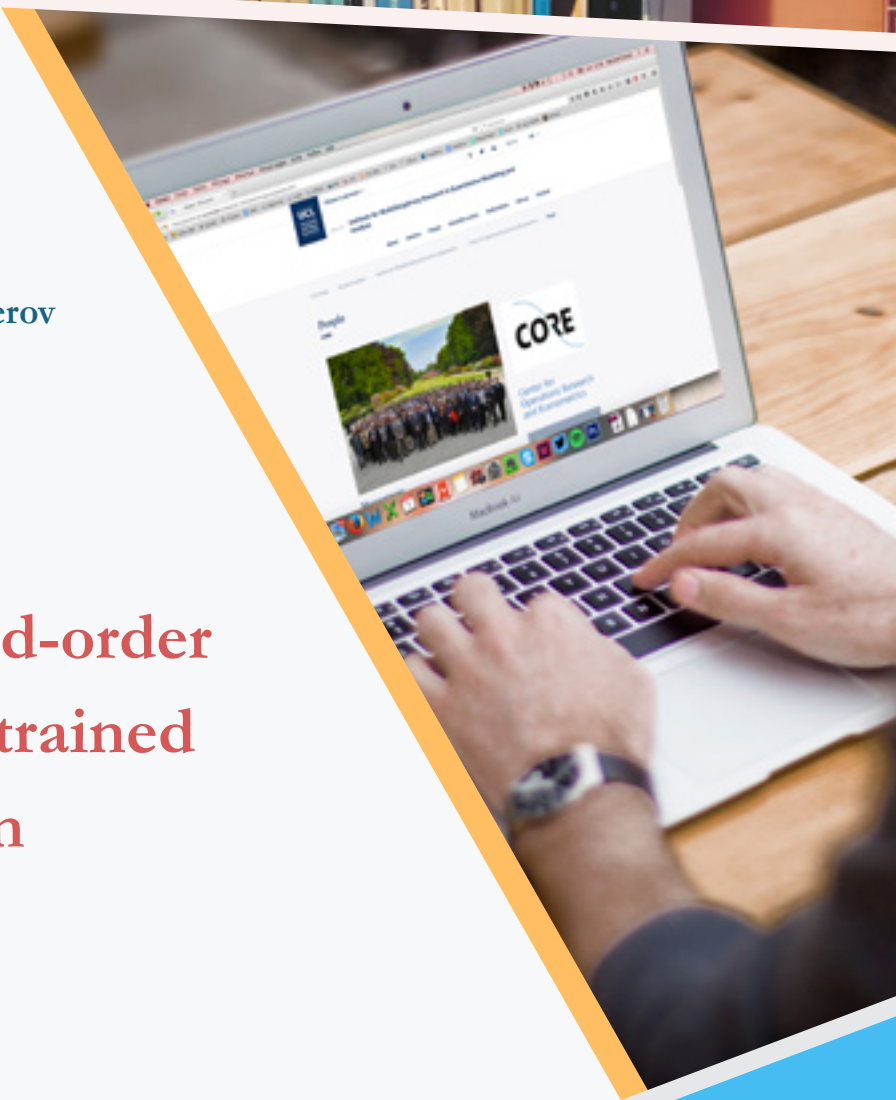
2018/32

DP

Pavel Dvurechensky and Yurii Nesterov

Global performance guarantees of second-order methods for unconstrained convex minimization



CORE
CENTER FOR OPERATIONS RESEARCH AND ECONOMETRICS

CORE

Voie du Roman Pays 34, L1.03.01

B-1348 Louvain-la-Neuve

Tel (32 10) 47 43 04

Email: immaq-library@uclouvain.be

https://uclouvain.be/en/research-institutes/lidam/core/discussion-papers.html

CORE DISCUSSION PAPER

2018/32

Global performance guarantees of second-order methods for unconstrained convex minimization

P. Dvurechensky* and Yu. Nesterov †

December 20, 2018

Abstract

In this paper we make an attempt to compare two distinct branches of research on second-order optimization methods. The first one studies self-concordant functions and barriers, the main assumption being that the third derivative of the objective is bounded by the second derivative. The second branch studies cubic regularized Newton methods with main assumption that the second derivative is Lipschitz continuous. We develop new theoretical analysis for a path-following scheme for general self-concordant function, as opposed to classical path-following scheme developed for self-concordant barriers. We show that the complexity bound for this scheme is better than for Damped Newton Method. Next, we analyze an important subclass of general self-concordant function, namely a class of strongly convex functions with Lipschitz continuous second derivative and show that for this subclass cubic regularized Newton Methods give even better complexity bound.

Keywords: self-concordant function, Damped Newton Method, Cubic Regularized Newton Method, path-following method.

*Weierstrass Institute for Applied Analysis and Stochastics,
Mohrenstr. 39, 10117 Berlin, Germany; e-mail: pavel.dvurechensky@wias-berlin.de.

†Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL),
34 voie du Roman Pays, 1348 Louvain-la-Neuve, Belgium, and National Research Institute Higher School of Economics, e-mail: Yurii.Nesterov@uclouvain.be.
Scientific results of this paper were obtained with support of RSF Grant 17-11-01027.

1 Introduction

Motivation. Local performance guarantees for the second-order methods are known since the paper [2] (1948), where the author proved a local quadratic convergence of the Newton method under some natural assumptions (non-degeneracy of the Hessian at solution and local Lipschitz continuity of the Hessian). However, in some sense the quadratic convergence is too fast: each step of such methods doubles the number of right digits in the approximate solution. Therefore, the questions on acceleration of these schemes were never raised in the literature (see [1]). Moreover, during many years the only complexity results on the global performance of the second-order methods were obtained in the framework of the theory of self-concordant functions and barriers (see [6], [4]).

The situation was changed after the paper [7], where the first global complexity bounds were obtained for a special cubic regularization of the Newton method. Namely, it was shown that for convex function with globally Lipschitz continuous Hessian the Cubic Newton converges in function value as $O(\frac{1}{k^2})$, where k is the iteration counter. Very soon it was shown that this method can be accelerated up to the rate $O(\frac{1}{k^3})$ using the technique of estimating sequences (see [5]).

Thus, at this moment there exist two independent frameworks for complexity analysis of the second-order methods. One is based on the affine-invariant theory of self-concordant functions. And the second one assumes bounded third derivatives of the objective in a fixed Euclidean norm. The main goal of this paper is to show that these classes of problems do intersect and we can compare efficiency of the corresponding methods. For that, we derive new complexity bounds for a path-following method as applied to minimization of a self-concordant function. This result is new since the known complexity bounds are related to self-concordant *barriers* (see, for example, Section 4.2 in [4]). We compare our bounds with the complexity results for different versions of the Cubic Newton Method on the class of strongly convex functions with Lipschitz continuous Hessian. It appears that such functions are self-concordant. Our conclusion is that the latter methods are much more efficient on this particular class of self-concordant functions.

Contents. In Section 2 we recall the properties of self-concordant functions extending them to the case of non-standard self-concordant functions and analyze complexity of Damped Newton Method. Section 3 is devoted to the description and new analysis of the path-following scheme for general self-concordant functions. In Section 4 we discuss modifications of Damped Newton Method and path-following scheme with adaptive choice of the stepsize. Section 5 contains complexity analysis of strongly convex functions with Lipschitz-continuous Hessian by cubic regularized Newton methods. Finally, in Section 6 we present numerical experiments to compare Damped Newton Method and path-following scheme for general self-concordant functions as well as adaptive variants of these methods.

