. (46)

2). Form the auxiliary objective $h_{k+1}(\cdot)$. Find a point v_{k+1} by method (22) such that $\|g\|_* \leq \delta$ for some $g \in \partial h_{k+1}(v_{k+1})$.

3). Set $x_{k+1} = \frac{b_{k+1}v_{k+1} + B_k x_k}{B_{k+1}}$.

Applying directly Theorem 3.2 and the corresponding Corollary 3.3 from [7], we get the following complexity bound.

Theorem 8 *Let the Hessian be Lipschitz continuous with constant $L_2(f)$. Let us set $\delta = \frac{1}{2 \cdot 3^{7/3}} \cdot \left(\frac{\epsilon}{L_2(f)}\right)^{2/3}$ in method (46), and let*

$$k = \left\lceil \left(2 \cdot 3^3\right)^{1/2} \cdot \left(\frac{L_2(f)\beta_\psi(x_0; x^*)}{\epsilon}\right)^{1/3} \right\rceil.$$

Then, $F(x_k) - F^* \leq \epsilon$. □

7 Experiments

In this section, let us present computational results for solving the unconstrained minimization problem,

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

with objective that is a smooth convex approximation of pointwise maximum:

$$f(x) := \mu \log\left(\sum_{i=1}^m \exp\left(\frac{\langle a_i, x\rangle - b_i}{\mu}\right)\right) \quad \approx \quad \max_{i=1}^m \left[\langle a_i, x\rangle - b_i\right].$$

The problems of this type are important in applications with *minimax strategies for matrix games* and ℓ_∞*-regression* [17].

The vectors $\{a_i \in \mathbb{R}^n\}_{i=1}^m$ and numbers $\{b_i \in \mathbb{R}\}_{i=1}^m$ are given data, while $\mu > 0$ is a fixed parameter of smoothing.

Let us fix matrix $B := \sum_{i=1}^m a_i a_i^T$, which we assume to be positive definite (otherwise, it is possible to reduce the dimensionality of the initial problem), and we use the following Euclidean norms:

$$\|x\| := \langle Bx, x\rangle^{1/2}, \qquad \|g\|_* \;:=\; \langle g, B^{-1}g\rangle^{1/2},$$

respectively for the variables and for the gradients. We also know the corresponding Lipschitz constant for the Hessian, that is (see, e.g. Example 1.3.6 in [6])

$$L_2 := \frac{2}{\mu^2}. \tag{47}$$

To generate the data, we sample random elements $\{\bar{a}_i \in \mathbb{R}^n, b_i \in \mathbb{R}\}_{i=1}^m$ from the uniform distribution on $[-1, 1]$, and form an auxiliary function

$$\bar{f}(x) := \mu \log\left(\sum_{i=1}^m \exp\left(\frac{\langle \bar{a}_i, x\rangle - b_i}{\mu}\right)\right).$$

Then, we set

$$a_i := \bar{a}_i - \nabla \bar{f}(0), \qquad 1 \leq i \leq m.$$

Thus we ensure to have the optimum at the origin, since $\nabla f(0) = 0$. We start the methods from $x_0 := (1, 1, \ldots, 1)$.

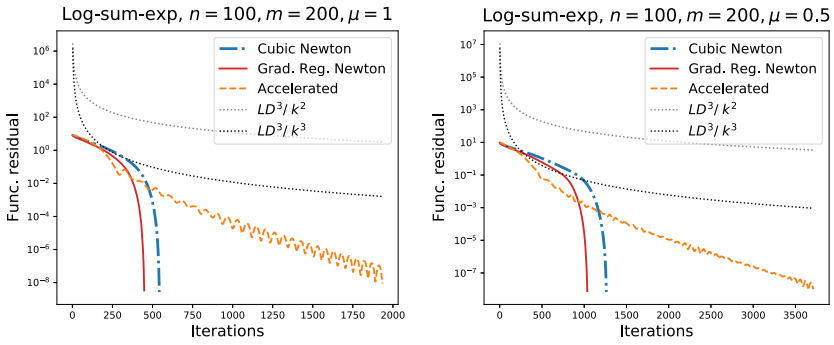

Fig. 1 Newton methods with Cubic and with Gradient regularization, and the accelerated scheme. Lipschitz constant is fixed

We study the performance of the Newton method with Gradient regularization and with Cubic regularization [19] on this problem. Also, we compare our accelerated scheme (46) with the basic methods.

We use the following scaling function for this problem, as in Example 1:

$$d(x) = \frac{1}{2}\|x\|^2 \; = \; \frac{1}{2}\langle Bx, x\rangle.$$

The subproblem in our methods is solved exactly by using the standard matrix inversion. For the Cubic Newton, one need to find a root of a one-dimensional nonlinear equation at each step (see Section 5 in [19]). To solve it, we apply the classical univariate Newton method and use the value $\epsilon = 10^{-8}$ as a target tolerance in terms of the function value.

