New versions of Newton method: step-size choice, convergence domain and under-determined equations

Article *in* Optimization Methods and Software · October 2019

DOI: 10.1080/10556788.2019.1669154

2 authors, including:

Boris T. Polyak
Institute of Control Sciences
266 PUBLICATIONS **9,543** CITATIONS

Optimization Methods and Software

ISSN: 1055-6788 (Print) 1029-4937 (Online) Journal homepage: https://www.tandfonline.com/loi/goms20

New versions of Newton method: step-size choice, convergence domain and under-determined equations

Boris Polyak & Andrey Tremba

To cite this article: Boris Polyak & Andrey Tremba (2019): New versions of Newton method: step-size choice, convergence domain and under-determined equations, Optimization Methods and Software, DOI: 10.1080/10556788.2019.1669154

To link to this article: https://doi.org/10.1080/10556788.2019.1669154

Published online: 02 Oct 2019.

OPTIMIZATION METHODS & SOFTWARE
https://doi.org/10.1080/10556788.2019.1669154

Taylor & Francis
Taylor & Francis Group

New versions of Newton method: step-size choice, convergence domain and under-determined equations

Boris Polyak and Andrey Tremba

Institute for Control Sciences, Moscow, Russia

ABSTRACT

Newton method is one of the most powerful methods for finding solutions of nonlinear equations and for proving their existence. In its 'pure' form it has fast convergence near the solution, but small convergence domain. On the other hand damped Newton method has slower convergence rate, but weaker conditions on the initial point. We provide new versions of Newton-like algorithms, resulting in combinations of Newton and damped Newton method with special step-size choice, and estimate its convergence domain. Under some assumptions the convergence is global. Explicit complexity results are also addressed. The adaptive version of the algorithm (with no a priori constants knowledge) is presented. The method is applicable for under-determined equations (with $m < n$, m being the number of equations and n being the number of variables). The results are specified for systems of quadratic equations, for composite mappings and for one-dimensional equations and inequalities.

ARTICLE HISTORY
Received 22 October 2018
Accepted 14 September 2019

KEYWORDS
Nonlinear equations; Newton method; under-determined equations; global convergence; adaptive algorithms; metric regularity

2010 MATHEMATICS SUBJECT CLASSIFICATIONS
49M15; 65H10; 90C30; 58C15

1. Introduction

Consider nonlinear equation

$$P(x) = 0, \tag{1}$$

written via the vector function $P : \mathbb{R}^n \to \mathbb{R}^m$. There exists the huge bunch of literature on solvability of such equations and numerical methods for their solution, see e.g. the classical monographs [4,22]. One of the most powerful methods is *Newton method*, going back to such giants as Newton, Cauchy, Fourier. The general form of the method is due to Kantorovich [14]; on history and references see [5,15,26,33]. The basic version of Newton method for (1) is applicable when $P(x)$ is differentiable and $P'(x)$ is invertible (this implies $m = n$):

$$x^{k+1} = x^k - P'(x^k)^{-1}P(x^k). \tag{2}$$

The method converges under some natural conditions, moreover it can be used for obtaining existence theorems for the solution (see references cited above). Unfortunately Newton method converges only locally: it requires a good initial approximation x^0 (so called 'hot

CONTACT Andrey Tremba atremba@ipu.ru

start'). Convergence conditions can be relaxed for *damped Newton method*

$$x^{k+1} = x^k - \alpha P'(x^k)^{-1} P(x^k)$$

with $0 < \alpha < 1$.

The advanced, generalized writing of Newton method is

$$x^{k+1} = x^k - \alpha_k z^k, \quad k = 0, 1, \ldots$$
$$z^k \in \operatorname{Arg} \min_z \{||z|| : P'(x^k)z = P(x^k)\}. \tag{3}$$

This variant relies on the solvability of the linear equation only, and it also admits non-constant step-size. It is applicable to under-determined systems of equations ($m < n$) and to non-linear equations in Banach space. The latter are outside of the scope of this paper but its analysis is essentially the same.

If $m = n$ and $P'(x^k)^{-1}$ exists, the method (3) coincides with classical Newton method for $\alpha_k = 1$ and damped Newton method for $\alpha_k = \alpha < 1$. Starting at some initial point x^0 the latter method converges to x^* under some additional constraints on residual $||P(x^0)||$ and function-related constants L, ρ, μ, α (see Theorems below for rigorous conditions).

In explicit form Newton method for $m \neq n$ has been written by Ben-Israel [1]:

$$x^{k+1} = x^k - P'(x^k)^{\dagger} P(x^k), \tag{4}$$

where $A^{\dagger}$ stands for Moore-Penrose pseudoinverse of A. However the results in [1] are mostly oriented on over-determined systems, and the assumptions of the theorems in [1] are hard to check. Other publications on under-determined equations include [11,19,20,23,31]. Moreover there exist numerous literature on more general settings: equalities plus inequalities [28,30], optimization problems [3,10] with more general algorithms, which can be applied to solving of equations as particular case. In next Section 2 we discuss the under-determined finite-dimensional case in more details.

There is a very similar problem statement to (1), made in terms of the equation with the variable right-hand side

$$g(x) = y, \tag{5}$$

with $g : \mathbb{R}^n \to \mathbb{R}^m$ having a known solution $\overline{x} : g(\overline{x}) = 0$ and the variable y as a parameter.

The question is: for which right-hand side part y the equation is still feasible and what are the solutions? This problem arises in finding image of a mapping $\{g(x) : x \in \mathbb{R}^n\}$, checking robustness/sensitivity of a solution or exploration problem of the image, etc. In general, this problem is hardly solvable, but we can provide *local* sufficient conditions of feasibility, imposed on y.

Trivially, Equation (5) can be written in the form (1) with $P_y(x) = g(x) - y = 0$ and $\|P_y(\overline{x})\| = \|y\|$. Thus conditions on feasibility of the equation $P_y(x) = 0$ are exactly conditions on feasibility (5) with respect to right-hand side y, and vice-versa.

Let us explain the connection between (1) and (5) deeper. There are few approaches to treating feasibility of an equation. One of them is to prove existence of the solution by providing *semi-local* existence theorems. These involve conditions in some point and/or around it, and prove that if such conditions hold, then a solution exists. It is not necessary to provide tools for finding this solution.

Another way is to explore a constructive, algorithmic way of solving the equation, starting at a point x^0, resulting in a sequence $\{x^k\}$, and prove convergence to a solution $x^k \to x^*$ (e.g. fixed-point theorems). The convergence conditions are typically tied to the sequence, including starting point x^0. The conditions do not necessarily coincide with the conditions of semi-local existence theorems. Moreover, the conditions ensuring *faster* convergence of the algorithm are typically more strict, than the conditions of the semi-local existence theorems.

In Newton method theory these approaches are closely connected, and semi-local theorems are often proved via convergence of variants of Newton method. We also show it below in Theorem 3.2. This relation becomes very clear in comparison of Equations (1) and (5). Naturally feasibility of (1) being solved by a Newton-like method (i.e. *convergence* of the Newton method started at x^0!) is stated in terms of norm of initial residual, say $\|P(x^0)\| \le s$. In terms of (5) the very same Newton-like method, being applied to the constructed $P_y(\cdot)$ and being started at the same point $\overline{x} = x^0$, converges for *any* fixed $y : \|y\| \le s$. This analysis claims *feasibility* of (5) for all such y. Therefore finding the *largest* set of possible y is essentially the same as the problem of finding the *broadest* residual range $P(x^0)$. Without loss of generality, through the paper we assume $\overline{x} = 0$, and switch between problems (1) and (5) as equivalent ones. The difference is clear from context.

We also examine some special cases of the nonlinear equations. One of them is the quadratic case, when all components of g are quadratic functions:

$$g_i(x) = \tfrac{1}{2}(A_i x, x) + (b_i, x), \quad A_i = A_i^{\mathrm{T}}, \quad b_i \in \mathbb{R}^n. \tag{6}$$

In this case we try to specify above results and design the algorithms to check feasibility of a vector $y \in R^m$.

The first goal of the present paper is to give explicit expressions of the method (3) for various norms and to provide simple, easily checkable conditions for convergence of the method. This also provides existence theorems: what is *a feasible set* Y such that $y \in Y$ implies solvability of (5).

The second goal is to develop constructive algorithms for choosing step-sizes α_k to achieve fast and *as global as possible* convergence. We suggest different strategies for constructing algorithms, study their properties and provide explicit convergence conditions for the method and demonstrate its potentially global convergence.

The main contributions of the paper are threefold.

(1) We propose the novel procedure for adjusting step-size α_k. The strategy guarantees wide range of convergence (in some cases the algorithm converges globally) and fast rate of convergence (local quadratic convergence, typical for pure Newton method). Moreover explicit formula for method's complexity is provided.
(2) The choice of norms in the algorithm can be different, thus we arrive to different versions of the algorithm. For instance, Euclidean norms imply explicit form of desired direction z^k (the same as in (4)) while ℓ_1 norm provides sparse approximations etc.
(3) We consider numerous applications, including under-determined cases, quadratic equations, one equation with n variables.

Few words on comparison with known results. In the paper [23] results on solvability of nonlinear equations in Banach spaces and on application of Newton-like methods have

been formulated in semi-local form. One of the results from [23] adopted to our notation and finite-dimensional case claims that if $P'(x)$ exists and is Lipschitz on a ball B of radius ρ centred in x^0 and estimate $||P'(x)^{\mathrm{T}}h||_* \geq \mu||h||_*,\ \mu > 0, \quad \forall\, h$ holds on B, then Equation (1) has a solution x^* provided $||P(x^0)|| < \rho/\mu$. Another result deals with convergence of Newton method; however the method is not provided in explicit form. The condition on the derivative has extension to non-differentiable functions and multi-valued mapping an is known as *metric regularity*, used for proving existence theorems in different cases.

The paper, which contains the closest results to ours, is [20]. Nesterov addresses the same problem (1) and his method (in our notation) has the form

$$x^{k+1} = x^k - z^k, \quad k = 0, 1, \ldots$$
$$z^k = \arg\min_z \{||P(x^k) - P'(x^k)z|| + M||z||^2\},$$

where M is the scalar parameter to be adjusted at each iteration. Nesterov's assumptions are close to ours and his results on solvability of equations and on convergence of the method are similar. The main difference is the method itself; it is not clear how to solve the auxiliary optimization problem in Nesterov's method, while finding z^k in our method can be implemented in explicit form. Other papers on under-determined equations mentioned above either do not specify the technique for solving the linearized auxiliary equation, or restrict analysis with Euclidean norm and/or pure Newton step-size $\alpha_k = 1$, see e.g. [16,23,29,31].

The rest of the paper is organized as follows. Next section is introductory to the case of under-determined systems. In Section 3 we remind few notions and results. Next, we prove simple solvability conditions for (1). In main Section 4 we propose few variants of general Newton algorithm (3), including adaptive ones and estimate their convergence rate. Some particular cases (scalar equations and inequalities, quadratic equations, problems with special structure) are treated in Section 5. Results of numerical simulation are exhibited in Section 6. Conclusion part finalizes the paper (Section 7).

2. Under-determined systems of equations

Under-determined equations attracted our attention by specific norm-dependency property. In case of $m < n$ norms in $\mathbb{R}^n$ (in the optimization sub-problem) and $\mathbb{R}^m$ (for residual) can be chosen arbitrarily, and they imply principally different forms and results of Newton method (3). Conditions on solvability and convergence look similar, but the results differ strongly.

Historically the case of under-determined equations ($m < n$) attracted much less attention than equations with the same number of equations and variables. The pioneering result is due to Graves [9] in more general setting of Banach spaces, for problem (5). Graves' theorem for finite-dimensional case claims, that if condition

$$||g(x^a) - g(x^b) - A(x^a - x^b)|| \leq C||x^a - x^b||$$

holds in the ball of radius ρ, centred at zero, for a matrix A with minimal singular value $\mu > C > 0$, then a solution of the Equation (5) exists provided $||y||$ is small enough, namely $||y|| \leq \rho(\mu - C)$. The solution can be found via a version of modified Newton method,

where next iteration requires solution of the linear equation with matrix A, see [6,18] for details. The condition above gives rise to the mentioned metric regularity property. However, finding the matrix A is still a problem itself.

First of all let us specify the subproblem of finding a vector z^k in (3) for different norms of $x \in \mathbb{R}^n$. We skip simple verifications of the statements from convex analysis.

(1) For $||x|| = ||x||_1$ vector z^k is a solution of the problem

$$\min\{||z||_1 : P'(x^k)z = P(x^k)\}.$$

(2) For $||x|| = ||x||_\infty$ vector z^k is a solution of the problem

$$\min\{||z||_\infty : P'(x^k)z = P(x^k)\}.$$

Both problems above can be easily reduced to linear programming.

(3) For $||x|| = ||x||_2$ vector z^k can be written explicitly

$$z^k = P'(x^k)^\dagger P(x^k).$$

In this case Newton method (3) coincide with (4). For $m \leq n$ Moore-Penrose pseudo-inverse of a matrix A is written as $A^\dagger = A^{\mathrm{T}}(AA^{\mathrm{T}})^{-1}$, if A has full row rank.

Thus in these (most important) cases algorithm (3) can be implemented effectively. Also the solution of the first two problems may be non-unique.