Notation. Given a function f with non-degenerate at any $x \in \mathbb{E}$ Hessian $\nabla^2 f(x)$, we denote

$$\|h\|_x \;=\; \langle \nabla^2 f(x)h, h\rangle^{1/2},\ h \in E, \quad \|g\|_x^* \;=\; \langle g, [\nabla^2 f(x)]^{-1} g\rangle^{1/2},\ g \in \mathbb{E}^*$$

and

$$\lambda_f(x) \;=\; \|\nabla f(x)\|_x^*, \quad x \in \mathbb{E}.$$

We define also two univariate functions $\omega(\tau) = \tau - \ln(1+\tau)$, $\omega_*(\tau) = -\tau - \ln(1-\tau)$, $\tau \geq 0$.

2 Minimizing self-concordant functions: Damped Newton Method

Let us start from a variant of definition of self-concordant function.

Definition 1 *Let function f from $\mathbb{C}^3$ be convex on $\mathbb{E}$. It is called a general self-concordant function if there exists a constant $M_f \geq 0$ such that for any point $x \in \mathbb{E}$ and direction $h \in E$ we have*

$$D^3 f(x)[h]^3 \quad \leq \quad 2M_f \langle \nabla^2 f(x)h, h\rangle^{3/2}. \tag{2.1}$$

If $M_f = 1$, then the function is called the standard self-concordat. □

It is clear that for any self-concordant function f, function

$$\tilde{f}(x) \quad \stackrel{\text{def}}{=} \quad M_f^2 f(x), \quad x \in \mathbb{E}, \tag{2.2}$$

is standard self-concordant. Standard self-concordant functions are more convenient for defining the self-concordant barriers (see [6]). However, in this paper we do not need the barrier property. Therefore, we will work directly with definition (2.1).

Taking into account normalization (2.2), we can rewrite all known properties of standard self-concordant functions for the general ones. Let us present the most important of them (see Section 4.1 in [4]). Denote

$$\|h\|_x \quad = \quad \langle \nabla^2 f(x)h, h\rangle^{1/2},\ h \in E, \quad \|g\|_x^* \ = \ \langle g, [\nabla^2 f(x)]^{-1} g\rangle^{1/2},\ g \in \mathbb{E}^*.$$

From now on, we assume that the Hessian $\nabla^2 f(x)$ is non-degenerate at any $x \in \mathbb{E}$. Denote

$$\lambda_f(x) \quad = \quad \|\nabla f(x)\|_x^*, \quad x \in \mathbb{E}. \tag{2.3}$$

For all $y \in E$ with $\|y - x\|_x < \frac{1}{M_f}$ we have

$$(1 - M_f\|y-x\|_x)^2 \nabla^2 f(x) \quad \preceq \quad \nabla^2 f(y) \ \preceq \ \tfrac{1}{(1-M_f\|y-x\|_x)^2} \nabla^2 f(x), \tag{2.4}$$

$$f(y) \quad \leq \quad f(x) + \langle \nabla f(x), y - x\rangle + \tfrac{1}{M_f^2}\omega_*(M_f\|y-x\|_x), \tag{2.5}$$

where $\omega_*(\tau) = -\tau - \ln(1-\tau)$. And for all $y \in \mathbb{E}$ we have

$$f(y) \quad \geq \quad f(x) + \langle \nabla f(x), y - x\rangle + \tfrac{1}{M_f^2}\omega(M_f\|y-x\|_x), \tag{2.6}$$

where $\omega(\tau) = \tau - \ln(1+\tau)$. Similarly, if $\delta \equiv \|\nabla f(x) - \nabla f(y)\|_x^* < \frac{1}{M_f}$, then

$$(1 - M_f\delta)^2 \nabla^2 f(x) \quad \preceq \quad \nabla^2 f(y) \ \preceq \ \tfrac{1}{(1-M_f\delta)^2} \nabla^2 f(x), \tag{2.7}$$

This relation follows from the trivial observation that the dual function

$$f_*(s) \quad = \quad \sup_{x \in \mathbb{E}} [\langle s, x\rangle - f(x)]$$

is also self-concordant on its domain with the same constant M_f.

Inequality (2.5) leads to the following bound.

Lemma 1 *Let $\lambda_f(x) < \frac{1}{M_f}$. Then*

$$f(x) - \min_{y \in \mathbb{E}} f(y) \quad \leq \frac{1}{M_f^2} \omega_*(M_f \lambda_f(x)). \tag{2.8}$$

Proof:
Note that for any $\tau \geq 0$ we have $\min_{y \in \mathbb{E}} \{\langle \nabla f(x), y - x \rangle : \ \|y - x\|_x = \tau\} = -\tau \lambda_f(x)$, and this minimum is attained at the point $y = x - \frac{\tau}{\lambda_f(x)} [\nabla^2 f(x)]^{-1} \nabla f(x)$. Therefore

$$\begin{aligned} \min_{y \in E} f(y) \quad &\overset{(2.6)}{\geq} \quad \min_{y \in \mathbb{E}} \{f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{M_f^2} \omega(M_f \|y - x\|_x)\} \\ &= \quad \min_{\tau \geq 0} \{f(x) - \tau \lambda_f(x) + \frac{1}{M_f^2} \omega(M_f \tau)\}. \end{aligned}$$

It remains to find optimal τ_* from the first-order optimality condition

$$\lambda_f(x) \quad = \quad \frac{1}{M_f} \cdot \frac{M_f \tau_*}{1 + M_f \tau_*}.$$

Thus, $\tau_* = \frac{\lambda_f(x)}{1 - M_f \lambda_f(x)}$. Note that $\tau_* > 0$ since $\lambda_f(x) < \frac{1}{M_f}$. We get inequality (2.8) by substituting this value into the objective function of the last minimization problem. $\square$

Minimizing the right-hand side of inequality (2.5) in y, we come to the following result.

Theorem 1 *Define*

$$x_+ \quad = \quad x - \frac{[\nabla^2 f(x)]^{-1} \nabla f(x)}{1 + M_f \lambda_f(x)}. \tag{2.9}$$

Then

$$f(x_+) \quad \leq \quad f(x) - \frac{1}{M_f^2} \omega(M_f \lambda_f(x)). \tag{2.10}$$

Moreover,

$$\lambda_f(x_+) \quad \leq \quad 2 M_f \lambda_f^2(x). \tag{2.11}$$

Proof:
Inequality (2.10) can be justified in the same way as it was done for inequality (2.8), using in the reasoning inequality (2.5) instead of (2.6).

Let us prove now inequality (2.11). Denote $\lambda = \lambda_f(x)$, $h = x_+ - x$, $r = \|h\|_x = \frac{\lambda}{1 + M_f \lambda}$. Hence, $\lambda = \frac{r}{1 - M_f r}$. Note that, since $r < \frac{1}{M_f}$,

$$\lambda_+^2 \quad \equiv \quad \langle \nabla f(x_+), [\nabla^2 f(x_+)]^{-1} \nabla f(x_+) \rangle \quad \overset{(2.4)}{\leq} \quad \frac{1}{(1 - M_f r)^2} \langle \nabla f(x_+), [\nabla^2 f(x)]^{-1} \nabla f(x_+) \rangle.$$

Without changing notation, we can associate with the Hessians symmetric positive-definite matrices. Then, denoting $G = [\nabla^2 f(x)]^{1/2} \succ 0$, we have

$$\begin{aligned} \nabla f(x_+) \quad &= \quad \nabla f(x) + \int_0^1 \nabla^2 f(x + \tau h) h d\tau \overset{(2.9)}{=} -(1 + M_f \lambda) \nabla^2 f(x) h + \int_0^1 \nabla^2 f(x + \tau h) h d\tau \\ &= \quad G \left[-(1 + M_f \lambda) I + G^{-1} \left(\int_0^1 \nabla^2 f(x + \tau h) d\tau \right) G^{-1} \right] G h. \end{aligned}$$