Regularization parameter is fixed according to the theory (47). The results are shown in Fig. 1. We see that both algorithms show reasonably good performance, which is better than the theoretical prediction of the global behaviour. The Newton method with Gradient regularization possesses the best convergence rate. Accelerated scheme has an improvement in the rate in the beginning, but the basic methods are better for the higher level of the accuracy due to their superlinear local convergence.

In the following experiment, we compare the uses of the fixed Lipschitz constant with the adaptive search procedure for our method. The results are shown in Fig. 2. We see that the adaptive methods show the best performance. At the same time, iterations of the Gradient regularization are much cheaper which results in better computational time.

Finally, we compare our approach with iterations of the *damped* Newton method with line search. For this problem, the Hessian is often degenerate, thus we use a small perturbation to correct the matrix. Namely, we consider the following iterations:

$$x_{k+1} = x_k - \alpha_k \Big(\nabla^2 f(x_k) + \tau B\Big)^{-1} \nabla f(x_k), \quad k \geq 0,$$

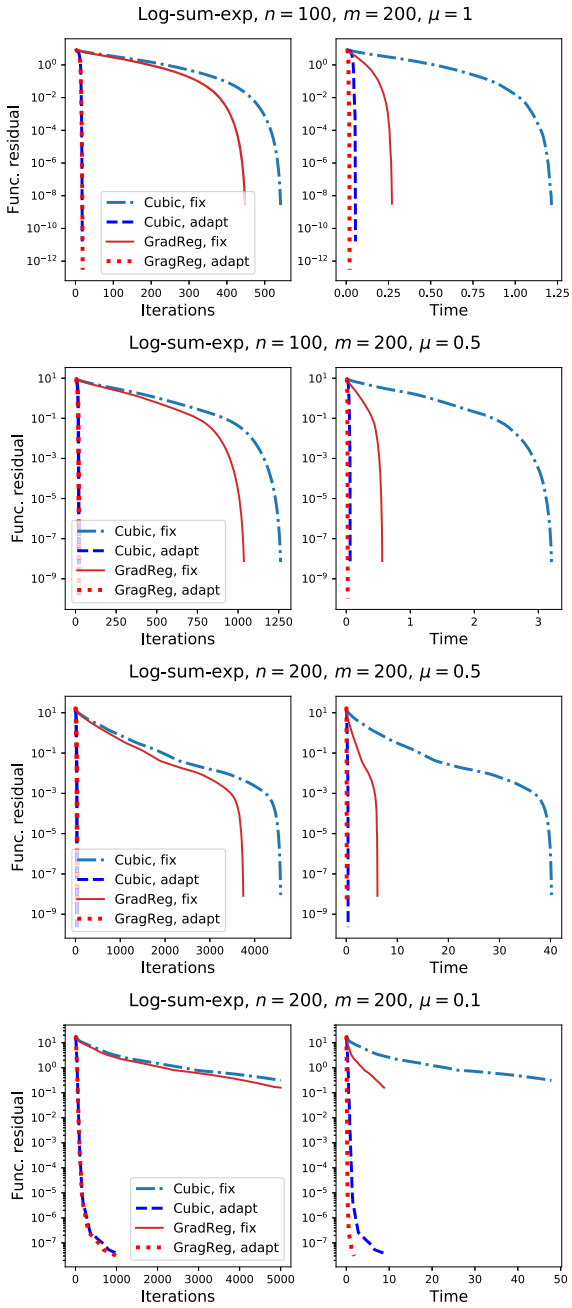

Fig. 2 The effect of using adaptive line search in the regularized Newton methods

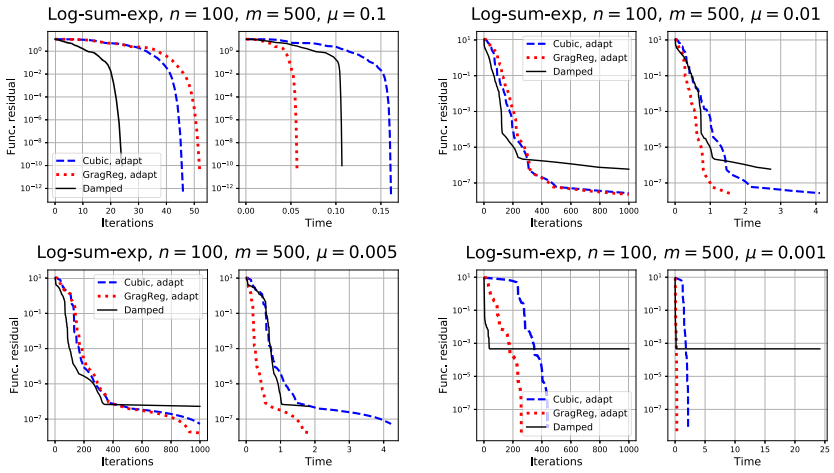

Fig. 3 Comparison of the regularized Newton methods with the damped Newton algorithm

where τ is a fixed small parameter (we set $\tau = 10^{-6}$ which was tuned to have the best performance), and α_k is chosen by the standard backtracking line search to satisfy the following condition:

$$f(x_k) - f(x_{k+1}) \geq \frac{\alpha_k}{2} \|\nabla f(x_k)\|_*^2.$$

The results are presented in Fig. 3. We see that the damped Newton method is sensitive to the choice of perturbation parameter τ, while the method with Gradient regularization shows the most robust and efficient performance for all problem instances.