An important case is the scalar one, i.e. $m = 1$. We specify general results for scalar equations and inequalities; the arising algorithms have much in common with unconstrained minimization methods. Finally we discuss nonlinear equations having some special structure. Then convergence results can be strongly enhanced.

3. Preliminaries and feasibility (existence) theorems

Key component in Newton method is the auxiliary convex optimization sub-problem, involving the linear constraint. Note that the constraint

$$Az = b, \quad b \in \mathbb{R}^m, \quad z \in \mathbb{R}^n \tag{7}$$

describes either a linear subspace, or the empty set. The classical result below (which goes back to Banach, see [14,18,20]) guarantees solvability of the linear Equation (7) and gives an estimate of its solution. We prefer to provide the direct proof of the result because it is highly clear and short in finite-dimensional case. Suppose that spaces $\mathbb{R}^n, \mathbb{R}^m$ are equipped with some norms, the dual norms are denoted $||\cdot||_*$ (for a linear functional c, associated with the vector of the same dimension, $\|c\|_* = \sup_{x:\|x\|=1}(c, x)$). Operator norm is subordinate with the vector norms, e.g. for $A : X \to Y$ we have $\|Ax\|_Y \leq \|A\|_{X,Y}\|x\|_X$. In most cases we do not specify vector norms; dual norms are obvious from the context. The adjoint operator A^* is identified with matrix A^{T}.

Lemma 3.1: *If $A \in \mathbb{R}^{m\times n}$ satisfies condition*

$$\|A^{\mathrm{T}}h\|_* \geq \mu_0\|h\|_*, \quad \mu_0 > 0, \tag{8}$$

for all $h \in \mathbb{R}^m$, then Equation (7) has a solution for all $b \in \mathbb{R}^m$, and all solutions of optimization problem

$$\widehat{z} \in \operatorname{Arg}\min\{\|z\| : Az = b\}$$

have bounded norms $\|\widehat{z}\| \leq \|b\|/\mu_0$.

Proof: Fix $b \in \mathbb{R}^m$ and denote $K = \{x \in \mathbb{R}^n : ||x|| \leq \|b\|/\mu_0\}$. This is a convex closed bounded set, and its linear image $Q = \{y \in \mathbb{R}^m : y = Ax, x \in K\}$ is convex closed bounded set as well. Suppose $b \notin Q$, then it can be strictly separated from Q: there exists $c \in \mathbb{R}^m : \max_{y\in Q}(c, y) < (c, b)$. But $\max_{y\in Q}(c, y) = \max_{x\in K}(c, Ax) = \max_{x\in K}(A^{\mathrm{T}}c, x) = (\|b\|/\mu_0)\|A^{\mathrm{T}}c\|_* \geq \|b\| \cdot \|c\|_*$, thus we get the contradiction: $\|b\| \cdot \|c\|_* < (c, b)$. Hence $b \in Q$, i.e. there exists $x \in \mathbb{R}^n : Ax = b, \|x\| \leq ||b||/\mu_0\}$. A solution with the least norm obeys the same inequality. ■

The Lemma is claiming that the matrix A has full row rank equal to m provided (8) holds. It is another way to say that the mapping $A : \mathbb{R}^n \to \mathbb{R}^m$ is *onto* mapping, i.e. covering all image space. In the case of Euclidean norms, parameter μ_0 is the smallest singular value of the matrix $\mu_0 = \sigma_m(A)$ (we denote singular values of a matrix in $\mathbb{R}^{m\times n}$ in decreasing order as $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_m$). In general case the conjugate operator A^* is used instead of A^{T}, and μ is the metric regularity constant.

Some results below will exploit the sum of double exponentials functions $H_k\colon [0, 1) \to \mathbb{R}_+$, cf. [23]:

$$H_k(\delta) = \sum_{\ell=k}^{\infty} \delta^{(2^\ell)}.$$

All functions $H_k(\cdot)$ are monotonically increasing and strictly convex. We also use two specific constants

$$c_1 = H_0\Big(\frac{1}{2}\Big) \approx 0.8164215,$$

and

$$c_2 = \max_{0\leq r\leq \frac{1}{4}} 2H_0\left(\frac{1}{2} - r\right) + 5r - 4r^2 - 2c_1 \approx 0.0036003. \tag{9}$$

Trivial approximations

$$0 \leq H_k(\delta) \leq \frac{\delta^{(2^k)}}{1 - \delta^{(2^k)}} = \frac{1}{\delta^{-(2^k)} - 1}. \tag{10}$$

may be used for polynomial lower and upper bounds of $H_0(\delta) = \delta + \delta^2 + \delta^4 + \ldots + \delta^{(2^{k-1})} + H_k(\delta)$ with arbitrary precision. We also use property $H_k(\delta^2) = H_{k+1}(\delta)$.

Below the problem of solvability of Equation (5) is addressed. We apply algorithm (3) in the form

$$x^{k+1} = x^k - \alpha_k z^k, \quad k = 0, 1, \dots$$
$$z^k \in \operatorname{Arg}\min_z \{||z|| : g'(x^k)z = g(x^k) - y\}. \tag{11}$$

with small α and prove that the iterations converge while the limit point is a solution. This technique follows the idea from [23]. Remind that $\mathbb{R}^n, \mathbb{R}^m$ are equipped with some norms, the dual norms are denoted $||\cdot||_*$.

Assumptions with respect to (5).

A. $g(0) = 0$ (i.e. $\overline{x} = 0$), $g(x)$ is differentiable in the ball $B = \{x \in \mathbb{R}^n : \|x\| \le \rho\}$, and its derivative $g'(x)$ satisfies Lipschitz condition in B:

$$\|g'(x^a) - g'(x^b)\| \le L\|x^a - x^b\|.$$

B. The following inequality holds for all $x \in B$ and some fixed $\mu > 0$:

$$\|g'(x)^{\mathrm{T}} h\|_* \ge \mu \|h\|_*, \quad \forall\, h \in \mathbb{R}^m.$$

C. $\|y\| < \mu\rho$.

Theorem 3.2: *If conditions* **A**, **B**, **C** *hold then there exists a solution* x^* *of* (5), *and* $\|x^*\| \le \|y\|/\mu$.

Proof: We apply algorithm (11) with $\alpha > 0$ small enough and $x^0 = 0$. The algorithm is well defined – condition **B** and Lemma 3.1 imply existence of solutions z^k provided that $x^k \in B$; this is true for $k = 0$ and will be validated recurrently. Standard formula

$$g(x + z) = g(x) + \int_0^1 g'(x + tz)z \, \mathrm{d}t$$

combined with condition **A** provides for $x = x^k$, $z = -\alpha z^k$ and $u_k = \|g(x^k) - y\|$ recurrent relation

$$u_{k+1} \le |1 - \alpha| u_k + \frac{L\alpha^2}{2}\|z^k\|^2.$$

Now condition **B** and Lemma 3.1 transform this estimate into

$$u_{k+1} \le |1 - \alpha| u_k + \frac{L\alpha^2 u_k^2}{2\mu^2}.$$

Choose $\alpha = \varepsilon\big(2\mu^2(Lu_0)^{-1}\big)\big(1 - u_0(\mu\rho)^{-1}\big)$ with small $\varepsilon < 1$ satisfying $0 < \alpha < 1$; it is possible due to condition **C**. From the above inequality we get $u_{k+1} \le u_k(1 - \alpha + \alpha\varepsilon(u_k/u_0)(1 - u_0(\mu\rho)^{-1})$. For $k = 0$ this implies $u_1 < u_0$ and recurrently $u_{k+1} < u_k$. We also get $u_{k+1} \le q u_k$, $q = 1 - \alpha + \alpha\varepsilon(1 - u_0(\mu\rho)^{-1}) < 1$. Thus $u_k \le q^k u_0$ and $u_k \to 0$ for $k \to \infty$.

On the other hand we have $\|x^{k+1} - x^k\| = \alpha\|z^k\| < \|z^k\| \leq \|g(x^k) - y\|/\mu = u_k/\mu \leq q^k u_0/\mu$. Hence for any k, s and for $k \to \infty$

$$\|x^{k+s} - x^k\| \leq \sum_{i=k}^{k+s-1} \|x^{i+1} - x^i\| \leq q^k \frac{u_0}{(1-q)\mu} \to 0.$$

It means that x^k is a Cauchy sequence and converges to some point $x^*(\varepsilon)$. We had $g(x^k) \to y$, thus continuity reasons imply $g(x^*(\varepsilon)) = y$. Now, for all iterations we got

$$\|x^k - x^0\| = \|x^k\| \leq \sum_{j=0}^{k-1} \|x^{j+1} - x^j\| \leq \alpha \frac{u_0}{\mu(1-q)}$$
$$= \frac{u_0}{\mu} \frac{1}{1 - \varepsilon(1 - \frac{u_0}{\mu\rho})} < \frac{u_0}{\mu} \frac{1}{1 - (1 - \frac{u_0}{\mu\rho})} = \rho.$$

Hence all iterations x^k remain in the ball B and our reasoning was correct. Finally under $\|x^k\| \leq (u_0/\mu)/\big(1 - \varepsilon(1 - u_0(\mu\rho)^{-1})\big)$ we take $\varepsilon \to 0$, leading to $\|x^k\| \leq u_0/\mu$, so its limit point $x^*(\varepsilon)|_{\varepsilon\to 0}$. The limit point $x^*(\varepsilon)|_{\varepsilon\to 0} = x^*$ is a solution as well and $\|x^*\| \leq u_0/\mu$. ■

Corollary 3.3: *If $\rho = \infty$ (that is conditions* **A**, **B** *hold on the entire space $\mathbb{R}^n$) then Equation* (5) *has a solution for an arbitrary right-hand side y.*

It is worth noting that if we apply pure Newton method (i.e. take $\alpha_k \equiv 1$), the conditions of its convergence are more restrictive: we need $\|y\| \leq 2\mu^2/L$, that is we guarantee only local convergence even for $\rho = \infty$. This is a corollary of Newton-Mysovskikh theorem [14], which proof is valid for under-determined case as well, cf. also [23].

Corollary 3.4: *If $m = n$ and Condition* **B** *is replaced with $\|g'(x)^{-1}\| \leq 1/\mu, x \in B$, then the statement of Theorem* 3.2 *holds true.*

In this case our method (11) reduces to classical Newton method (2).

There exist numerous results on solvability of (5). Some of them are stronger than Theorem 3.2 and are based on general notion of metric regularity [7,12]. We provided the proof based on our technique to exhibit its applicability to existence theorems.

4. Main algorithms

Here it is more convenient to use the main equation in form (1). In previous Section we proved solvability of equation by use of the algorithm with constant $\alpha_k \equiv \alpha > 0$; choosing α smaller we obtained larger solvability domain.

However, in this Section our goal is different – to reach the fastest convergence to a solution. For this purpose different strategies for design of step-sizes are needed. The basic

policy is as follows. First, we rewrite assumptions in new notation. We remark that the assumptions in context of equation $P(x) = 0$ are very much the same as **A**, **B**.

A′. $P(x)$ is differentiable in the ball $B = \{x \in \mathbb{R}^n : \|x - x^0\| \le \rho\}$, and its derivative $P'(x)$ satisfies Lipschitz condition in B:

$$\|P'(x^a) - P'(x^b)\| \le L\|x^a - x^b\|.$$

B′. The following inequality holds for all $x \in B$ and some $\mu > 0$:

$$\|P'(x)^{\mathrm{T}} h\|_* \ge \mu \|h\|_*, \quad \forall\, h \in \mathbb{R}^m.$$

If **A**′, **B**′ hold true, we have the same recurrent inequalities for $u_k = \|P(x^k)\|$:

$$u_{k+1} \le |1 - \alpha_k| u_k + \frac{L\alpha_k^2 \|z^k\|^2}{2}, \tag{12}$$

$$u_{k+1} \le |1 - \alpha_k| u_k + \frac{L\alpha_k^2 u_k^2}{2\mu^2}, \tag{13}$$

the second one being just continuation of the first one based on the estimate $\|z^k\| \le u_k/\mu$, compare with the calculations in the proof of Theorem 3.2. Now we can minimize right-hand sides of these inequalities over α_k; it is natural to expect that such choice of step-size imply the fastest convergence of u_k to zero and thus the fastest convergence of iterations x_k to the solution. One of the main contributions of this paper is careful analysis of the resulting method.

If one applies such policy based on inequality (13), optimal α depends on μ, L (Algorithm 1 below). The values are hard to estimate in most applications, thus the method would be hard for implementation. Fortunately, we can modify the algorithm using parameter adjustment (Algorithm 2). On the other hand the same policy based on (12) requires just the value of Lipschitz constant L, which is commonly available (Algorithm 3).

Thus we arrive to an algorithm which we call *Newton method* while in fact it is blended *pure Newton* with *damped Newton* with special rule for damping. In some relation it reminds *Newton method* for minimization of self-concordant functions [21]. Despite its simplicity, the idea of minimizing upper bound seems to be unexplored (or long forgotten) in Newton method theory with respect to equations. Authors found similar choice of step-size in [5], in different conditions and without explicit convergence bounds.