Note that

$$\int_0^1 \nabla^2 f(x+\tau h)d\tau \overset{(2.4)}{\preceq} \int_0^1 \tfrac{1}{(1-\tau M_f r)^2}\nabla^2 f(x)d\tau \;=\; \tfrac{1}{1-M_f r}\nabla^2 f(x) \;=\; (1+M_f\lambda)\nabla^2 f(x).$$

Thus, denoting $H = [\cdot]$, we can see that $H \preceq 0$. On the other hand,

$$\int_0^1 \nabla^2 f(x+\tau h)d\tau \overset{(2.4)}{\succeq} \int_0^1 (1-\tau M_f r)^2\nabla^2 f(x)d\tau \;=\; (1 - M_f r + \tfrac{1}{3}M_f^2 r^2)\nabla^2 f(x).$$

Thus, $H \succeq [-(1+M_f\lambda) + (1-M_f r)]I = [-(1+M_f\lambda) + \frac{1}{1+M_f\lambda}]I \succeq -2\lambda M_f I$, and we conclude that

$$\begin{aligned}\lambda_+^2 \;&\leq\; \tfrac{1}{(1-M_f r)^2}(\|GHGh\|_x^*)^2 \;=\; \tfrac{1}{(1-M_f r)^2}\langle GH^2Gh, h\rangle \leq \tfrac{(2M_f\lambda)^2}{(1-M_f r)^2}\langle G^2 h, h\rangle \\ &=\; \tfrac{(2M_f\lambda)^2}{(1-M_f r)^2}\langle \nabla^2 f(x)h, h\rangle \;=\; \tfrac{(2M_f\lambda)^2 r^2}{(1-M_f r)^2} \;=\; 4M_f^2\lambda^4. \qquad \square\end{aligned}$$

Now we can analyze the efficiency of Damped Newton Method

$$x_{k+1} \;=\; x_k - \tfrac{[\nabla^2 f(x_k)]^{-1}\nabla f(x_k)}{1+M_f\lambda_f(x_k)}, \quad k \geq 0. \tag{2.12}$$

as applied to the following minimization problem:

$$f^* \;=\; \min_{x\in\mathbb{E}} f(x), \tag{2.13}$$

where f is a general self-concordant function. We assume that the Hessian of this function at any point is positive definite and that the solution x^* of problem (2.13) do exist.

Our goal is to find an approximate solution of problem (2.13). In view of inequality (2.11), method (2.12) starts converging quadratically when it enters the region

$$\mathbb{Q} \;=\; \left\{x \in \mathbb{E} :\; \lambda_f(x) \leq \tfrac{1}{2M_f}\right\}.$$

This convergence is very fast and, in view of inequality (2.8), any reasonable accuracy in function value can be reached in a small number of iterations. Therefore, the main computational time is spent when $\lambda_f(x_k) \geq \frac{1}{2M_f}$. Denote by N the last iteration such that

$$\lambda_f(x_k) \;\geq\; \tfrac{1}{2M_f}, \quad k = 0, \ldots, N.$$

Then, in view of inequality (2.10), we have

$$N \leq \frac{\Delta(x_0)}{\omega\left(\frac{1}{2}\right)}, \quad \Delta(x_0) \overset{\text{def}}{=} M_f^2(f(x_0) - f^*). \tag{2.14}$$

Let us show that $\Delta(x_0)$ is a natural complexity measure of our problem class. In order to see this, let us attribute to our objects some physical units. Denote the units for measuring the function value by μ_f, and the units for measuring the argument by μ_x. Then, the units for measuring the gradient is $\mu_g = \mu_f/\mu_x$. The Hessian is measured

in $\mu_h = \mu_f/\mu_x^2$, and the third derivative is measured in $\mu_t = \mu_f/\mu_x^3$. Thus, in view of definition (2.1), the units for measuring the constant M_f are

$$\mu_s \;=\; \mu_t\mu_x^3/(\mu_h\mu_x^2)^{3/2} \;=\; \mu_f^{-1/2}.$$

Note that the number of iterations is an integer number with no physical dimension (scalar). Therefore, for using the constant M_f in the bounds for the number of iterations, it must be multiplied by something having physical dimension $\mu_f^{1/2}$. The simplest way to do this is to define the characteristic $\Delta(x_0)$ as in (2.14). In the sequel, we will use $\Delta(x_0)$ as the main characteristic of complexity of problem (2.13). By the way, it is important that we can use the characteristics of our problem as arguments of nonlinear univariate functions only by transforming them in a scalar form. For example, the values $M_f\lambda_f(x)$ and $M_f\|h\|_x$ has no physical dimension.

3 Minimizing self-concordant functions: path-following scheme

Let us estimate the complexity of solving the problem (2.13) by a path-following scheme. Let us start from some $x_0 \in \mathbb{E}$. Define the central path $x(t)$, $0 \le t \le 1$, by the following equation:

$$\nabla f(x(t)) \;=\; t\nabla f(x_0). \tag{3.1}$$

Clearly, $x(1) = x_0$ and $x(0) = x^*$. Note that this is a trajectory of minimizers of the following parametric family of general self-concordant functions:

$$x(t) \;=\; \arg\min_{x\in\mathbb{E}} \left\{ f_t(x) \stackrel{\rm def}{=} f(x) - t\langle \nabla f(x_0), x\rangle \right\}, \quad 0 \le t \le 1. \tag{3.2}$$

Let us present here a part of the theory of path-following schemes, which is relevant for general self-concordant functions (see Sections 4.2.4 and 4.2.5 in [4]). Note that the full justification of these methods is done only for self-concordant barriers.

Let us introduce two constants

$$\beta \;=\; 0.026, \quad \gamma \;=\; 0.1125 < \tfrac{\sqrt{\beta}}{1+\sqrt{\beta}} - \beta. \tag{3.3}$$

We say that point x satisfies an *approximate centering condition* if

$$\lambda_{f_t}(x) \;\equiv\; \|\nabla f(x) - t\nabla f(x_0)\|_x^* \;\le\; \tfrac{\beta}{M_f}. \tag{3.4}$$

Consider the following iterate:

$$(t_+, x_+) \;=\; \mathcal{P}(t,x) \;\equiv\; \left\{ \begin{array}{rcl} t_+ & = & t - \frac{\gamma}{M_f\|\nabla f(x_0)\|_x^*}, \\ & & \\ x_+ & = & x - [\nabla^2 f(x)]^{-1}(\nabla f(x) - t_+\nabla f(x_0)). \end{array} \right. \tag{3.5}$$

The following statement is just Theorem 4.2.8 from [4], but we prove it for the sake of completeness.

Theorem 2 *If the pair (x,t) satisfies (3.4), and β, γ are chosen such that*

$$|\gamma| \leq \frac{\sqrt{\beta}}{1+\sqrt{\beta}} - \beta \tag{3.6}$$

then the pair (x_+, t_+) satisfies (3.4) too.

Proof:
Let us denote $\lambda_0 = \|\nabla f(x) - t\nabla f(x_0)\|_x^*$, $\lambda_1 = \|\nabla f(x) - t_+\nabla f(x_0)\|_x^*$ and $\lambda_+ = \|\nabla f(x_+) - t\nabla f(x_0)\|_x^*$. Then $\lambda_0 \leq \frac{\beta}{M_f}$ and

$$\lambda_1 = \left\| \nabla f(x) - t\nabla f(x_0) + \frac{\gamma}{M_f \|\nabla f(x_0)\|_x^*} \nabla f(x_0) \right\|_x^* \leq \frac{\beta + |\gamma|}{M_f}.$$

Since x_+ is obtained from x as a step of Newton Method for the function $f_{t_+}(x)$, by Theorem 4.1.12 from [4],

$$\lambda_+ \leq M_f \left(\frac{\lambda_1}{1 - M_f \lambda_1} \right)^2.$$

The statement of the theorem follows from the fact that inequality $M_f \left(\frac{\lambda_1}{1-M_f\lambda_1} \right)^2 \leq \frac{\beta}{M_f}$ is equivalent to inequality $\lambda_1 \leq \frac{1}{M_f} \frac{\sqrt{\beta}}{1+\sqrt{\beta}}$. □

Let us derive from this fact a complexity bounds of the path-following scheme as applied to the problem (2.13).