8 Discussion

In this paper, we have analyzed the global behaviour of the Newton method with a general Bregman regularizer, whose regularization parameter is chosen to be proportional to the square root of the current gradient norm.

We demonstrated that our scheme works with the composite form of the convex optimization problem. For the Euclidean norms, this approach can be seen as a *relaxation* of the Cubically regularized Newton method, achieving the same global convergence rates.

A significant advantage of the gradient regularization scheme is a simpler structure of the subproblem, which does not need auxiliary one-dimensional minimizations that are required in the cubic regularization. As a consequence, the subproblem becomes suitable for the large-scale case as for employing the Conjugate Gradient methods.

It is a favorable feature of our methods that regularization parameter always depends on the *current* iterate only. Therefore, it seems to be convenient for the use in stochastic

optimization. We believe that this property could fit well with the broad family of stochastic second-order methods based on the Cubic regularization (see [4, 12, 14]). We keep the development of such schemes for further investigation.

Another important direction is an extension of our results to nonconvex optimization problems. It seems to be a challenging question since our current analysis heavily relies on positive semidefiniteness of the Hessian. It is needed to ensure a bound for the step length (see Lemma 1). Therefore, to tackle nonconvex problems, some modifications of our analysis have to be made.

Acknowledgements We are very thankful to Associate Editor and two anonymous referees for valuable comments that significantly improved the initial version of this paper.

References

1. Ben-Tal, A., Nemirovski, A.: Lectures on Modern Convex Optimization (2020). SIAM, Philadelphia (2021)
2. Carmon, Y., Duchi, J.C.: Gradient descent efficiently finds the cubic-regularized non-convex Newton step. arXiv preprint arXiv:1612.00547 (2016)
3. 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)
4. Cartis, C., Scheinberg, K.: Global convergence rate analysis of unconstrained optimization methods based on probabilistic models. Math. Program. **169**(2), 337–375 (2018)
5. Dennis, J.E., Jr., Schnabel, R.B.: Numerical Methods for Unconstrained Optimization and Nonlinear Equations. SIAM, Philadelphia (1996)
6. Doikov, N.: New second-order and tensor methods in convex optimization. Ph.D. thesis, Université catholique de Louvain (2021)
7. Doikov, N., Nesterov, Y.: Contracting proximal methods for smooth convex optimization. SIAM J. Optim. **30**(4), 3146–3169 (2020)
8. Doikov, N., Nesterov, Y.: Minimizing uniformly convex functions by cubic regularization of newton method. J. Optim. Theory Appl. 1–23 (2021)
9. Fan, J.Y, Ai, W.B., Zhang, Q.Y.: A line search and trust region algorithm with trust region radius converging to zero. J. Comput. Math. 865–872 (2004)
10. Grapiglia, G.N., Nesterov, Y.: Tensor methods for finding approximate stationary points of convex functions. Optim. Methods Softw. 1–34 (2020)
11. Gratton, S., Toint, P.L.: Adaptive regularization minimization algorithms with non-smooth norms and Euclidean curvature. arXiv preprint arXiv:2105.07765 (2021)
12. Hanzely, F., Doikov, N., Richtárik, P., Nesterov, Y.: Stochastic subspace cubic Newton method. In: International Conference on Machine Learning, pp. 4027–4038. PMLR (2020)
13. Khanh, P.D., Mordukhovich, B., Phat, V.T., Tran, B.D.: Globally convergent coderivative-based generalized Newton methods in nonsmooth optimization. arXiv preprint arXiv:2109.02093 (2021)
14. Kovalev, D., Mishchenko, K., Richtárik, P.: Stochastic Newton and cubic Newton methods with simple local linear-quadratic rates. arXiv preprint arXiv:1912.01597 (2019)
15. Martínez, J.M., Raydan, M.: Cubic-regularization counterpart of a variable-norm trust-region method for unconstrained minimization. J. Glob. Optim. **68**(2), 367–385 (2017)

16. Mishchenko, K.: Regularized Newton method with global $O(1/k^2)$ convergence. In: Proceedings of the Beyond First-Order Methods in ML Systems Workshop at the 38th International Conference on Machine Learning, vol. 139. PMLR (2021)
17. Nesterov, Y.: Smooth minimization of non-smooth functions. Math. Program. **103**(1), 127–152 (2005)
18. Nesterov, Y.: Lectures on Convex Optimization, vol. 137. Springer, Berlin (2018)
19. Nesterov, Y., Polyak, B.: Cubic regularization of Newton's method and its global performance. Math. Program. **108**(1), 177–205 (2006)
20. Polyak, R.: Regularized Newton method for unconstrained convex optimization. Math. Program. **120**(1), 125–145 (2009)