4.1. Newton method with known constants

If both constants L and μ are known, then the step-size is taken as the minimizer of right-hand side of (13):

$$\alpha_k = \arg\min_{\alpha} \left(|1 - \alpha| \cdot \|P(x^k)\| + \frac{L\alpha^2 \|P(x^k)\|^2}{2\mu^2} \right) = \min\left\{1, \frac{\mu^2}{L\|P(x^k)\|}\right\}. \tag{14}$$

Algorithm 1 (Basic Newton method)

$$z^k \in \operatorname{Arg} \min_{P'(x^k)z=P(x^k)} \|z\|,$$

$$x^{k+1} = x^k - \min\left\{1, \frac{\mu^2}{L\|P(x^k)\|}\right\} z^k, \ k \geq 0. \tag{15}$$

The algorithm is well-defined, as soon $\|P(x^k)\| = 0$ means that a solution is already found (formally $z^k = 0$, $\alpha_k = 1$ thereafter). We remind that in calculation of z^k any vector norm in $\mathbb{R}^n$ can be used, also any vector norm in $\mathbb{R}^m$ can be used for $\|P(x^k)\|$, and constants L, μ must comply with these norms.

The update step in (15) can be written in less compact but more illustrative form:

$$\begin{aligned} x^{k+1} &= x^k - \frac{\mu^2}{L\|P(x^k)\|} z^k, && \text{if } \|P(x^k)\| \geq \frac{\mu^2}{L} && \text{(Stage 1 step)}, \\ x^{k+1} &= x^k - z^k, && \text{otherwise} && \text{(Stage 2 step)}. \end{aligned}$$

The latter case is a pure Newton step while the primal one is a damped Newton step. Direction z^k calculation is the same in both stages. The result on convergence and rate of convergence is given below. We use upper ($\lceil\cdot\rceil$) and lower ($\lfloor\cdot\rfloor$) rounding to integer; constant $c_1 \approx 0.8164$ was introduced in Section 3. The theorem is followed by the corollary with simpler statements.

Theorem 4.1: *Suppose that Assumptions* $\mathbf{A}', \mathbf{B}'$ *hold and*

$$\|P(x^0)\| \leq \frac{\mu^2}{L} F_1^{\text{inv}}\left(\frac{L}{\mu}\rho\right), \tag{16}$$

where $F_1^{\text{inv}}(\cdot)$ *is the inverse function for the continuous strictly increasing function* $F_1(w)$*, given by*

$$F_1(w) = \begin{cases} 2H_0\left(\dfrac{w}{2}\right), & 0 \leq w \leq 1, \qquad \text{(17a)} \\ \lceil 2w \rceil - 2 + 2H_0\left(\dfrac{1}{2} - \dfrac{\lceil 2w \rceil - 2w}{4}\right), & w > 1. \qquad \text{(17b)} \end{cases}$$

Then the sequence $\{x^k\}$ *generated by Algorithm 1 converges to a solution* $x^* : P(x^*) = 0$*. The values* $\|P(x^k)\|$ *are monotonically decreasing, and there are not more than*

$$k_{\max} = \max\left\{0, \left\lceil \frac{2L}{\mu^2}\|P(x^0)\| \right\rceil - 2\right\} \tag{18}$$

iterations at Stage 1, then followed by Stage 2 steps. At k-th step the following estimates for the rate of convergence hold:

$$\|P(x^k)\| \leq \|P(x^0)\| - \frac{\mu^2}{2L}k, \qquad k < k_{\max}, \qquad \text{(19a)}$$

$$\|x^k - x^*\| \le \frac{\mu}{L}\left(k_{\max} - k + 2H_0\left(\frac{\overline{w}}{2}\right)\right), \qquad k < k_{\max}, \tag{19b}$$

$$\|P(x^k)\| \le \frac{2\mu^2}{L}\left(\frac{\overline{w}}{2}\right)^{(2^{(k-k_{\max})})}, \qquad k \ge k_{\max}, \tag{19c}$$

$$\|x^k - x^*\| \le \frac{2\mu}{L} H_{k-k_{\max}}\left(\frac{\overline{w}}{2}\right), \qquad k \ge k_{\max}. \tag{19d}$$

where $\overline{w} = (L/\mu^2)\|P(x^0)\| - k_{\max}/2 = \min\{(L/\mu^2)\|P(x^0)\|,\ 1 - \frac{1}{2}\lceil 2(L/\mu^2)\|P(x^0)\|\rceil + (L/\mu^2)\|P(x^0)\|\} \in [0, 1)$.

The Theorem's statement may look quite involved, but both functions $H_0(\delta)$ and $F_1^{\text{inv}}(p)$ are easily calculated in practice. In the interval of interest $\delta \in [0, \frac{1}{2}]$, the former function has rational approximation (10). The latter function can be evaluated with needed accuracy via binary search, as soon $F_1(w)$ is monotonically increasing on R_+.

Proof: Assume that all $x^k \in B$, $k \ge 0$. Below we state condition enabling this assumption. Using $w_k = (L/\mu^2)\|P(x^k)\|$ as the objective function, we rewrite (13) with generic step-size α as

$$w_{k+1} \le |1 - \alpha| w_k + \frac{1}{2}\alpha^2 w_k^2. \tag{20}$$

Its optimum over α is at $\alpha_k = 1/w_k < 1$, if $w_k > 1$; and $\alpha_k = 1$ otherwise; it is exactly (14).

During Stage 1 of damped Newton steps ($\alpha_k < 1$), the objective function monotonically decreases as

$$w_{k+1} \le w_k - \frac{1}{2}. \tag{21}$$

There are at most $k_{\max} = \max\{0,\ \lceil 2w_0\rceil - 2\}$ iterations in the phase, say $\bar{k}$ ones, resulting in $w_{\bar{k}} \le 1$. As soon w_k reaches this unit threshold, the algorithm switches to Stage 2, pure Newton steps. Then recurrent relation (20) becomes

$$w_{k+1} \le \frac{1}{2} w_k^2, \quad k \ge \bar{k}.$$

so we can write

$$w_{\bar{k}+\ell} \le 2\left(\frac{w_{\bar{k}}}{2}\right)^{(2^\ell)}, \quad \ell \ge 0. \tag{22}$$

For the second phase $\|x^{i+1} - x^i\| = \|z^i\| \le \|P(x^i)\|/\mu = (\mu/L) w_i$ due Lemma 3.1, and for $\ell_2 \ge \ell_1 \ge 0$ holds

$$\|x^{\bar{k}+\ell_2} - x^{\bar{k}+\ell_1}\| \le \sum_{i=\ell_1}^{\ell_2-1} \|x^{\bar{k}+i+1} - x^{\bar{k}+i}\| \le \frac{2\mu}{L}\left(H_{\ell_1}\left(\frac{w_{\bar{k}}}{2}\right) - H_{\ell_2}\left(\frac{w_{\bar{k}}}{2}\right)\right). \tag{23}$$

The sequence $\{x^k\}$ is a Cauchy sequence because $H_j(w_{\bar{k}}/2) \le H_j(\frac{1}{2}) \to_{j\to\infty} 0$. It converges to a point x^* : $\|P(x^*)\| = \lim_{k\to\infty}\|P(x^k)\| = 0$ due to continuity of P, with

$$\|x^{\bar{k}+\ell} - x^*\| \le \frac{2\mu}{L} H_\ell\left(\frac{w_{\bar{k}}}{2}\right), \quad \ell \ge 0. \tag{24}$$

Next we are to estimate distance from points x^k in Stage 1 to the limit solution point x^*. One-step distance for $k < \overline{k}$ is bounded by a constant: $\|x^{k+1} - x^k\| = \alpha_k \|z^k\| \leq \alpha_k w_k \mu / L = \mu / L$, and altogether

$$\|x^k - x^*\| \leq \|x^{\overline{k}} - x^*\| + \sum_{i=k}^{\overline{k}-1} \|x^{i+1} - x^i\| \leq \frac{\mu}{L}\left(\overline{k} - k + 2H_0\left(\frac{w_{\overline{k}}}{2}\right)\right), \quad k < \overline{k}. \tag{25}$$

Note that the formula also coincides with the upper bound (24) at $k = \overline{k}$. Exact number $\overline{k}$ of the steps in the first phase is not known, but we can replace it with the upper bound $k_{\max}$ in all estimates (21)–(25), due to monotonic decrease of $\{w_k\}$. We also have an upper bound for $w_{k_{\max}} \leq \overline{w} = w_0 - \frac{1}{2}k_{\max} = w_0 - (\max\{0, \lceil 2w_0 \rceil - 2\})/2 \in [0, 1]$. Substituting $w_k = (L/\mu^2)\|P(x^k)\|$ back we arrive to Theorem 4.1 bounds (19).

Finally we are to check our primal assumption of the algorithm-generated points x^k being within B. This is guaranteed by one of two conditions, depending on whether the Algorithm starts from Stage 1 step or Stage 2 step.

In the first case $w_0 > 1$, and $\|x^0 - x^k\|$ is bounded similarly to (25) as

$$\begin{aligned}
\|x^0 - x^k\| &\leq \sum_{i=0}^{k-1} \|x^{i+1} - x^i\| \leq \sum_{i=0}^{\overline{k}-1} \|x^{i+1} - x^i\| + \sum_{i=\overline{k}}^{\infty} \|x^{i+1} - x^i\| \leq \\
&\leq \frac{\mu}{L}\left(\overline{k} + 2H_0\left(\frac{w_{\overline{k}}}{2}\right)\right) \leq \frac{\mu}{L}\left(k_{\max} + 2H_0\left(\frac{w_{k_{\max}}}{2}\right)\right) \leq \\
&\leq \frac{\mu}{L}\left(k_{\max} + 2H_0\left(\frac{\overline{w}}{2}\right)\right) \\
&= \frac{\mu}{L}\left(\lceil 2w_0 \rceil - 2 + 2H_0\left(\frac{1}{2} - \frac{\lceil 2w_0 \rceil - 2w_0}{4}\right)\right).
\end{aligned} \tag{26}$$

Here we also used $k_{\max} = \lceil 2w_0 \rceil - 2 > 0$ and the upper bound $w_{k_{\max}} \leq \overline{w}$. In other words, given $w_0 = (L/\mu^2)\|P(x^0)\| > 1$, for $\{x^k\} \in B$ it is sufficient to satisfy $\rho L/\mu \geq F_1(w_0)$. This corresponds to (17b) part.

In the second case we have $w_0 \leq 1$, and the algorithm makes pure Newton steps with $\alpha_k \equiv 1$ from the beginning. Then $\overline{k} = 0$, $w_{k_{\max}} = w_0$ and from (23) follows

$$\|x^k - x^0\| \leq \frac{2\mu}{L}\left(H_0\left(\frac{w_0}{2}\right) - H_k\left(\frac{w_0}{2}\right)\right) \leq \frac{2\mu}{L} H_0\left(\frac{w_0}{2}\right), \quad k \geq 0.$$

Therefore if $w_0 \leq 1$, then inequality $\|x^0 - x^k\| \leq \rho$, $k \geq 0$ is satisfied if $\rho L/\mu \geq 2H_0(w_0/2) = F_1(w_0)$. This corresponds to (17a) part.

Gluing the cases $w_0 \leq 1$ and $w_0 > 1$ we arrive to the sufficient condition $\rho \geq (\mu/L)F_1\big((L/\mu^2)\|P(x^0)\|\big)$, resulting in $x^k \in B$. Due to $F_1(w)$ being strictly increasing this condition is equivalent to (16). ■

Result on the rate of convergence means, roughly speaking, that after no more than $k_{\max}$ iterations one has very fast (quadratic) convergence. For good initial approximations $k_{\max} = 0$, and pure Newton method steps are performed from the very start.

Corollary 4.2: *If $\rho = \infty$ (that is conditions $\mathbf{A}', \mathbf{B}'$ hold on the entire space $\mathbb{R}^n$), then Algorithm 1 converges to a solution of (1) for any $x^0 \in \mathbb{R}^n$.*

The following corollary provides simpler tight relaxed condition for Theorem 4.1. The idea is to develop an upper bound for (17), resulting in a lower bound on (16).

Corollary 4.3: *Condition (16) can be replaced with piece-wise linear one*

$$\|P(x^0)\| \le \frac{\mu^2}{L} F_1^{\text{inv,lower}}\left(\frac{L}{\mu}\rho\right) = \frac{\mu^2}{L} \times \begin{cases} \dfrac{1}{2(2c_1-1)}\dfrac{L}{\mu}\rho, & 0 \le \rho \le (2c_1-1)\dfrac{\mu}{L}, \\ \dfrac{L}{2\mu}\rho + 1 - c_1, & \rho > (2c_1-1)\dfrac{\mu}{L}. \end{cases}$$

Proof: In order to find *a lower* bound for $F_1^{\text{inv}}(\cdot)$, we are to prove *an upper* bound for the function $F_1(w)$, (17). Both bounds are continuous strictly increasing functions. First we notice, that both (17a) and (17b) coincide at $w \in [\frac{1}{2}, 1]$, and it can be rewritten through the different junction point $w = \frac{1}{2}$ instead of $w = 1$

$$F_1(w) = \begin{cases} 2H_0\left(\dfrac{w}{2}\right), & 0 \le w \le \dfrac{1}{2}, \\ \lceil 2w \rceil - 2 + 2H_0\left(\dfrac{1}{2} - \dfrac{\lceil 2w \rceil - 2w}{4}\right), & w > \dfrac{1}{2}. \end{cases}$$

Due to convexity on interval $[0, \frac{1}{2}]$, function $H_0(\delta)$ is bounded by a secant segment:

$$2H_0\left(\frac{w}{2}\right) \le 2H_0\left(\frac{1}{4}\right) \cdot (2w) = 2(2c_1 - 1)w.$$

Here we used property $H_0(\frac{1}{4}) = H_0(\frac{1}{2}) - \frac{1}{2}$, which follows from the identity $H_k(x) = x^{(2^k)} + H_k(x^2)$, $x \in [0, 1)$; and constant $c_1 = H_0(\frac{1}{2})$ is introduced in Section 3.