Theorem 3 *Consider the following process:*

$$t_0 = 1,\ x_0 \in \mathbb{E}, \quad (t_{k+1}, x_{k+1}) \ = \ \mathcal{P}(t_k, x_k), \quad k \geq 0. \tag{3.7}$$

Assume that $\lambda(x_k) \geq \frac{1}{2M_f}$ for all $k = 0, \ldots, N$. Then

$$t_N \ \leq \ \exp\left\{ -\frac{\gamma(\gamma - 2\beta)N^2}{2M_f^2 (f(x_0) - f^*)} \right\}. \tag{3.8}$$

Proof:
Denote $c = -\nabla f(x_0)$. Then

$$t_{k+1} \ \overset{(3.5)}{=} \ t_k - \frac{\gamma}{M_f \|c\|_{x_k}^*} \ = \ t_k \left(1 - \frac{\gamma}{M_f t_k \|c\|_{x_k}^*} \right) \ \leq \ t_k \exp\left\{ -\frac{\gamma}{M_f t_k \|c\|_{x_k}^*} \right\}.$$

Thus, $t_N \leq \exp\left\{ -\frac{\gamma}{M_f} S_N \right\}$, where $S_N = \sum\limits_{k=0}^{N} \frac{1}{t_k \|c\|_{x_k}^*}$.

Let us estimate the value S_N from below. Note that

$$x_k - x_{k+1} \ \overset{(3.5)}{=} \ [\nabla^2 f(x_k)]^{-1} \left(t_k c + \nabla f(x_k) - \frac{\gamma c}{M_f \|c\|_{x_k}^*} \right). \tag{3.9}$$

Therefore,

$$r_k \ \overset{\text{def}}{=} \ \|x_k - x_{k+1}\|_{x_k} \ \overset{(3.4)}{\leq} \ \frac{\beta + \gamma}{M_f}. \tag{3.10}$$

On the other hand, $\frac{\beta^2}{M_f^2} \stackrel{(3.4)}{\geq} \lambda_f^2(x_k) + 2t_k\langle \nabla f(x_k), [\nabla^2 f(x_k)]^{-1}c\rangle + t_k^2(\|c\|_{x_k}^*)^2$. Hence,

$$-\langle \nabla f(x_k), [\nabla^2 f(x_k)]^{-1}c\rangle \;\; \geq \;\; \frac{1}{2t_k}\left[\lambda_f^2(x_k) + t_k^2(\|c\|_{x_k}^*)^2 - \frac{\beta^2}{M_f^2}\right]. \tag{3.11}$$

Therefore, denoting $\lambda_k = \|\nabla f(x_k) - t_k \nabla f(x_0)\|_{x_k}^*$,

$$\begin{aligned}
f(x_k) - f(x_{k+1}) &\stackrel{(2.5)}{\geq} \langle \nabla f(x_k), x_k - x_{k+1}\rangle - \tfrac{1}{M_f^2}\omega_*(M_f r_k) \\
&\stackrel{(3.9)}{=} \langle \nabla f(x_k), [\nabla^2 f(x_k)]^{-1}\left(t_k c + \nabla f(x_k) - \tfrac{\gamma c}{M_f\|c\|_{x_k}^*}\right)\rangle - \tfrac{1}{M_f^2}\omega_*(M_f r_k) \\
&= \lambda_k^2 - t_k\langle c, [\nabla^2 f(x_k)]^{-1}(t_k c + \nabla f(x_k))\rangle + \langle \nabla f(x_k), [\nabla^2 f(x_k)]^{-1}\left(\tfrac{-\gamma c}{M_f\|c\|_{x_k}^*}\right)\rangle - \tfrac{1}{M_f^2}\omega_*(M_f r_k) \\
&\geq \lambda_k^2 - t_k\|c\|_{x_k}^*\lambda_k - \tfrac{\gamma}{M_f\|c\|_{x_k}^*}\langle \nabla f(x_k), [\nabla^2 f(x_k)]^{-1}c\rangle - \tfrac{1}{M_f^2}\omega_*(M_f r_k) \\
&\stackrel{(3.11)}{\geq} \lambda_k^2 - t_k\|c\|_{x_k}^*\lambda_k + \tfrac{\gamma}{2M_f t_k\|c\|_{x_k}^*}\left[\lambda_f^2(x_k) + t_k^2(\|c\|_{x_k}^*)^2 - \tfrac{\beta^2}{M_f^2}\right] \\
&\stackrel{(3.10)}{\geq} \tfrac{\gamma - 2M_f\lambda_k}{2M_f}t_k\|c\|_{x_k}^* + \rho_k \stackrel{(3.4)}{\geq} \tfrac{\gamma-2\beta}{2M_f}t_k\|c\|_{x_k}^* + \rho_k,
\end{aligned} \tag{3.12}$$

where $\rho_k = \frac{\gamma}{2M_f t_k\|c\|_{x_k}^*}\left[\lambda_f^2(x_k) - \frac{\beta^2}{M_f^2}\right] - \frac{1}{M_f^2}\omega_*(\beta+\gamma)$.

Our next goal is to show that $\rho_k \geq 0$. Note that $t_k\|c\|_{x_k}^* \stackrel{(3.4)}{\leq} \lambda_f(x_k) + \frac{\beta}{M_f}$. Since $\lambda_f(x_k) \geq \frac{1}{2M_f}$, we have

$$\rho_k \;\; \geq \;\; \tfrac{\gamma}{2M_f}\left[\lambda_f(x_k) - \tfrac{\beta}{M_f}\right] - \tfrac{1}{M_f^2}\omega_*(\beta+\gamma) \geq \tfrac{\gamma(1-2\beta)}{4M_f^2} - \tfrac{1}{M_f^2}\omega_*(\beta+\gamma)$$

Using the values (3.3), by direct computation we can see that the right-hand side of this inequality is positive.

Thus, we have proved that $f(x_k) - f(x_{k+1}) \geq \frac{\gamma-2\beta}{2M_f}t_k\|c\|_{x_k}^*$. Therefore,

$$\begin{aligned}
S_N &\geq \sum_{k=0}^{N} \tfrac{\gamma-2\beta}{2M_f(f(x_k)-f(x_{k+1}))} \\
&\geq \tfrac{\gamma-2\beta}{2M_f}\min_{\tau\in\mathbb{R}_+^{N+1}}\left\{\sum_{i=1}^{N+1}\tfrac{1}{\tau^{(i)}} : \sum_{i=1}^{N+1}\tau^{(i)} = f(x_0) - f(x_{N+1})\right\} \\
&= \tfrac{(\gamma-2\beta)(N+1)^2}{2M_f(f(x_0)-f(x_{N+1}))}. \qquad \square
\end{aligned}$$

Let us estimate now the number of iterations, which are necessary for method (3.7) to enter the region of quadratic convergence $\mathbb{Q}$. Denote

$$D = \max_{x,y\in\mathrm{dom} f}\{\|x-y\|_{x_0} : \; f(x) \leq f(x_0),\; f(y) \leq f(x_0)\}.$$

Theorem 4 *Let sequence $\{x_k\}_{k\geq 0}$ be generated by the method (3.7). Then for all*

$$N \;\geq\; \left[\frac{2\Delta(x_0)}{\gamma(\gamma-2\beta)}\ln\frac{M_f D\omega^{-1}(\Delta(x_0))}{\omega\left(\frac{(1-\beta)(1-2\beta)}{2}\right)}\right]^{1/2} \tag{3.13}$$

we have $x_N \in \mathbb{Q}$.

Proof:
Indeed,

$$f(x(t_k)) - f^* \leq \langle \nabla f(x(t_k)), x(t_k) - x^*\rangle \overset{(3.1)}{=} t_k\langle \nabla f(x_0), x(t_k) - x^*\rangle \leq t_k\lambda_f(x_0)D,$$

where we used that $f(x_{k+1}) \leq f(x_k)$, $k \geq 0$, see (3.12). Since $\omega(M_f\lambda_f(x_0)) \overset{(2.10)}{\leq} M_f^2(f(x_0) - f^*)$, we have

$$\tfrac{1}{M_f^2}\omega(M_f\lambda_f(x(t_k))) \overset{(2.10)}{\leq} f(x(t_k)) - f^* \;\leq\; \tfrac{t_k}{M_f}\omega^{-1}(\Delta(x_0))D.$$

Note that $\|\nabla f(x_k) - \nabla f(x(t_k))\|^*_{x_k} \overset{(3.1)}{=} \|\nabla f(x_k) - t_k\nabla f(x_0)\|^*_{x_k} \overset{(3.4)}{\leq} \frac{\beta}{M_f} < \frac{1}{M_f}$. Therefore,

$$\begin{aligned}\lambda_f(x_k) &\overset{(3.4)}{\leq} t_k\|\nabla f(x_0)\|^*_{x_k} + \tfrac{\beta}{M_f} \overset{(3.1)}{=} \langle \nabla f(x(t_k)), [\nabla^2 f(x_k)]^{-1}\nabla f(x(t_k))\rangle^{1/2} + \tfrac{\beta}{M_f}\\ &\overset{(2.7)}{\leq} \tfrac{1}{1-\beta}\lambda_f(x(t_k)) + \tfrac{\beta}{M_f}.\end{aligned}$$

Thus, inclusion $x_k \in \mathbb{Q}$, is ensured by inequality $\lambda_f(x(t_k)) \leq \frac{(1-\beta)(1-2\beta)}{2M_f}$. Consequently, we need

$$\tfrac{t_k}{M_f}\omega^{-1}(\Delta(x_0))D \;\leq\; \tfrac{1}{M_f^2}\omega\left(\tfrac{(1-\beta)(1-2\beta)}{2}\right)$$

It remains to use inequality (3.8). □

As we can see from the estimate (3.13), up to a logarithmic factor, the number of iterations of the path-following scheme is proportional to $\Delta^{1/2}(x_0)$. This is much better than the guarantee (2.14) for the Damped Newton Method (2.12). However, as we will see in Section 5, for some special subclasses of self-concordant functions the performance estimate (3.13) can be significantly improved.

Note that, in the complexity bound (3.13), the constant $\left[\frac{2}{\gamma(\gamma-2\beta)}\right]^{1/2} \leq 17.1$. The choice of the parameters β and γ is governed by the following aspects. First, from Theorem 2, these parameters should satisfy (3.6). Second, ρ_k in the proof of Theorem 3 should be non-negative. Third, the complexity in (3.13) is proportional to $(\gamma(\gamma-2\beta))^{-1/2}$, which is desired to be as small as it is possible. This motivates the following maximization problem for optimal choice of β and γ.

$$\max \gamma(\gamma - 2\beta) \;\text{ s.t.} \tag{3.14}$$

$$\frac{\sqrt{\beta}}{1+\sqrt{\beta}} - \beta - \gamma \geq 0 \tag{3.15}$$

$$\frac{\gamma(1-2\beta)}{4} - \omega_*(\beta+\gamma) \geq 0. \tag{3.16}$$

Figure 1 illustrates this optimization problem and the optimal objective value

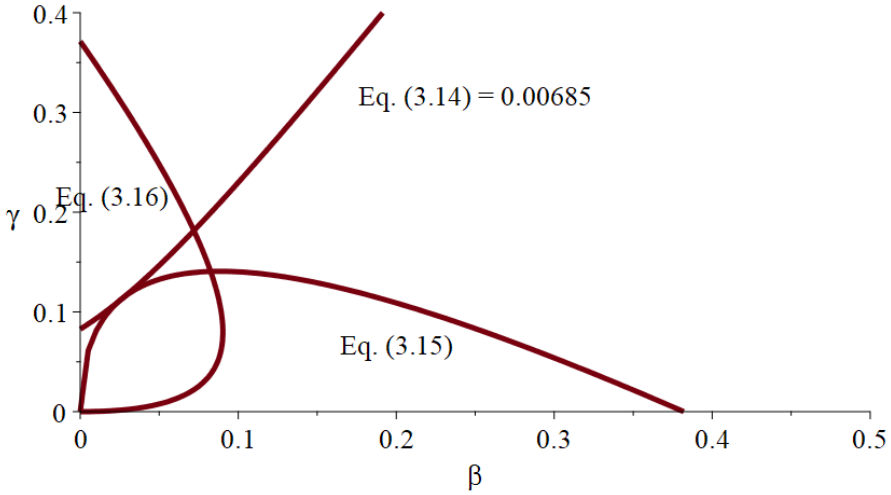


Figure 1: Optimal choice of β and γ.

4 Implementation details

The results of previous section prescribe specific values for the accuracy of following the central path β and stepsize γ. Nevertheless, if we use a larger stepsize γ and after the Newton step the approximate centering condition holds, we can continue to follow the path. This leads to an adaptive path-following scheme.

Adaptive Path-Following Scheme

- Set initial point x_0, initial value of the penalty parameter $t_0 = 1$, initial stepsize value $\gamma_{-1} \leq 0.1125$.
- Iteration $k \geq 0$. Find the minimum value $i_k \geq 0$ s.t.

$$\begin{aligned} t_+ &= t - \tfrac{2^{1-i_k}\gamma_{k-1}}{M_f \|\nabla f(x_0)\|_x^*}, \\ x_+ &= x_k - [\nabla^2 f(x_k)]^{-1}(\nabla f(x_k) - t_+ \nabla f(x_0)). \end{aligned} \tag{4.1}$$

  satisfy approximate centering condition

$$\|\nabla f(x_+) - t_+ \nabla f(x_0)\|_{x_+}^* \leq \frac{\beta}{M_f}.$$

- Set $\gamma_k = 2^{1-i_k}\gamma_{k-1}$, $x_{k+1} = x_+$, $t_{k+1} = t_+$.

By the results of the previous section, there exists $\hat{\gamma}$ s.t. the Newton step with the stepsize $\hat{\gamma}$ outputs a point satisfying approximate centering condition. Hence, the search for i_k is finite and $\gamma_k = 2^{1-i_k}\gamma_{k-1} \geq \frac{\hat{\gamma}}{2}$. Hence, the number of Newton steps can be estimated as follows

$$\sum_{j=0}^{k} i_k = \sum_{j=0}^{k} \left(1 + \log_2 \frac{\gamma_{j-1}}{\gamma_j}\right) = k + 1 + \log_2 \frac{\gamma_{-1}}{\gamma_k} = k + \log_2 \frac{2\gamma_{-1}}{\hat{\gamma}}.$$

As we see, the price for the adaptivity is reasonable taking into account that the practical performance of the adaptive algorithm is better since the penalty parameter t grows faster.

Damped Newton method (2.9) has also an adaptive stepsize extension. Since the Damped Newton method is obtained by minimization of (2.5), the adaptive choice of the stepsize is based on this inequality and we obtain the adaptive Damped Newton method as follows.