Next we show that $F_1(w) \le 2(w + c_1 - 1)$ for $w \ge \frac{1}{2}$. Indeed, following continuous function is periodic on w, $w \ge \frac{1}{2}$. It can be also written through the variable $r = (1 + 2w - \lceil 2w \rceil)/2 \in (0, \frac{1}{2}]$.

$$\begin{aligned} F_1(w) - 2(w + c_1 - 1) &= 2\left(\frac{\lceil 2w \rceil - 2w}{2} + H_0\left(\frac{1}{2} - \frac{\lceil 2w \rceil - 2w}{4}\right) - H_0\left(\frac{1}{2}\right)\right) \\ &= 2\left(-r + H_0\left(\frac{1}{4} + \frac{r}{2}\right) - H_0\left(\frac{1}{4}\right)\right) \\ &= 2\left(H_0\left(\frac{1}{4} + \frac{r}{2}\right) - \left(H_0\left(\frac{1}{4}\right) + r\right)\right) \le 0. \end{aligned} \tag{27}$$

In the second row identity $H_0(\frac{1}{2}) = \frac{1}{2} + H_0(\frac{1}{4})$ is used. The last inequality is due to convexity of $H_0(\delta)$, which is under secant segment on corresponding interval $[\frac{1}{4}, \frac{1}{2}]$.

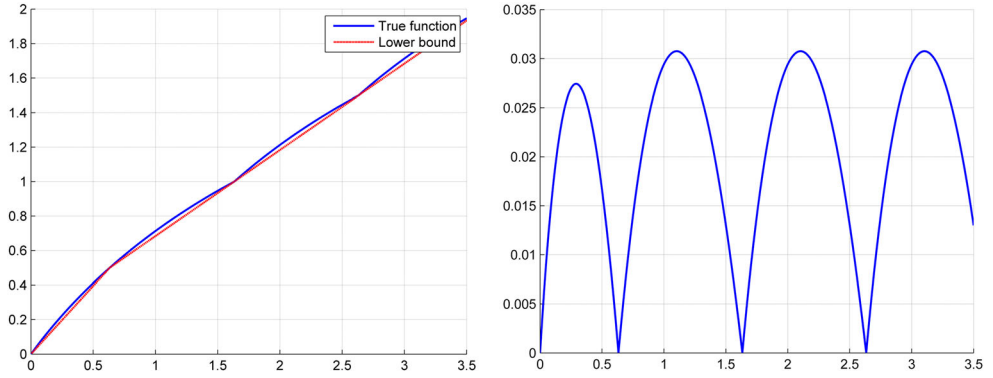

Figure 1. Algorithm 1: function $F_1^{\text{inv}}(p)$ and lower bound $F_1^{\text{inv,lower}}(p)$ (left), residual $F_1^{\text{inv}}(p) - F_1^{\text{inv,lower}}(p)$ (right).

Thus we have the monotonically increasing piece-wise linear upper bound for (17)

$$F_1(w) \le F_1^{\text{upper}}(w) = \begin{cases} 2(2c_1 - 1)w, & 0 \le w \le \dfrac{1}{2}, \\ 2(w + c_1 - 1), & w > \dfrac{1}{2}. \end{cases}$$

Its inverse $F_1^{\text{inv,lower}}$ with property $F_1^{\text{inv,lower}}(F_1^{\text{upper}}(w)) \equiv w, w \ge 0$ is the piece-wise linear lower bound for F_1^{inv}:

$$F_1^{\text{inv}}(p) \ge F_1^{\text{inv,lower}}(p) = \begin{cases} \dfrac{1}{2(2c_1 - 1)}p, & 0 \le p \le 2c_1 - 1, \\ \dfrac{p}{2} + 1 - c_1, & p > 2c_1 - 1. \end{cases}$$

Substituting back $p = \rho L/\mu$ and $w = (L/\mu^2)\|P(x^0)\|$ results in the Corollary statement. ■

On Figure 1 the bound and its residual are plotted. From the proof it is clear that the bounds are tight.

The linear upper (lower) bounds of $F_1(\cdot), F_1^{\text{inv}}(\cdot)$ for the interval $w \in [0, \frac{1}{2}]$ were chosen for consistency with linear bounds on $w \in [\frac{1}{2}, \infty)$. Bound residual on these two intervals are also in the same order, cf. Figure 1. In Corollary 4.8 of Theorem 4.5 we present refined, quadratic approximation for $F_1(w) = 2H_0(w/2),\ w \le \frac{1}{2}$.

Corollary 4.4: *The upper bounds* (19) *may be simplified as well, using* $\overline{w} \le 1$ *and thus* $H_0(\overline{w}/2) \le H_0(\frac{1}{2}) = c_1$*:*

$$\|x^k - x^*\| \le \frac{\mu}{L}(k_{\max} - k + 2c_1), \quad k < k_{\max},$$

$$\|P(x^k)\| \le \frac{\mu^2}{L}\frac{1}{2^{(2^{k-k_{\max}})-1}}, \quad k \ge k_{\max},$$

$$\|x^k - x^*\| \le \frac{2\mu}{L}H_{k-k_{\max}}\left(\frac{1}{2}\right) \le \frac{2\mu}{L}\frac{1}{2^{(2^{k-k_{\max}})} - 1}, \quad k \ge k_{\max}.$$

4.2. Adaptive Newton method

Presented Algorithm 1 explicitly uses two constants μ and L but both enter into the algorithm as one parameter $\beta = \mu^2/L$. There is a simple modification allowing adaptively change an estimate of the parameter.

Input of the algorithm is an initial point x^0, the approximation β_0 and the shrinkage parameter $0 < q < 1$.

Algorithm 2 (Adaptive Newton method)

1. Calculate
$$z^k \in \operatorname{Arg} \min_{P'(x^k)z = P(x^k)} \|z\|,$$
$$\alpha_k = \min\left\{1, \frac{\beta_k}{\|P(x^k)\|}\right\},$$
$$u_{k+1} = \|P(x^k - \alpha_k z^k)\|.$$
2. If either
$$\alpha_k < 1 \text{ and } u_{k+1} < u_k - \frac{\beta_k}{2},$$
or
$$\alpha_k = 1 \text{ and } u_{k+1} < \frac{1}{2\beta_k} u_k^2,$$
holds, go to Step 4. Otherwise
3. apply update rule $\beta_k \leftarrow q\beta_k$ and return to Step 1 without increasing the counter.
4. Take
$$x^{k+1} = x^k - \alpha_k z^k,$$
set $\beta_{k+1} = \beta_k$, increase counter $k \leftarrow k + 1$, and go to Step 1.

Properties of Algorithm 2 are similar to Algorithm 1. We omit the formal proof of convergence; it follows the lines of the proof of Theorem 4.1 with respect to the properties:

- Algorithm 2 runs real steps at Step 4 and some number of fictitious steps resulting in update rule Step 3;
- β_k is non-increasing sequence;
- if $\beta_k < \beta$ (the actual constant of the objective function), then Step 3 won't appear and β_k won't decrease anymore. It means that there are at most $\widehat{k} = \max\{0, \lceil \log_{1/q}(\beta_0/\beta) \rceil\}$ check steps. Minimal possible value of β_k is $\beta_{\min} = q^{\widehat{k}}\beta_0$, and the number of Stage 1 steps is limited by $\widehat{k}_{\max} = \max\{0, \lceil 2\|P(x^0)\|/\beta_{\min} \rceil - 2\}$ as well;

- if Step 4 is performed with $\beta_k > \beta$ due to validity of a condition in Step 2, then $\|P(x^{k+1})\|$ decreases *more* than at the corresponding step with 'optimal' step-size $\alpha_k = \min\{1,\ \beta/\|P(x^k)\|\}$ (calculated with 'true' value β).

Let us mention two other versions of adaptive Newton method. The first one uses increasing updates (e.g. $\beta_{k+1} = q_2\beta_k$ with $q_2 > 1$) in the end of Step 4, thus adapting the constant to current x^k. Also other decrease policies can be applied for β_k in Step 3.

The alternative to the Algorithm 2 is line-search or Armijo-like rules for choosing step-size α_k to minimize objective function $\|P(x^k - \alpha z^k)\|$ directly. It is known that this approach eventually leads to the quadratic convergence rate with pure Newton steps as well, but without any estimates [2]. The difference between is the following: in the proposed Algorithm 2 parameter β is being *monotonically tuned* to the global problem-specific value, rather than trial-and-error procedure is performed at *every* iteration in the Armijo-like approach. We compare the alternatives in Example 1.

4.3. Method for L known

Constant μ, used in Assumptions $\mathbf{B}, \mathbf{B}'$, is rarely accessible. As said in the beginning of the section, we can use more accurate upper bound (12) instead of (13) for step-size choice. It results in the algorithm, which uses the Lipschitz constant only. The optimal step-size in this case is

$$\alpha_k^* = \arg\min_{\alpha}\left(|1-\alpha|\cdot\|P(x^k)\| + \frac{L\alpha^2\|z^k\|^2}{2}\right) = \min\left\{1, \frac{\|P(x^k)\|}{L\|z^k\|^2}\right\}. \tag{28}$$

Algorithm 3 (L-Newton method)

$$z^k \in \operatorname{Arg}\min_{P'(x^k)z = P(x^k)} \|z\|,$$

$$x^{k+1} = x^k - \min\left\{1, \frac{\|P(x^k)\|}{L\|z^k\|^2}\right\} z^k,\ k \geq 0.$$

The algorithm is well-defined, as condition $\|z^k\| = 0$ holds only if $P(x^k) = 0$, i.e. a solution was found at the previous step. Formally we put $z^k = 0,\ \alpha_k = 1$ and $x^{k+1} = x^k$ thereafter.

For the Algorithm we also present similar convergence theorem and set of corollaries. We emphasize that while constant μ is still used in the bounds, Algorithm 3 does not depend on it.

Theorem 4.5: *Suppose that Assumptions* $\mathbf{A}', \mathbf{B}'$ *hold and*

$$\|P(x^0)\| \leq \frac{\mu^2}{L} F_2^{\text{inv}}\left(\frac{L}{\mu}\rho\right), \tag{29}$$

where $F_2^{\mathrm{inv}}(\cdot)$ is the inverse function for the continuous strictly increasing function $F_1(w)$ given by

$$F_2(w) = \begin{cases} 2H_0\left(\frac{w}{2}\right), & 0 \le w \le 1, \quad (30a) \\ \frac{(\lceil 2w\rceil - 2)(4w - \lceil 2w\rceil + 3)}{4} + 2H_0\left(\frac{1}{2} - \frac{\lceil 2w\rceil - 2w}{4}\right), & w > 1. \quad (30b) \end{cases}$$

Then the sequence $\{x^k\}$ generated by Algorithm 3 converges to a solution $x^* : P(x^*) = 0$. The values $\|P(x^k)\|$ are monotonically decreasing, at k-th step the following estimates for the rate of convergence hold:

$$\|P(x^k)\| \le \|P(x^0)\| - \frac{\mu^2}{2L}k, \quad k < k_{\max}, \quad (31a)$$

$$\|x^k - x^*\| \le \frac{\mu}{L}\left(\frac{(4w_0 - \lceil 2w_0\rceil + 3 - k)(\lceil 2w_0\rceil - 2 - k)}{4} + 2H_0\Big(\frac{\overline{w}}{2}\Big)\right), k < k_{\max}, \quad (31b)$$

$$\|P(x^k)\| \le \frac{2\mu^2}{L}\Big(\frac{\overline{w}}{2}\Big)^{(2^{(k-k_{\max})})}, \quad k \ge k_{\max}, \quad (31c)$$

$$\|x^k - x^*\| \le \frac{2\mu}{L}H_{k-k_{\max}}\Big(\frac{\overline{w}}{2}\Big), \quad k \ge k_{\max}. \quad (31d)$$

where $\overline{w} = (L/\mu^2)\|P(x^0)\| - k_{\max}/2 = \min\{(L/\mu^2)\|P(x^0)\|,\ 1 - \frac{1}{2}\lceil 2(L/\mu^2)\|P(x^0)\|\rceil + (L/\mu^2)\|P(x^0)\|\} \in [0, 1)$.

For proving the theorem we need the simple proposition about real sequences.

Proposition 4.6: *Consider two non-negative real sequences $w_k \ge 0, v_k \ge 0,\ k \ge 0$, and functions $h_k(v), f(v), k \ge 0$. Let $f(\cdot)$ be monotonically increasing function, being also a majorant function for $h_k(\cdot)$ with respect to $\{w_k\}$. Namely, we require $h_k(w_k) \le f(w_k)$. If $w_0 \le v_0$ and the sequences satisfy*

$$w_{k+1} \le h_k(w_k),$$
$$v_{k+1} = f(v_k),$$

then $w_k \le v_k,\ k \ge 0$.