Adaptive Damped Newton Method

- Set initial value of the stepsize τ_{-1} and initial point x_0.
- k-th iteration. Find the minimum value $i_k \geq 0$ s.t.

$$x_{k+1} = x_k - \frac{2^{1-i_k}\tau_{k-1}}{1 + M_f\lambda_f(x_k)}[\nabla^2 f(x_k)]^{-1}\nabla f(x_k) \tag{4.2}$$

  satisfies

$$f(x_{k+1}) \leq f(x_k) - \frac{2^{1-i_k}\tau_{k-1}(\lambda(x_k))^2}{1 + M_f\lambda_f(x_k)} + \omega_*\left(\frac{2^{1-i_k}\tau_{k-1}\lambda(x_k)}{1 + M_f\lambda_f(x_k)}\right).$$

- Set $\tau_k = 2^{1-i_k}\tau_{k-1}$.

The overhead for the adaptivity can be estimated similarly to the path-following scheme.

5 Minimizing strongly convex functions

Let $B = B^* \succ 0$ maps $\mathbb{E}$ to $\mathbb{E}^*$. Define Euclidean metric

$$\|x\|^2 \quad = \quad \langle Bx, x\rangle^{1/2}, \quad x \in \mathbb{E}.$$

In this section, we consider the following minimization problem

$$\min_{x\in\mathbb{E}} f(x), \tag{5.1}$$

where f is a strongly convex function:

$$f(y) \quad \geq \quad f(x) + \langle \nabla f(x), y - x\rangle + \tfrac{1}{2}\sigma_f\|y - x\|^2, \quad x, y \in \mathbb{E}, \tag{5.2}$$

where $\sigma_f > 0$. We also assume that function f belongs to $\mathbb{C}^3(\mathbb{E})$ and its Hessian is Lipschitz continuous:

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq H_f \|x - y\|, \quad x, y \in \mathbb{E}. \tag{5.3}$$

Lemma 2 *Under assumptions above, function f is self-concordant with*

$$M_f = \frac{H_f}{2\sigma_f^{3/2}}. \tag{5.4}$$

Proof:
Indeed, for any point $x \in \mathbb{E}$ and direction $h \in \mathbb{E}$ we have

$$D^3 f(x)[h]^3 \overset{(5.3)}{\leq} H_f \left[\|h\|^2\right]^{3/2} \overset{(5.2)}{\leq} H_f \left[\frac{1}{\sigma_f}\langle \nabla^2 f(x)h, h\rangle\right]^{3/2}.$$

It remains to use definition (2.1). □

Thus, problem (5.1) can be solved by methods (2.12) and (3.7). The corresponding complexity bounds can be given in terms of the complexity measure

$$\Delta(x_0) = \frac{H_f^2}{\sigma_f^3}(f(x_0) - f^*).$$

As we have seen, the first method needs $O(\Delta(x_0))$ iterations. Complexity bound of the second method is of the order $\widetilde{O}(\Delta^{1/2}(x_0))$. Let us show that for our particular subclass of self-concordant functions these bounds can be significantly improved.

Our methods are based on the *cubic regularization* of the Newton method. Let us define quadratic approximation of f at point $x \in \mathbb{E}$:

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

Then

$$|f(y) - Q(x, y)| \leq \frac{H_f}{6}\|y - x\|^3, \quad y \in \mathbb{E}. \tag{5.5}$$

This inequality justifies the cubic Newton step:

$$T_M(x) = \arg\min_{y \in \mathbb{E}} \{Q(x, y) + \tfrac{1}{6} M \|y - x\|^3\}. \tag{5.6}$$

As it was shown in [7], the method iterating these steps converges for functions with Lipschitz continuous Hessian as $O(\frac{1}{k^2})$, where k is the iteration counter.

Define the region of quadratic convergence of the Cubic Newton Method in terms of the function value:

$$\mathbb{Q}_f = \left\{x \in \mathbb{E} : \; f(x) - f^* \leq \frac{\sigma_f^3}{2H_f^2} \overset{(5.4)}{=} \frac{1}{8M_f^2}\right\}$$

(see (6.4) in [5]). Let us check how many iterations we need for entering this region by different schemes based on the cubic Newton step. Assume our method has the following rate of convergence:

$$\begin{aligned} f(x_k) - f^* &\leq \frac{cH_f\|x_0 - x^*\|^3}{k^p} \overset{(5.2)}{\leq} \frac{cH_f}{k^p}\left(\frac{2}{\sigma_f}(f(x_0) - f^*)\right)^{3/2} \\ &\overset{(5.4)}{=} \frac{2^{5/2} c M_f}{k^p}(f(x_0) - f^*)^{3/2}, \end{aligned} \tag{5.7}$$

where c is an absolute constant and $p > 0$. Thus, we need

$$O\left(\left[M_f^3(f(x_0) - f^*)^{3/2}\right]^{1/p}\right) \;=\; O\left(\Delta^{\frac{3}{2p}}(x_0)\right)$$

iterations for entering the region of superlinear convergence $\mathbb{Q}_f$. For Cubic Newton method we have $p = 2$ (see [7]). Thus, it ensures complexity $O(\Delta^{3/4}(x_0))$. For the accelerated Cubic Newton method [5] we have $p = 3$. Thus, it needs $O(\Delta^{1/2}(x_0))$ iterations (which is slightly better than (3.13)). However, note that there exists a powerful tool for accelerating these schemes, the *restarting strategy*.

Let us define k_p as the first integer, for which the right-hand side of inequality (5.7) is smaller than $\frac{1}{2}(f(x_0) - f^*)$:

$$\frac{2^{5/2}cM_f}{k^p}(f(x_0) - f^*)^{3/2} \;\leq\; \tfrac{1}{2}(f(x_0) - f^*).$$

Clearly $k_p = O\left(\left[M_f(f(x_0) - f^*)^{1/2}\right]^{1/p}\right) = O\left(\Delta^{\frac{1}{2p}}(x_0)\right)$.

This value can be used in the following multi-stage scheme.

Multi-stage Acceleration Scheme (5.8)

At the first stage, we perform $t_1 = \lceil k_p \rceil$ iterations of our method starting from the point $y_0 = x_0$. It generates the point y_1, which is the starting point for the next stage. Its length is $\left\lceil \frac{k_p}{2^{1/p}} \right\rceil$, etc.

In general, kth stage starts from the point y_{k-1} and its length is $t_k = \left\lceil \frac{k_p}{2^{(k-1)/(2p)}} \right\rceil$. Method stops when $y_k \in \mathbb{Q}_f$.

Theorem 5 *The total number of stages T in the optimizations strategy (5.8) satisfies inequality*

$$T \;\leq\; 4 + \log_2 \Delta(x_0). \tag{5.9}$$

The total number of the lower-level iterations N in this scheme does not exceed

$$4 + \log_2 \Delta(x_0) + \tfrac{2^{1/(2p)}}{2^{1/(2p)}-1} k_p.$$

Proof:

Let us prove by induction that $f(y_k) - f^* \leq (\frac{1}{2})^k (f(y_0) - f^*)$. For $k = 0$ this is true. Assume that this is also true for some $k \geq 0$. Note that $t_{k+1}^p \geq (\frac{1}{2})^{k/2} k_p^p$. Therefore,

$$\tfrac{f(y_{k+1})-f^*}{f(y_k)-f^*} \;\leq \tfrac{2^{5/2}cM_f}{t_{k+1}^p}(f(y_k) - f^*)^{1/2} \;\leq\; \tfrac{k_p^p (f(y_k)-f^*)^{1/2}}{2t_{k+1}^p (f(x_0)-f^*)^{1/2}} \;\leq \tfrac{1}{2}\left[\tfrac{2^k (f(y_k)-f^*)}{f(x_0)-f^*}\right]^{1/2} \;\leq\; \tfrac{1}{2}.$$

Thus, the total number of stages satisfies inequality $\left(\frac{1}{2}\right)^{T-1}(f(x_0) - f^*) \geq \frac{1}{8M_f^2}$. Finally,

$$N \;=\; \sum_{k=1}^{T} t_k \;\leq\; T + k_p \sum_{k=0}^{T-1} \left(\tfrac{1}{2}\right)^{\frac{k}{2p}} \;\leq\; T + k_p \sum_{k=0}^{\infty} \left(\tfrac{1}{2}\right)^{\frac{k}{2p}} \;=\; T + \frac{k_p}{1-\left(\frac{1}{2}\right)^{1/(2p)}}.$$

□

Applying Theorem 5 to different second-order methods based on the Cubic Regularization, we get the following complexity bounds.