The proposition is trivially proved by induction step $w_{k+1} \le h_k(w_k) \le f(w_k) \le f(v_k) = v_{k+1}$.

The proof of Theorem 4.5 resembles the proof of Theorem 4.1, and it uses majorization idea. Main issue is due to different step-size, now there is no clear separation between damped and pure Newton steps.

Proof: We compare two discrete processes, both starting with the same value $v_0 = w_0$. The first sequence is generated by recurrent equality $v_{k+1} = f(v_k)$, where

$$f(v) = \min_{\alpha}\left(|1 - \alpha|v + \frac{\alpha^2}{2}v^2\right) = \begin{cases} v - \frac{1}{2}, & v > 1, \\ \frac{1}{2}v^2, & v \le 1, \end{cases}$$

is monotonically increasing function on $v \geq 0$. The second process is $w_k = (L/\mu^2)\|P(x^k)\|$, with $\{x^k\}$ generated by Algorithm 3. Assume that all $x^k \in B$ and thus Assumptions $\mathbf{A}', \mathbf{B}'$ hold. Then due to main inequality (12) and step-size (28)

$$w_{k+1} \leq h_k(w_k) = \min_{\alpha} \left(|1-\alpha| w_k + \frac{\alpha^2}{2} \left(\frac{L\|z^k\|}{\mu} \right)^2 \right).$$

Here we used $\|z^k\|$ in parametric part $a_k^2 \geq 0$ within introduced function $h_k(w) = \min_\alpha(|1-\alpha|w + a_k^2\alpha^2/2)$. From $a_k = (L/\mu)\|z^k\| \leq (L/\mu^2)\|P(x^k)\| = w_k$ by Assumption $\mathbf{B}'$, the functions $|1-\alpha|w + a_k^2\alpha^2/2$ under minimization in $h_k(\cdot)$ are majorized by corresponding functions $|1-\alpha|v + w_k^2\alpha^2/2 \leq |1-\alpha|v + v^2\alpha^2/2$ for $v \geq w_k$. Minimums of the functions over α (implicitly dependent on w_k, v) also satisfy $h_k(v) \leq f(v)$ whenever $v \geq w_k$. It follows that $h_k(w_k) \leq f(w_k)$.

Therefore sequences $\{w_k\}$ and $\{v_k\}$, alongside with functions $h_k(\cdot), f(\cdot)$ satisfy Proposition 4.6, given $v_0 = w_0 = (L/\mu^2)\|P(x^0)\|$. Now we have upper bound on w_k through v_k for all $k \geq 0$. Next we closely follow the lines and calculations of the proof of Theorem 4.1.

Analysis of $\{v_k\}$ is the same as analysis of the upper bound in Theorem 4.1. First, v_k decreases, and the number of steps until v_k reaches 1 is the same $k_{\max}$, given by (18). Explicit expressions on v_k are

$$\begin{aligned} v_k &= v_0 - \frac{1}{2}k, && k \leq k_{\max}, \\ v_k &= 2\left(\frac{v_{k_{\max}}}{2}\right)^{(2^{k-k_{\max}})}, && k > k_{\max}. \end{aligned}$$

where $v_{k_{\max}} = \overline{w} = v_0 - k_{\max}/2 = \min\{w_0, 1 - (\lceil 2w_0 \rceil - 2w_0)/2\}$ (remind that $v_0 = w_0$ by definition). These expressions result in the bounds (31a) and (31c) on $\|P(x^k)\|$, which are the same as in Theorem 4.1.

For the late steps with $k \geq k_{\max}$ we have $w_k \leq v_k \leq 1$, and thus $\alpha_k = 1$ (just because $\|P(x^k)\|/(L\|z^k\|^2) \geq \mu^2/(L\|P(x^k)\|) = 1/w_k \geq 1$). Then points $x^k, k \geq k_{\max}$ form a Cauchy sequence like (23), and obey similar to (24) bound:

$$\begin{aligned} \|x^{k_{\max}+\ell} - x^*\| &\leq \sum_{i=\ell}^{\infty} \|z^k_{\max} + i\| \leq \frac{\mu}{L} \sum_{i=\ell}^{\infty} w_{k_{\max}} + i \\ &\leq \frac{\mu}{L} \sum_{i=\ell}^{\infty} v_{k_{\max}} + i = \frac{2\mu}{L} H_\ell \left(\frac{\overline{w}}{2}\right), \quad \ell \geq 0. \end{aligned}$$

This is (31d) bound, by the way the same as (19d) of Theorem 4.1.

What is the main difference from Theorem 4.1 proof, is the distance counting until $k_{\max}$-th step. In this case we assume $k_{\max} = \lceil 2w_0 \rceil - 2 > 0$. Due to the Algorithm's step-size choice now we cannot be sure, whether α_k^* be always less than 1 thereafter or not. For $i < k_{\max}$ we have $\|x^{i+1} - x^i\| = \alpha_i^*\|z^i\| \leq \|z^i\| \leq \|P(x^i)\|/\mu = w_i\mu/L \leq v_i\mu/L = (v_0 - \frac{1}{2}i)\mu/L$, and arrive to (31b):

$$\|x^k - x^*\| \leq \|x^{k_{\max}} - x^*\| + \sum_{i=k}^{k_{\max}-1} \|x^{i+1} - x^i\|$$

$$\le 2\frac{\mu}{L}H_0\Big(\frac{\overline{w}}{2}\Big) + \frac{\mu}{L}\sum_{i=k}^{k_{\max}-1}\Big(v_0 - \frac{1}{2}i\Big)$$
$$= \frac{\mu}{L}\left(\frac{(4v_0 - k_{\max} + 1 - k)(k_{\max} - k)}{4} + 2H_0\Big(\frac{\overline{w}}{2}\Big)\right)$$
$$= \frac{\mu}{L}\left(\frac{(4w_0 - \lceil 2w_0\rceil + 3 - k)(\lceil 2w_0\rceil - 2 - k)}{4} + 2H_0\Big(\frac{\overline{w}}{2}\Big)\right).$$

In (31b) we used explicit formula for $\overline{w} = 1 + w_0 - (\lceil 2w_0\rceil)/2$ in case $k_{\max} > 0$.

The last part of the proof is checking assumption $x^k \in B$, i.e. $\|x^0 - x^k\| \le \rho$. From the derivation of bound (31b) above we have

$$\|x^0 - x^k\| \le \sum_{i=0}^{k-1}\|x^{i+1} - x^i\| \le \sum_{i=0}^{k_{\max}-1}\|x^{i+1} - x^i\| + \sum_{i=k_{\max}}^{\infty}\|x^{i+1} - x^i\|$$
$$\le \frac{\mu}{L}\left(\frac{(4w_0 - \lceil 2w_0\rceil + 3)(\lceil 2w_0\rceil - 2)}{4} + 2H_0\Big(\frac{\overline{w}}{2}\Big)\right), \quad k \ge 0,$$

in case $k_{\max} > 0$, i.e. $w_0 \ge 1$, and $\|x^0 - x^k\| \le (2\mu/L)H_0(w_0/2)$ otherwise (from derivation of bound (31de)). Thus sufficient condition for $x^k \in B$ is

$$\rho \ge \frac{\mu}{L}F_2\Big(\frac{L}{\mu^2}\|P(x^0)\|\Big),$$

which is equivalent to (29). ■

Like for Algorithm 1, we can state few corollaries: on global convergence and some simple bounds.

Corollary 4.7: *If $\rho = \infty$ (that is conditions $\mathbf{A}', \mathbf{B}'$ hold on the entire space $\mathbb{R}^n$) then Algorithm 3 converges to a solution of (1) for any $x^0 \in \mathbb{R}^n$.*

There is a simpler tight relaxed condition for Theorem 4.5.

Corollary 4.8: *Condition (29) can be replaced with*

$$\|P(x^0)\| \le \frac{\mu^2}{L}F_2^{\text{inv,lower}}\Big(\frac{L}{\mu}\rho\Big)$$
$$= \frac{\mu^2}{L} \times \begin{cases} \frac{1}{4} - 2c_3 + \sqrt{\Big(2c_3 - \frac{1}{4}\Big)^2 + \frac{2L}{\mu}\rho}, & 0 \le \rho \le c_3\frac{\mu}{L}, \\ -\frac{1}{4} + \sqrt{\frac{L}{\mu}\rho - c_3 + \frac{9}{16}}, & \rho > c_3\frac{\mu}{L}, \end{cases}$$

where constant $c_3 = 2c_1 + c_2 - 1 \approx 0.66885$.

The idea of the proof is the same as of Corollary 4.3: we derive upper bound $F_2^{\text{upper}}(w)$ for function $F_2(w)$ (30). Then the sufficient condition for points x^k generated by Algorithm 3 being inside B is $\rho \geq (\mu/L)F_2^{\text{upper}}\big((L/\mu^2)\|P(x^0)\|\big)$. Its inverse function $F_2^{\text{inv,lower}} : F_2^{\text{inv,lower}}(F_2^{\text{upper}}(w)) \equiv w, w \geq 0$ is a lower bound for $F_2^{\text{inv}}(\cdot)$ then.

Proof: Notice that both components of $F_2(w)$ are the same in $w \in [\frac{1}{2}, 1]$, so

$$F_2(w) = \begin{cases} 2H_0\Big(\dfrac{w}{2}\Big), & 0 \leq w \leq \dfrac{1}{2}, \\ \dfrac{(\lceil 2w\rceil - 2)(4w - \lceil 2w\rceil + 3)}{4} + 2H_0\left(\dfrac{1}{2} - \dfrac{\lceil 2w\rceil - 2w}{4}\right), & w > \dfrac{1}{2}. \end{cases}$$

Let's begin with the case $w > \frac{1}{2}$. We introduce the auxiliary function $r(w) = (2w - \lceil 2w\rceil)/4 \in (0, \frac{1}{4}]$, related with the fractional part. This function is periodic on w, and

$$\begin{aligned} F_2(w) &= \Big(w + \frac{1}{4}\Big)^2 + 2H_0\Big(\frac{1}{2} - r(w)\Big) - \Big(\frac{5}{4} - 2r(w)\Big)^2 \\ &\leq \Big(w + \frac{1}{4}\Big)^2 + 2c_1 + c_2 - \frac{25}{16}, \quad w > \frac{1}{2}. \end{aligned}$$

Here we used definition (9) of constant c_2, introduced in Section 3.

By the definition of $H(\cdot)$ one can select terms up to quadratic in $H_0(\delta) = \delta + \delta^2 + H_2(\delta)$, thus $H_2(\frac{1}{4}) = H_0(\frac{1}{4}) - \frac{5}{16} = H_0(\frac{1}{2}) - \frac{1}{2} - \frac{5}{16} = c_1 - \frac{13}{16}$. From convexity we have the upper linear bound for $H_2(\delta) \leq (4c_1 - \frac{13}{4})\delta,\ \delta \in [0, \frac{1}{4}]$, and consequently for $2H_0(w/2)$:

$$\begin{aligned} 2H_0\Big(\frac{w}{2}\Big) &= 2\frac{w}{2} + 2\frac{w^2}{4} + 2H\Big(\frac{w}{2}\Big) \leq \frac{w^2}{2} + \Big(4c_1 - \frac{9}{4}\Big)w \\ &\leq \frac{w^2}{2} + \Big(4c_1 + 2c_2 - \frac{9}{4}\Big)w, \quad 0 \leq w \leq \frac{1}{2}. \end{aligned}$$

In the last inequality we manually added a small positive linear term $2c_2 w$ for continuity of the resulting upper bound. Combining two parts, we arrive to the increasing continuous upper bound for $F_2(\cdot)$:

$$F_2(w) \leq F_2^{\text{upper}}(w) = \begin{cases} \dfrac{w^2}{2} + \Big(4c_1 + 2c_2 - \dfrac{9}{4}\Big)w, & 0 \leq w \leq \dfrac{1}{2}, \\ \Big(w + \dfrac{1}{4}\Big)^2 + 2c_1 + c_2 - \dfrac{25}{16}, & w > \dfrac{1}{2}. \end{cases}$$

Using definition of $c_3 = F_2^{\text{upper}}(\frac{1}{2}) = 2c_1 + c_2 - 1$, inverse of this function is the lower bound for $F_2^{\text{inv}}(\cdot)$

$$F_2^{\text{inv}}(p) \geq F_2^{\text{inv,lower}}(p) = \begin{cases} \dfrac{1}{4} - 2c_3 + \sqrt{\Big(2c_3 - \dfrac{1}{4}\Big)^2 + 2p}, & 0 \leq p \leq c_3, \\ -\dfrac{1}{4} + \sqrt{p - c_3 + \dfrac{9}{16}}, & p > c_3. \end{cases}$$

■

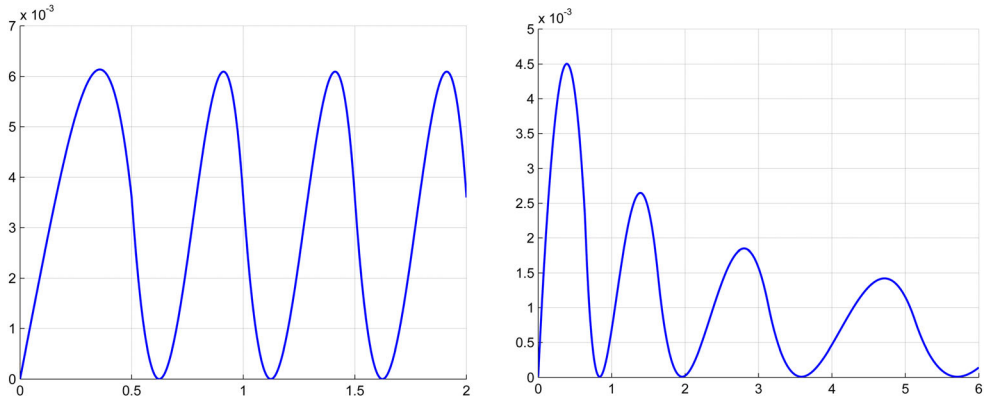

Figure 2. Algorithm 3: residual of upper bound $F_2^{\text{upper}}(w) - F_2(w)$ (left) and residual of lower bound $F_2^{\text{inv}}(p) - F_2^{\text{inv,lower}}(p)$ (right).

The presented bounds are sharp and quite exact. Visually paired graphics of $F_2(w)$ and $F_2^{\text{upper}}(w)$, $F_2^{\text{inv}}(p)$ and $F_2^{\text{inv,lower}}(p)$ looks the same, and its residuals are plotted in Figure 2.

Surprisingly enough, in practice the Algorithm 3 (and its adaptive modification) sometimes converges faster than Algorithm 1, possibly because direction-wise (along z^k) Lipschitz constant is less or equal than uniform Lipschitz constant of Assumption $\mathbf{A}'$, and the convergence rate can be better.

The idea of adaptive algorithm with estimates L_k works as well for Algorithm 3; including its modifications with increasing L_k.

4.4. Pure Newton method

For comparison let us specify convergence conditions of pure Newton method ($\alpha_k = 1$).

Theorem 4.9: *Let conditions* $\mathbf{A}', \mathbf{B}'$ *hold. If* $\delta = \big(L/(2\mu^2)\big)\|P(x^0)\| < 1$ *and* $(2\mu/L)H_0(\delta) \le \rho$, *then pure Newton method converges to a solution* x^* *of* (1), *and*

$$\|P(x^k)\| \le \frac{2\mu^2}{L}\delta^{(2^k)}, \quad \|x^k - x^*\| \le \frac{2\mu}{L}H_k(\delta).$$

It coincides with Corollary 1 of [23], proven in the Banach space setup (a misprint in [23] is corrected here). For $m = n$ case the result is a minor extension of Mysovskikh's theorem [14].

5. Special cases

In the section we outline few important cases in more detail, namely solving equations with special structure, solving scalar equations or inequalities, solvability of quadratic equations.

5.1. Structured problems

The problem is to solve equation $g(x) = y$ where $g(x)_i = \varphi(c_i^{\mathrm{T}} x),\ c_i \in R^n, i = 1, \dots m$. Here $\varphi(t)$ is a twice differentiable scalar function,

$$|\varphi'(t)| \geq \mu_\varphi > 0, \quad |\varphi''(t)| \leq L_\varphi, \quad \forall\, t.$$

It is not hard to see that Assumptions **A**, **B** hold on the entire space $\mathbb{R}^n$ and Algorithm 1 converges, with Theorem 4.1 and Corollary 4.3 providing rate of convergence. The rate of convergence depends on estimates for μ, L, which can be written as functions of μ_φ, L_φ and minimal and maximal singular values $\sigma_{\min}, \sigma_{\max}$ of matrix C with columns c_i (we suppose that C has full rank, thus $\sigma_{\min} > 0$). Indeed after simple calculations (see expression for $g'(x)$ with C below) we get

$$L \leq \sigma_{\max}^2 L_\varphi, \quad \mu \geq \sigma_{\min} \mu_\varphi. \tag{33}$$

However the special structure of the problem allows to get much sharper results. Let's use notation $P(x) = g(x) - y$. Indeed $P'(x) = g'(x) = D(x)C^{\mathrm{T}}$, $D(x) = \operatorname{diag}(\varphi'(c_i^{\mathrm{T}} x))$ and repeating the proof of Theorem 3.2 we get the equality

$$P(x^{k+1}) = |1 - \alpha| P(x^k) - \alpha \int_0^1 (D_t - D) C^{\mathrm{T}} z^k \, \mathrm{d}t, \quad \alpha \geq 0,$$

where $D_t = D(x^k - \alpha t z^k), D = D(x^k)$. Thus (recall $u_k = \|P(x^k)\|$)

$$u_{k+1} \leq |1 - \alpha| u_k + \alpha ||C^{\mathrm{T}} z^k|| \int_0^1 ||D_t - D|| \, \mathrm{d}t \leq |1 - \alpha| u_k + \frac{L_\varphi \alpha^2 ||C^{\mathrm{T}} z^k||^2}{2}$$

Identity between spectral norm of a diagonal matrix and Euclidean norm of vector on the diagonal is used in the last line, followed by element-wise Lipschitz property of $\varphi'(\cdot)$: $\|D_t - D\| = \|[\varphi'(c_i^{\mathrm{T}}(x^k - \alpha t z^k)) - \varphi'(c_i^{\mathrm{T}} x^k)]\| \leq \|[|\varphi'(c_i^{\mathrm{T}}(x^k - \alpha t z^k)) - \varphi'(c_i^{\mathrm{T}} x^k)|]\| \leq \|[L_\varphi \alpha t |c_i^{\mathrm{T}} z^k|]\| = L_\varphi \alpha t \|[c_i^{\mathrm{T}} z^k]\| = L_\varphi \alpha t \|C^{\mathrm{T}} z^k\|$ for $t \geq 0$. Thus $\int_0^1 ||D_t - D|| \, \mathrm{d}t \leq \frac{1}{2} L_\varphi \alpha \|C^{\mathrm{T}} z^k\|$.

But $P'(x^k) z^k = P(x^k)$, thus $DC^{\mathrm{T}} z^k = P(x^k)$, $C^{\mathrm{T}} z^k = D^{-1} P(x^k)$ and hence $||C^{\mathrm{T}} z^k|| \leq u_k / \mu_\varphi$. We arrive to the inequality, very similar to (13), but with different constant γ

$$u_{k+1} \leq |1 - \alpha| u_k + \gamma \frac{\alpha^2 u_k^2}{2}, \quad \gamma = \frac{L_\varphi}{\mu_\varphi^2}.$$

Hence $u_{k+1} \leq u_k - \frac{1}{2}\gamma$ at Stage 1, thus this inequality does not depend on C! As the result we get estimates for the rate of convergence which are the same for ill-conditioned and well-conditioned matrices C. Of course this estimate is much better than standard one with $\gamma = (\sigma_{\max}/\sigma_{\min})^2 L_\varphi / \mu_\varphi^2$ which follows from (33).

This example is just an illustrating one (explicit solution of the problem can be found easily), but it emphasizes the role of special structure in equations to solve. Numerical experiments with such problems are provided below, in Section 6.3.

5.2. One-dimensional case

Suppose we solve one equation with n variables:

$$f(x) = 0, \quad f : \mathbb{R}^n \to \mathbb{R}.$$

Here 0 is *not a minimal value* of f, thus it is not a minimization problem! Nevertheless, our algorithms will remind some minimization methods. This case has some specific features compared with arbitrary m. For instance calculation of z^k may be done explicitly. Norm in image space is absolute value $|\cdot|$, and ℓ_p norms in pre-image space $\mathbb{R}^n$, $p \in \{1, 2, \infty\}$ can be considered. Then

$$\begin{aligned}
z^k &= \frac{f(x^k)\ \mathrm{sign}\,(\nabla f(x^k)_i)}{\|\nabla f(x^k)\|_\infty} e^i,\ i \in \mathrm{Arg}\max_i |\nabla f(x^k)_i|, && \text{in case of } \ell_1\text{−norm},\\
z^k &= \frac{f(x^k)}{\|\nabla f(x^k)\|_2^2} \nabla f(x^k), && \text{in case of Euclidean norm},\\
z^k &= \frac{f(x^k)}{\|\nabla f(x^k)\|_1} \mathrm{sign}\,(\nabla f(x^k)), && \text{in case of } \ell_\infty\text{−norm},
\end{aligned}$$

where $e^j = (0, \ldots, 0, 1, 0, \ldots, 0)^{\mathrm{T}}$ is j-th orth vector, and $\mathrm{sign}\,(\cdot)$ function is coordinate-wise sign function, $\mathrm{sign} : \mathbb{R}^n \to \{-1, 1\}^n$.

Constant μ (and μ_0) are also calculated explicitly via conjugate (dual) vector norm as $\mu = \min_{x \in B} \|\nabla f(x)\|_*$, $\mu_0 = \|\nabla f(x^0)\|_*$. For any norms $\|z^k\| = |f(x^k)|/\|\nabla f(x^k)\|_*$, and in Algorithm 3 damped Newton step is performed iff $\|\nabla f(x^k)\|_*^2 < L|f(x^k)|$, otherwise pure Newton step is made.

If we choose ℓ_1 norm, the method becomes coordinate-wise one. Thus, if we start with $x^0 = 0$ and perform few steps (e.g. we are in the domain of attraction of pure Newton algorithm) we arrive to a *sparse* solution of the equation.

In Euclidean case a Stage 1 step (damped Newton) of Algorithm 3 is

$$x^{k+1} = x^k - \frac{1}{L}\,\mathrm{sign}\,(f(x^k))\nabla f(x^k),$$

which is exactly gradient minimization step for function $|f(x^k)|$. Stage 2 (pure Newton) step is

$$x^{k+1} = x^k - \frac{f(x^k)}{\|\nabla f(x^k)\|_2^2} \nabla f(x^k).$$

This reminds well-known subgradient method for minimization of convex functions. However in our case we do not assume any convexity properties, and the direction may be either gradient or anti-gradient in contrast with minimization methods!

5.3. Quadratic equations

Proceed to a specific nonlinear equation, namely the quadratic one. Then the function $g(x)$ may be written componentwise as (6), with gradients

$$\nabla g_i(x) = A_i x + b_i \in \mathbb{R}^n, \quad i = 1, \ldots, m.$$

Obviously $g(0) = 0$, the question is solvability of $g(x) = y$. There are some results on construction of the entire set of feasible points $Y = \{y : g(x) = y\} = g(\mathbb{R}^n)$, including

its convexity, see e.g. [24]. We focus on local solvability, trying to derive the largest ball inscribed in Y.

The derivative matrix $g'(x)$ is formed row-wise as

$$g'(x) = \begin{bmatrix} \nabla g_1(x)^{\mathrm{T}} \\ \vdots \\ \nabla g_m(x)^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} x^{\mathrm{T}} A_1 + b_1^{\mathrm{T}} \\ \vdots \\ x^{\mathrm{T}} A_m + b_m^{\mathrm{T}} \end{bmatrix} \in \mathbb{R}^{m \times n}.$$

One has $g'(0) = H$, H being $m \times n$ matrix with rows b_i. We suppose H has rank m (recall that $m \leq n$), then its smallest singular value $\sigma_{\min}(H) > 0$ serves as μ_0.

The derivative $g'(x)$ is linear on x, thus it has uniform Lipschitz constant L on $\mathbb{R}^n$, and assumption **A** holds everywhere. There are several estimates for the Lipschitz constants, for example (for ℓ_2 norm)

$$L \leq L_1 = \sqrt{\lambda_{\max}\left(\sum_{i=1}^{m} A_i^{\mathrm{T}} A_i\right)}$$

from [25], where $\lambda_{\max}$ is the maximal eigenvalue of a matrix. Other estimates can be obtained via elaborate convex semidefinite optimization problem (SDP), cf. [32] for details.

Quadratic equations play significant role in power system analysis, because power flow equations are quadratic, see [17]. It is of interest to compare our estimates with some known results on solvability of power flow equations [34].

5.4. Solving systems of inequalities

Below we address some tricks to convert systems of inequalities into systems of equations.

First, if one seeks a solution of a system of inequalities

$$g_i(x) \leq 0, \quad i = 1, \ldots, m, \quad x \in \mathbb{R}^{\ell},$$

then by introducing slack variables the problem is reduced to solution of the under-determined system of equations

$$g_i(x) + x_{\ell+i}^2 = 0, \quad i = 1, \ldots, m, \quad x \in \mathbb{R}^n, \; n = \ell + m.$$

Similarly finding a feasible point for linear inequalities $x \geq 0, Ax = b$, $x \in \mathbb{R}^n$, $b \in \mathbb{R}^m$ can be transformed to the under-determined system

$$\sum_{j=1}^{n} A_{ij} z_i^2 = b_i, \quad i = 1, \ldots, m, \quad z \in \mathbb{R}^n.$$

The efficiency of such reductions is unclear a priori and should be checked by intensive numerical study.

6. Numerical tests

We have performed several experiments to check effectiveness of the proposed approach for solving Equations (1) and to compare it with known ones.

The first two experiments relate to the classical case $n = m$, i.e. the number of variables equals the number of equations. Algorithm 2 with adaptive parameter estimation is compared with well-known 'backstepping' Armijo-like techniques for the damped Newton method. Namely, the competitor is the step-size proposed in [2].

$$\gamma_k = q^j : \|P(x^k + q^j z^k)\| \leq (1 - cq^j)\|P(x^k)\| \tag{34}$$

with some shrinkage parameter $q \in (0, 1)$ and relaxation parameter $c \in (0, 1)$. This method in some sense is similar to our Algorithm 2, but there are differences in step-size choice.