- **Cubic Newton Method [7].** For this method $p = 2$. Therefore, the complexity bound of this scheme, as applied in the framework of multi-stage method (5.8) is of the order $O(\Delta^{1/4}(x_0))$. In fact, this method does not need a restarting strategy. Thus, Theorem 5 provides the Cubic Newton with a better way of estimating its rate of convergence.
- **Accelerated Newton Method [5].** For this method $p = 3$. Hence, the complexity bound of the corresponding multi-stage scheme (5.8) becomes $O(\Delta^{1/6}(x_0))$.
- **Optimal second-order method [3].** For this method $p = 3.5$. Therefore, the corresponding complexity bound is $\tilde{O}(\Delta^{1/7}(x_0))$. However, this method includes an expensive line-search procedure. Consequently, its practical efficiency should be worse than the efficiency of the method from the previous item. Note that the theoretical gap in the complexity estimates of these methods is negligibly small, of the order of $O(\Delta^{1/42}(x_0))$.

Remark 1 *For the restarting procedure, the knowledge of f^* is not necessary. Indeed, if we know only a lower bound $\tilde{f}$ such that $f^* \geq \tilde{f}$, we obtain that after*

$$\tilde{k}_p = O\left(\left[M_f(f(x_0) - \tilde{f})^{1/2}\right]^{1/p}\right)$$

steps of the inner method the right-hand side of inequality (5.7) is smaller than $\frac{1}{2}(f(x_0) - f^*)$ since $\tilde{k}_p \geq k_p$. Then, the overall number of lower-level iterations is of the order of $\tilde{k}_p = O\left(\tilde{\Delta}(x_0)\right)^{\frac{1}{2p}}$, where $\tilde{\Delta}(x_0) = M_f^2(f(x_0) - \tilde{f})$.

As we can see, the methods considered in this section have much better complexity bounds for problem (5.1) than the the methods based on the framework of self-concordant functions. A possible explanation of this phenomena is that these methods use a more precise model of the objective function, which is based on two independent inequalities (5.2) and (5.3) instead of single inequality (2.1). Nevertheless, unlike Damped Newton Method and Path-Following Scheme, the methods in this section are not applicable for a wide class of self-concordant functions - self-concordant barriers, a typical example being log-barrier $-\ln x$. The reason is that this function neither strongly convex, nor have Lipshitz-continuous Hessian.

6 Computational experiments

In the experiments, we consider two problems, the first one being the regularized logistic regression and the second being the dual problem to a feasibility problem on the cube.

6.1 Regularized logistic regression

Regularized logistic regression is the following problem

$$\min_{x\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \ln\left(1+\exp(-\langle a_i, x\rangle)\right) + \frac{\kappa}{2}\|x\|_2^2,$$

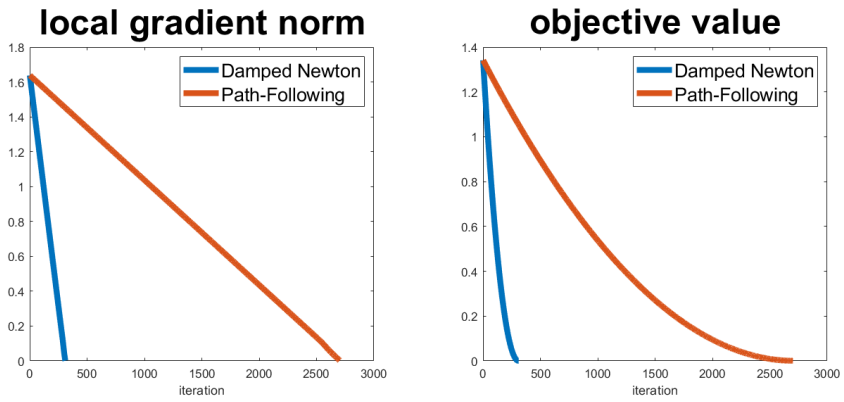
where $a_i \in \mathbb{R}^m$, $i = 1, ..., n$. As it is shown in [8], this function is self-comcordant with $M_f = \frac{\max\{\|a_i\|_2,\, i=1,...,n\}}{2\sqrt{\kappa}}$. We use different values of regularization parameter, namely $\kappa \in \{10^{-1}, 10^{-1}\}$ and random starting point. We compare the Damped Newton method (2.9) and the path-following scheme (3.7) using the binary classification dataset downloaded from [9] at `https://www.csie.ntu.edu.tw/~cjlin/libsvm/`. We also compare the adaptive versions (4.2) of the Damped Newton method and (4.1) of the path-following scheme. Figure 2 illustrates convergence of these two methods. As we can see, adaptive choice of stepsize accelerates path-following scheme so that it becomes much faster than the Damped Newton Method. Interestingly, the value of γ_k in the experiments was up to 15, whereas the value of τ_k did not exceed 2. Table 1 contains the results of experiments with fixed stepsize and different random initial points and different values of the regularization parameter κ. As we can see the theoretical estimate for the number of iterations to enter the region of quadratic convergence is much larger for Damped Newton Method. At the same time, both methods require much smaller number of iterations than in theory and Damped Newton Method is faster than the path-following scheme. The situation is opposite for the case of adaptive choice of the stepsize, see Table 2.

κ	10^{-1}	10^{-1}	10^{-1}	10^{-1}		10^{-4}	10^{-4}	10^{-4}	10^{-4}
DN (experiment)	575	328	998	388		294	866	76	806
DN (theory)	1708582	547205	5189850	769250		443171	3918329	27776	3388730
PF (experiment)	5044	2851	8797	3382		2552	7630	626	7094
PF (theory)	31464	17061	56987	20497		15222	49047	3360	45386

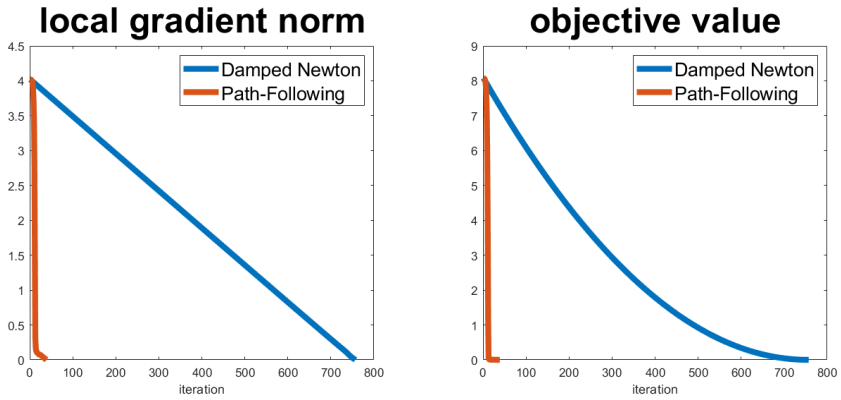
Table 1: Number of iterations until $\lambda_f(x_k) \leq \frac{1}{2M_f}$ in different runs of the experiment, fixed stepsize.