Two other examples relate to under-determined case, i.e. $n > m$. Example 3 is the illustration how to employ structure of the data as explained in Subsection 5.1. We show that such approach strongly accelerates convergence. Final Example 4 is an optimal control problem. Exploiting L_1 norms we construct *sparse* controls for the minimal-fuel problem.

6.1. Example 1

We studied the Fletcher–Powell system of equations [8]:

$$\sum_{j=1}^{n} A_{ij} \sin x_j + B_{ij} \cos x_j = E_i, \quad i = 1, 2, \ldots, n \tag{35}$$

for various dimensions n. The data are generated as proposed in original paper [8]: matrices $A, B \in \mathbb{R}^{n\times n}$ are random, then for some $x^* \in \mathbb{R}^n$ right-hand sides $E \in \mathbb{R}^n$ are calculated. The arising system of equations may have many solutions, however we have the guarantee that it is solvable. Then a set of 1000 initial points x^0 were randomly sampled (multistart policy). From each of the initial points, Newton algorithms were run with a) Armijo-kind step-size (34), and b) β-adaptive algorithm (Algorithm 2). Parameters of the algorithms were chosen as $\beta_0 = 100, q = 0.95, c = 0.8$.

Each run has a 'success' or 'fail' result. 'Success' means that accuracy $||P(x^k)|| < 10^{-8}$ is achieved, while 'failure' is marked when either step-size threshold is $\gamma_k < 10^{-13}$ attained or maximal number of iterations ($N = 10000$) is performed. For each dimension n data of algorithms' outcomes were aggregated as following. For a random sample of equations (there were 100 for each n) 'success ratio' r was calculated among 1000 initial points as ratio of success runs and total number of runs: $N_{success}/1000$, both for Armijo-like approach (r_{Armijo}), and for our Algorithm 2 ($r_{Alg.\,2}$). To emphasize comparison, we used ratio of ratios $r_{Alg.\,2}/r_{Armijo}$ as indicator. Then $r_{Alg.\,2}/r_{Armijo}$ values were imaged as the box-and-whisker plot for all dimensions (Figure 3). The middle line in a (quartile) box is the median, whiskers' lengths are set to 0.05 and 0.95 percentiles, outlier data is dot-plotted - as soon there are 100 points (for each of the dimensions), there are exactly 5 upper and 5 lower outliers).

We see that with high probability the ratio is larger than 1 for large dimensions. As a conclusion, our method finds a solution more often than Armijo-like approach.

The similar analysis was done for the function calls. For each of the samples the numbers of function calls (these can be many in one Newton step) are averaged over initial

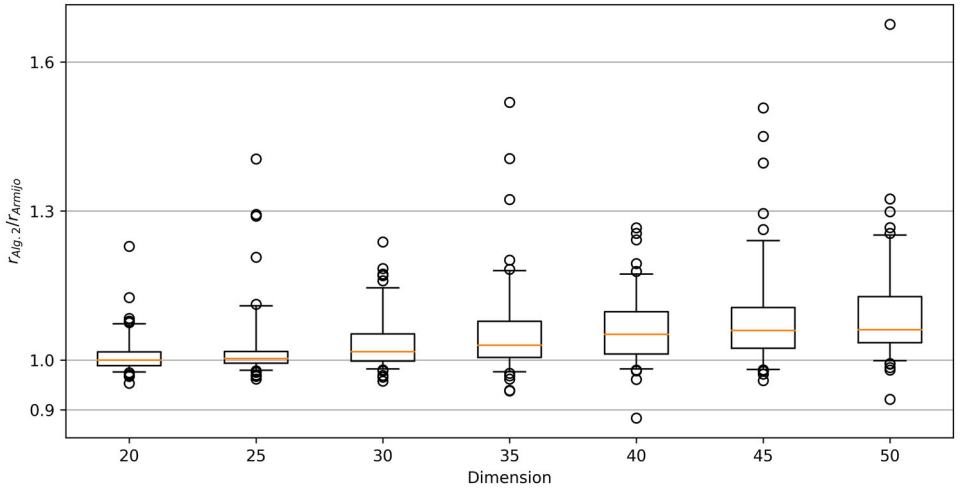

Figure 3. Box-and-whisker plot for the ratios of success ratios $r_{Alg.\,2}/r_{Armijo}$ for all dimensions (Ex. 1).

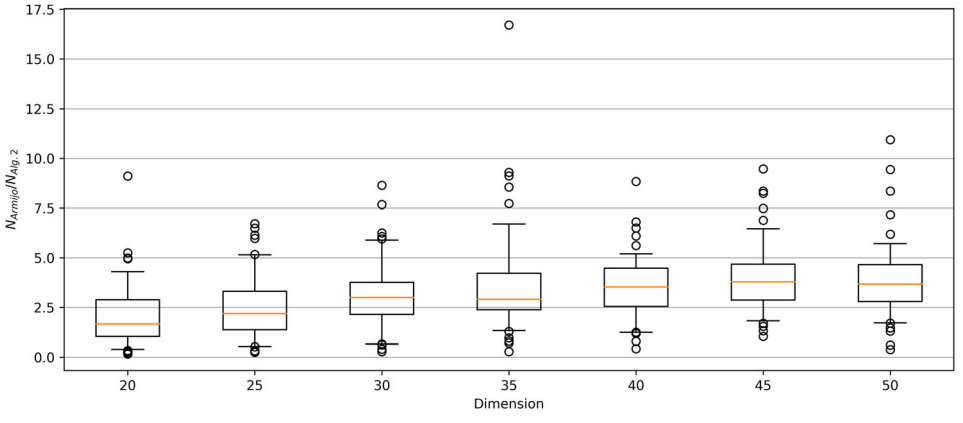

Figure 4. Box-and-whisker plot for the ratios of function calls $N_{Armijo}/N_{Alg.\,2}$ for all dimensions (Ex. 1).

conditions for all runs, resulting in values $N_{Alg.\,2}, N_{Armijo}$. The ratios of the averaged function calls for Armijo-like step-size (34) and Algorithm 2 ($N_{Armijo}/N_{Alg.\,2}$) were aggregated on the box-and-whisker plot on Figure 4. Typically Algorithm 2 admits much less functions evaluations compared with Armijo-like algorithm.

6.2. Example 2

The original problem is equality-constrained optimization:

$$\min_{x:h(x)=0} f(x)$$

with scalar differentiable functions f, h, and $x \in R^n$. By use of Lagrange multiplier rule it is reduced to the solution of equations

$$P(X) = \begin{bmatrix} \nabla f(x) + \nu \nabla h(x) \\ h(x) \end{bmatrix} = 0$$

with new variable $X = [x^{\mathrm{T}}, \nu]^{\mathrm{T}} \in R^{n+1}$. Note that we are interested at any solution of these equations, that is we do not distinguish minimum points and stationary points. The derivative is a block matrix

$$P'(X) = \begin{bmatrix} \nabla^2 f(x) + \nu \nabla^2 h(x) & \nabla h(x) \\ (\nabla h(x))^{\mathrm{T}} & 0 \end{bmatrix}$$

We address the simplest case: minimization of a quadratic function (with symmetric A) on Euclidean unit sphere:

$$\min_{||x||^2=1} \frac{1}{2}(Ax, x) + (b, x)$$

Let the constraint be defined by $h(x) = \frac{1}{2}(x, x) - \frac{1}{2}$, then

$$P(X) = \begin{bmatrix} Ax + b + \nu x \\ \frac{1}{2}x^{\mathrm{T}}x - \frac{1}{2} \end{bmatrix}, \quad P'(X) = \begin{bmatrix} A + \nu I_n & x \\ x^{\mathrm{T}} & 0 \end{bmatrix}. \tag{36}$$

The experiment were run for different dimensions ($n = 20, 25, 30, 35, 40, 45, 50$). For each dimension 100 problems were randomly generated, and 1000 initial points were randomly chosen for each problem. Then Algorithm 2 and Armijo-like step-size algorithm (34) were run as in the first example. The parameters and stopping criteria were the same as in Example 1, except for the parameter $q = 0.85$, and number of Newton steps were bounded by 1000. Matrix A of the quadratic objective function is a positive semidefinite matrix, formed as $A = \frac{1}{2}MM^{\mathrm{T}}$. The auxiliary matrix $M \in \mathbb{R}^{n \times n+4}$ has coefficients, uniformly distributed on $[-0.9, 2.1]$. Coefficients of the linear term b_i are picked up from the scaled Gaussian variables: $b_i \sim 0.1\mathcal{N}(0, 1)$. The initial conditions were chosen for the extended variable $X = [x^{\mathrm{T}}, \nu]^{\mathrm{T}}$ as following:

- the first n components (x^0, corresponding to the original variable x) are sampled from the uniform distribution on Euclidean sphere with radius 1,
- the last, $n + 1$ component (Lagrange multiplier ν) is chosen as the 'best approximation', i.e. as the minimizer of the residual $\|P(X)\|^2 = \|Ax^0 + b + \nu x^0\|^2$. The explicit solution depends on the first n components (x^0),

$$\nu_0 = -(x^0)^{\mathrm{T}}Ax^0 - (x^0)^{\mathrm{T}}b.$$

Same as in Example 1, success ratios $r_{Armijo}, r_{Alg.\,2}$ and number of function calls $N_{Armijo}, N_{Alg.\,2}$ were calculated (averaged over initial points). On Figure 5 the ratios of $r_{Alg.\,2}/r_{Armijo}$, and on Figure 6 the ratios $N_{Armijo}/N_{Alg.\,2}$ were gathered in box-and-whisker plot.

Here it appears, that the success ration of Algorithm 2 prevails over Armijo-like approach only at dimension 50 (still the ratio is about 1, meaning that both algorithms behave quite similar); however, the average of function evaluations of our algorithm still lower for all cases.

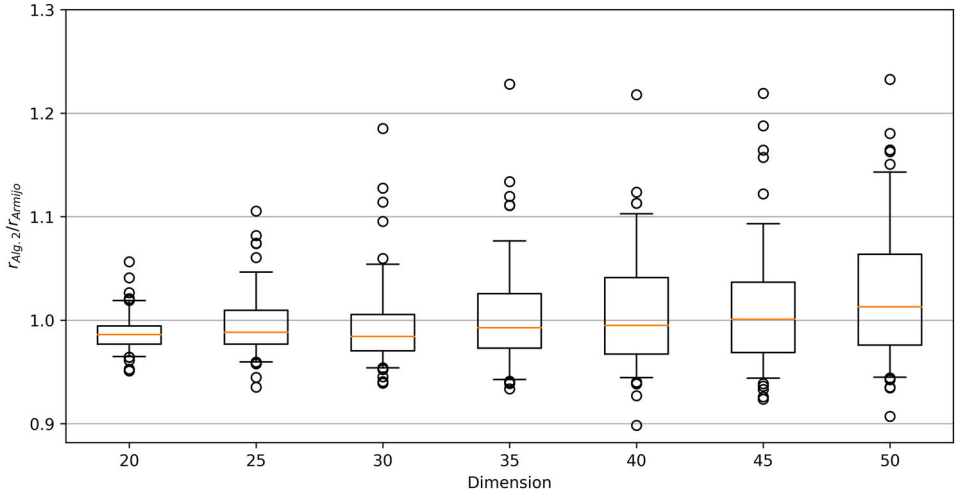

Figure 5. Box-and-whisker plot for the ratios of success ratios $r_{Alg.\ 2}/r_{Armijo}$ for all dimensions (Ex. 2).

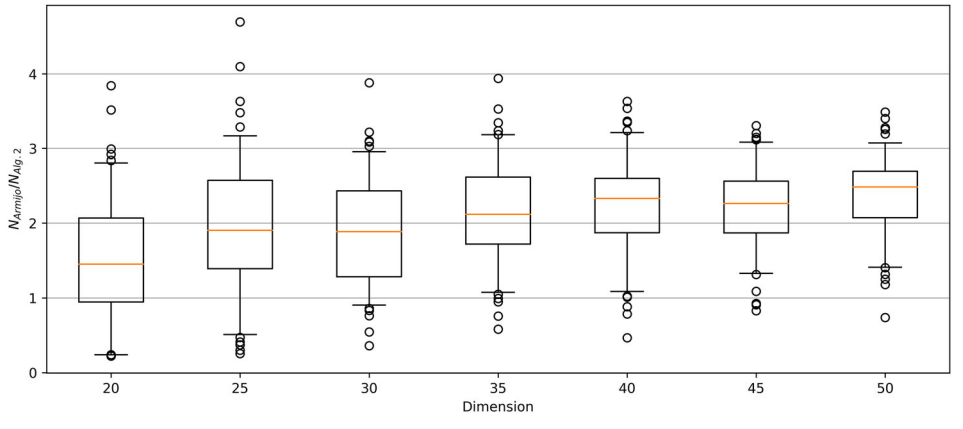

Figure 6. Box-and-whisker plot for the ratios of function calls $N_{Armijo}/N_{Alg.\ 2}$ for all dimensions (Ex. 2).

6.3. Example 3

The problem is described in Section 5.1; it is to solve $g(x) = y$ where $g(x)_i = \varphi(c_i^{\mathrm{T}} x),\ c_i \in R^n, i = 1, \dots m$. Here $\varphi(t)$ is twice differentiable scalar function,

$$|\varphi'(t)| \geq \mu_\varphi > 0, \quad |\varphi''(t)| \leq L_\varphi, \quad \forall\, t.$$

It has been explained in Section 5.1 that the special structure of the problem allows to get much sharper results.

Here we restrict ourselves with the single example to demonstrate how the methods work for medium-size problems ($n = 40$, $m = 21$). The equations have special structure as in Section 5.1:

$$P_i(x) = \varphi(c_i^{\mathrm{T}} x - b_i) - y_i, \quad x \in \mathbb{R}^n, \quad y \in \mathbb{R}^m,$$

where

$$\varphi(t) = \frac{t}{1+e^{-|t|}}, \quad \varphi'(t) = \frac{1+(1+|t|)e^{-|t|}}{(1+e^{-|t|})^2}.$$

Matrix C with rows c_i, vectors b, y were generated randomly. For function $\varphi(t)$ we have $\mu_\varphi = \max_t \varphi'(t) \geq 0.5$, $L_\varphi = \max_t |\varphi''(t)| \leq 2$ for all t. Thus if we do not pay attention to the special structure of the problem we have $\mu \geq 0.5\sigma_{\min}(C)$, $L \leq 2\sigma_{\max}(C)$. On the other hand if we take into account the structure we can replace μ^2/L ($= 0.0012$ in the example) in Algorithm 1 (see Subsection 4.3) with $\mu_\varphi^2/L_\varphi = 0.125$.

The results of simulations are as follows. When we apply Algorithm 1 with values L, μ, it requires 6000 iterations to achieve the accuracy $||P(x^k)|| < 10^{-12}$, while the same algorithm with L_φ, μ_φ requires just 70 iterations. The similar result holds for Algorithm 3: the version exploiting L requires 30 iterations, exploiting L_φ - just 5 iterations. All algorithms which are not based on information on these constants (pure Newton, adaptive Algorithm 2) also converge in 5 iterations.

These results demonstrate how sensitive can be the proposed algorithms to a priori data and to the special structure of equations.

6.4. Example 4.

This test is devoted to the underdetermined systems of equations and, specifically, sparsity property, arising in optimal control problems. The behaviour of a pendulum with force control u is given by the second-order differential equation

$$\ddot{\phi} + \alpha\dot{\phi} + \beta \sin\phi = u.$$

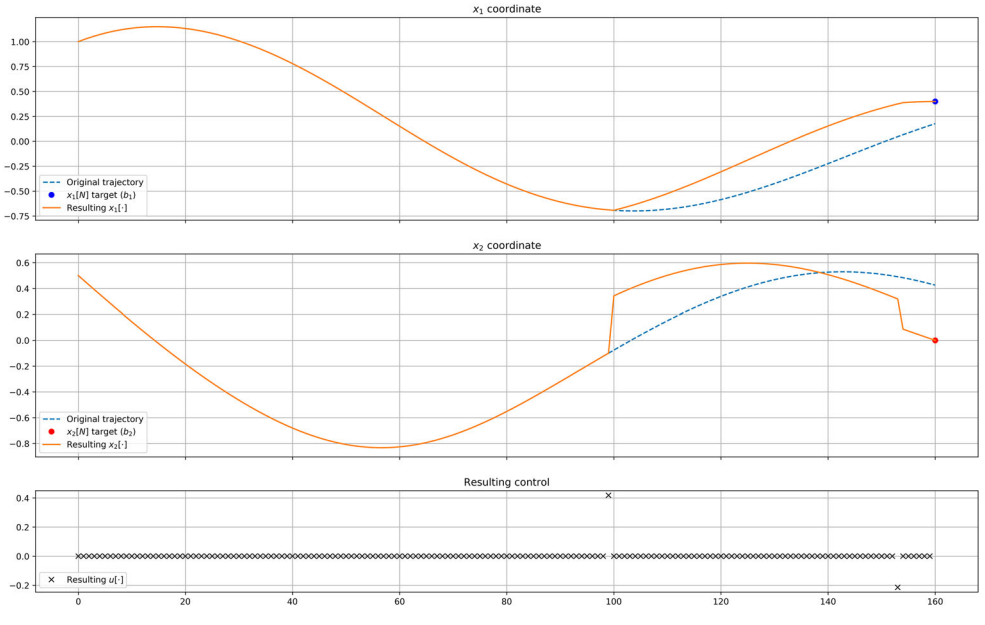

Figure 7. Solution with two-impulse control, [27].

Given some initial condition $\phi(0), \dot{\phi}(0)$, the goal is to drive the pendulum to the specified terminal position and angular speed $[\phi(T), \dot{\phi}(T)]^{\mathrm{T}} = b \in \mathbb{R}^2$ for the fixed time T. The secondary goal is to have sparse control and the least control capacity $\int_0^{\mathrm{T}} |u(t)|\, \mathrm{d}t$.

The model was discretized on the interval $[0, T]$. The discretized control U has dimension $N = T/h - 1$, where h is the discretization step. The problem is to solve two equations $[\phi_d(T, U), \psi_d(T, U)]^{\mathrm{T}} = b$, where ϕ_d, ψ_d are the discrete counterparts of $\phi, \dot{\phi}$, in N dimensional variable U.

The problem was solved by exploiting Algorithm 2 with specific choice of norm, namely ℓ_1-norm. First, it represents a discretized control capacity ($\|u\| = \sum_{i=0}^{N} |U^{(i)}|$). Second, it is known for its property of finding sparse solution. The initial approximation was $U = 0$. Newton method converges in 3 steps, resulting in 5 non-zero components of control (i.e. the control should be applied at 5 time instants only). Moreover, the first Newton step reveals 2 components (time instants), which are sufficient to get to the goal, see Figure 7. Thus 2-impulse control (with impulses at $t = 98$ and $t = 153$) solves the problem.

Details on the simulation can be found in [27].

7. Conclusions and future research

New solvability conditions for under-determined equations (with wider solvability set) are proposed. The algorithms for finding a solution are easy to implement, they combine weaker assumptions on initial approximations and fast convergence rate. No convexity assumptions are required. The algorithms have large flexibility in using prior information, various norms and problem structure. It is worth mentioning that we do not try to convert the problem into optimization one. Combination of damped/pure Newton method is a contribution for solving classic $n = m$ problems as well.

There are numerous directions for future research.

(1) We suppose that the auxiliary optimization problem for finding direction z^k is solved exactly. Of course an approximate solution of the sub-problem suffices.
(2) The algorithms provide a solution of the initial problem which is not specified a priori. Sometimes we are interested in the solution closest to x^0, i.e. $\min_{P(x)=0} \|x - x^0\|$. An algorithm for this purpose is of interest.
(3) More general theory of structured problems (Section 5.1) is needed.
(4) It is not obvious how to introduce regularization techniques into the algorithms.

Acknowledgments

The authors thank Yuri Nesterov and Alexander Ioffe for helpful discussions and references; the comments of the anonymous reviewers are highly acknowledged.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Russian Science Foundation under Grant 16-11-10015.

Notes on contributors

Boris Polyak (PhD, Doctor of Science) is a chief researcher at Ya.Z. Tsypkin Laboratory, Institute for Control Sciences of Russian Academy of Sciences, Moscow, Russia, and Professor of Moscow University of Physics and Engineering. His research interests include mathematical programming, nonsmooth optimization, stochastic estimation and optimization, linear and nonlinear analysis and design, robust stability, chaos control and Monte-Carlo simulation.

Andrey Tremba (PhD) is a senior reseracher at Ya.Z.Tsypkin Laboratory, Institute for Control Sciences of Russian Academy of Sciences, Moscow, Russia. His research interests include nonlinear optimization, robust stability and control.

ORCID

Boris Polyak http://orcid.org/0000-0002-1898-2984
Andrey Tremba http://orcid.org/0000-0001-5783-7600

References

[1] A. Ben-Israel, *A Newton-Raphson method for the solution of systems of equations*, J. Math. Anal. Appl. 15 (1966), pp. 243–252.
[2] O.P. Burdakov, *Some globally convergent modifications of Newton's method for solving systems of nonlinear equations*, Soviet Math. Dokl. 22 (1980), pp. 376–379.
[3] J. Burke and S.-P. Han, *A Gauss-Newton approach to solving generalized inequalities*, Math. Oper. Res. 11 (1986), pp. 632–643.
[4] J.E. Dennis and R.B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, SIAM, Philadelphia, 1996.
[5] P. Deuflhard, *Newton Methods for Nonlinear Problems. Affine Invariance and Adaptive Algorithms*, Springer Series in Computational Mathematics, Vol. 35, Springer, Berlin, 2004.
[6] A.L. Dontchev, *The Graves theorem revisited*, J. Convex Anal. 3 (1996), pp. 45–53.
[7] A.L. Dontchev and R.T. Rockafellar, *Implicit Function and Solution Mapping*, 2nd ed, Springer, New York, 2014.
[8] R. Fletcher and M.J.D. Powell, *A rapidly convergent descent method for minimization*, Comput. J. 6 (1963), pp. 163–168.
[9] L.M. Graves, *Some mapping theorems*, Duke Math. J. 17 (1950), pp. 111–114.
[10] W.W. Hager, *Analysis and implementation of a dual algorithm for constrained optimization*, J. Optimiz. Theory App. 79 (1993), pp. 427–461.
[11] W.M. Hüßler, *A Kantorovich-type convergence analysis for the Gauss-Newton-method*, Numerische Mathematik. 48 (1986), pp. 119–125.
[12] A.D. Ioffe, *Variational Analysis of Regular Mappings*, Springer, 2017.
[13] L.V. Kantorovich, *The method of successive approximations for functional analysis*, Acta. Math. 71 (1939), pp. 63–97.
[14] L.V. Kantorovich and G.P. Akilov, *Functional Analysis*, 2nd ed., Pergamon Press, Oxford, 1982.
[15] C.T. Kelley, *Solving Nonlinear Equations with Newton's Method*, SIAM, Philadelphia, 2003.
[16] Y. Levin and A. Ben-Israel, *A Newton method for systems of m equations in n variables*, Nonlinear Anal. 47 (2001), pp. 1961–1971.
[17] J. Machowski, J. Bialek and J. Bumby, *Power System Dynamics. Stability and Control*, 2nd ed., John Wiley & Sons Ltd., 2012.
[18] G.G. Magaril-Il'yaev and V.M. Tikhomirov, *Newton's method, differential equations and the Lagrangian principle for necessary extremum conditions*, Proc. Steklov Inst. Math., 262 (2008), pp. 149–169.
[19] M.Z. Nashed and X. Chen, *Convergence of Newton-like methods for singular operator equations using outer inverses*, Numerische Mathematik. 66 (1993), pp. 235–257.

[20] Yu. Nesterov, *Modified Gauss-Newton scheme with worst case guarantees for global performance*, Optimiz. Meth. Softw. 22 (2007), pp. 469–483.
[21] Yu. Nesterov and A. Nemirovskii, *Interior-point Polynomial Algorithms in Convex Programming*, SIAM, Philadelphia, 1994.
[22] J.M. Ortega and W.C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, SIAM, Philadelphia, 2000.
[23] B.T. Polyak, *Gradient methods for solving equations and inequalities*, USSR Comput. Math. and Math. Phys. 4 (1964), pp. 17–32.
[24] B.T. Polyak, *Quadratic transformations and their use in optimization*, J. Optimiz. Theory Appl. 99 (1998), pp. 553–583.
[25] B.T. Polyak, *Convexity of nonlinear image of a small ball with applications to optimization*, Set-Valued Anal. 9 (2001), pp. 159–168.
[26] B.T. Polyak, *Newton-Kantorovich method and its global convergence*, J. Math. Sci. 133 (2006), pp. 1513–1523.
[27] B. Polyak and A. Tremba, *Sparse solutions of optimal control via Newton method for underdetermined systems*, J. Global Optimiz. (2019), doi:10.1007/s10898-019-00784-z.
[28] B.N. Pshenichnyi, *Newton's method for the solution of systems of equalities and inequalities*, Math. Notes. Academy. Sci. USSR 8 (1970), pp. 827–830.
[29] A. Prusinska and A.A. Tret'yakov, *On the existence of solutions to nonlinear equations involving singular mappings with non-zero p-kernel*, Set-Valued Anal. 19 (2011), pp. 399–416.
[30] S.M. Robinson, *Extension of Newton's method to nonlinear functions with values in a cone*, Numerische Mathematik. 19 (1972), pp. 341–347.
[31] H.F. Walker, *Newton-like methods for underdetermined systems*, in *Computational Solution of Nonlinear Systems of Equations*, E.L. Allgower, K. Georg, eds., Lecture Notes in Applied Mathematics, Vol. 26, AMS, Providence, RI, 1990, pp. 679–699.
[32] Y. Xia, *On local convexity of quadratic transformations*, J. Oper. Res. Soc. China 8 (2014), pp. 341–350.
[33] T. Yamamoto, *Historical developments in convergence analysis for Newton's and Newton-like methods*, J. Comput. Appl. Math. 124 (2000), pp. 1–23.
[34] S. Yu, H.D. Nguyen and K.S. Turitsyn, *Simple certificate of solvability of power flow equations for distribution systems*, Power & Energy Society General Meeting, IEEE. 2015.