κ	10^{-1}	10^{-1}	10^{-1}	10^{-1}		10^{-4}	10^{-4}	10^{-4}	10^{-4}
DN (experiment)	203	653	74	902		353	67	570	97
DN (theory)	207792	2215056	26092	4240539		645250	22004	1689036	46643
PF (experiment)	23	27	21	29		36	30	37	34
PF (theory)	10106	36158	3249	51164		18642	2954	31267	4470

Table 2: Number of iterations until $\lambda_f(x_k) \leq \frac{1}{2M_f}$ in different runs of the experiment, adaptive stepsize.



(a) Methods with fixed stepsize



(b) Methods with adaptive stepsize

Figure 2: Convergence of Damped Newton Method and path-following scheme

6.2 Dual feasibility problem

In this subsection we consider the following problem

$$\text{Find } x \text{ s.t. } \|x\|_\infty \leq 1 \text{ and } Ax = b,$$

where $x \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, $A \in \mathbb{R}^{m\times n}$. To solve this problem, we introduce a barrier for the set $\|x\|_\infty \leq 1$ and minimize it over the affine manifold given by linear constraints $Ax = b$

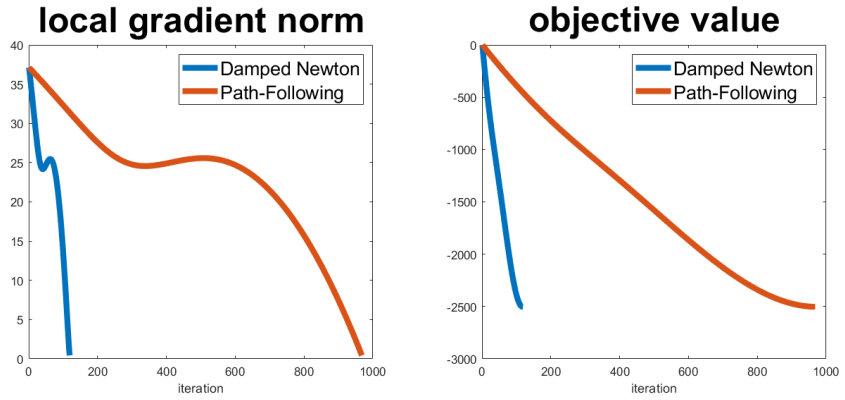
$$\min \sum_{i=1}^n \psi(x_i) \text{ s.t. } Ax = b,$$

where $\psi(t) := (-|t| - \ln(1 - |t|))$. Introducing Lagrange multiplier $y \in \mathbb{R}^m$, we construct the dual problem

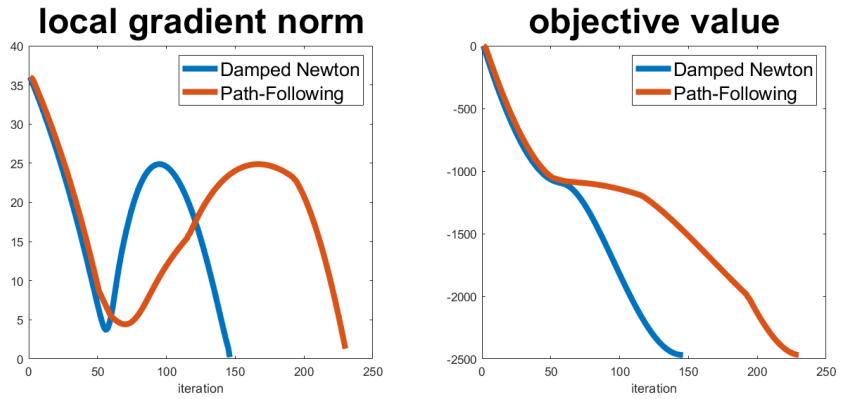
$$\begin{aligned} & \min_x \max_y \textstyle\sum_{i=1}^n \psi(x_i) + \langle y, Ax - b\rangle \\ = \; & \max_y \left\{-\langle b, y\rangle + \min_x \left\{\textstyle\sum_{i=1}^n \psi(x_i) + \langle A^T y, x\rangle\right\}\right\} \\ = \; & -\langle b, y\rangle - \textstyle\sum_{i=1}^n \psi^*(a_i^T y), \end{aligned}$$

where $\psi^*(\tau) = |\tau| - \ln(1 + |\tau|)$ is the Fenchel conjugate for $\psi(t)$, $a_i \in \mathbb{R}^m$, $i = 1, ..., n$ are columns of A. Formally, $\psi^*(\tau)$ is not three times continuously differentiable. Nevertheless, repeating the same arguments as in Section 4.1.4 in [4], we obtain that inequalities (2.4) – (2.7) hold and we can apply the whole theory described above.

Similarly to the previous subsection, we compare the Damped Newton method (2.9) and the path-following scheme (3.7) and their adaptive versions (4.2) and (4.1) respectively. The data A, b was generated randomly and the starting point was chosen to be zero. Figure 3 illustrates convergence of these four methods. Unlike logistic regression, adaptivity does not lead to a large gain for this problem. Table 3 contains the results of experiments with fixed stepsize and different random data for different problem sizes m, n. As we can see the theoretical estimate for the number of iterations to enter the region of quadratic convergence is much larger for Damped Newton Method. At the same time, both methods require much smaller number of iterations than in theory and Damped Newton Method is faster than the path-following scheme. Table 4 illustrates the behavior of the adaptive counterparts of these methods.



(a) Methods with fixed stepsize



(b) Methods with adaptive stepsize

Figure 3: Convergence of Damped Newton Method and path-following scheme

	$n = 1000, m = 100$					$n = 5000, m = 1000$			
DN (experiment)	64	69	67	65		121	118	123	121
DN (theory)	5029	5740	5425	5435		26135	25777	26546	26796
PF (experiment)	463	504	490	478		978	955	997	982
PF (theory)	1582	1700	1650	1647		3671	3653	3717	3724

Table 3: Number of iterations until $\lambda_f(x_k) \leq \frac{1}{2M_f}$ in different runs of the experiment, fixed stepsize.

	$n = 1000, m = 100$					$n = 5000, m = 1000$			
DN (experiment)	66	65	65	65		124	118	141	123
DN (theory)	5415	5190	5359	5222		25722	26442	26990	26398
PF (experiment)	147	145	143	145		180	174	219	184
PF (theory)	1647	1611	1636	1614		3774	3661	4059	3763

Table 4: Number of iterations until $\lambda_f(x_k) \leq \frac{1}{2M_f}$ in different runs of the experiment, adaptive stepsize.

References

[1] Conn, A., Gould, N., Toint, P. Trust Region Methods, SIAM, 2000

[2] Kantorovich, L.V. On Newtons method for functional equations, Dokl. Akad. Nauk SSSR, 59(7), 1948

[3] Monteiro, R. and Svaiter, B. An Accelerated Hybrid Proximal Extragradient Method for Convex Optimization and Its Implications to Second-Order Methods, SIAM Journal on Optimization, 23(2), 2013

[4] Nesterov, Yu. Introductory Lectures on Convex Optimization: a basic course, Kluwer Academic Publishers, Massachusetts, 2004

[5] Nesterov, Yu. Accelerating the cubic regularization of Newton's method on convex problems, Mathematical Programming, 112(1), 2008

[6] Nesterov, Yu., Nemirovskii, A. Interior-point polynomial algorithms in convex programming, SIAM, 1994

[7] Nesterov, Yu., Polyak, B. Cubic regularization of Newton method and its global performance, Mathematical Programming, 108(1), 2006

[8] Sun, T., Tran-Dinh, Q. Generalized self-concordant functions: a recipe for Newton-type methods. Mathematical Programming, 2018. `https://link.springer.com/article/10.1007/s10107-018-1282-4`

[9] Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines. `https://dl.acm.org/citation.cfm?id=1961199`