B. T. Polyak, Newton–Kantorovich method and its global convergence,
Зап. научн. сем. ПОМИ, 2004, том 312, 256–274

https://www.mathnet.ru/eng/znsl783

Записки научных
семинаров ПОМИ
Том 312, 2004 г.

B. T. Polyak

NEWTON–KANTOROVICH METHOD AND ITS GLOBAL CONVERGENCE

Abstract. In 1948, L. V. Kantorovich extended the Newton method for solving nonlinear equations to functional spaces. This event cannot be overestimated: the Newton–Kantorovich method became a powerful tool in numerical analysis as well as in pure mathematics. We address basic ideas of the method in the historical perspective and focus on some recent applications and extensions of the method and some approaches to overcoming its local nature.

1. Introduction

In 1948, L. V. Kantorovich published a seminal paper [1] where an extension of Newton method for functional spaces was provided. The results were also included in the survey paper [2]. Further developments of the method can be found in [3–7] and in the monograph [8, 9]. This contribution of Kantorovich to one of the fundamental techniques in numerical analysis and functional analysis can not be overestimated; we shall try to analyze it in the historical perspective.

The paper is organized as follows. Section 2 provides the idea of Newton method and the history of its development. Kantorovich's contribution will be addressed in Sec. 3. Newton method in its basic form possesses just local convergence, its global behavior and modifications to achieve global convergence will be discussed in Secs. 4 and 5. The case of underdetermined systems is worth of special consideration (Sec. 6). In its original form, Newton–Kantorovich method is destined for solving equations. However, it has numerous applications in unconstrained (Sec. 7) and constrained (Sec. 8) optimization. For instance, modern polynomial-time interior point methods for convex optimization are based on Newton method. Some extensions of the method and directions for future research are described in Sec. 9.

2. Idea and history of the method

The basic idea of Newton method is very simple: it is linearization.

Suppose $F : R^1 \to R^1$ is a differentiable function, and we are solving the equation

$$F(x) = 0. \tag{1}$$

Starting from the initial point x_0 we can construct a linear approximation of $F(x)$ in the neighborhood of $x_0 : F(x_0 + h) \approx F(x_0) + F'(x_0)h$ and solve the arising linear equation. Thus we arrive to the recurrent method

$$x_{k+1} = x_k - F'(x_k)^{-1}F(x_k), \quad k = 0, 1, \dots . \tag{2}$$

This is the method proposed by Newton in 1669. To be more precise, Newton dealt with polynomials only; in the expression of $F(x+h)$ he discarded higher order terms in h. The method was illustrated on the example $F(x) = x^3 - 2x - 5 = 0$. The starting approximation for the root was $x = 2$. Then $F(2+h) = h^3 + 6h^2 + 10h - 1$, neglecting higher order terms Newton got linear equation $10h - 1 = 0$. Thus the next approximation is $x = 2 + 0.1 = 2.1$ and the process can be repeated for this point. Fig. 1 demonstrates that the convergence to the root of $F(x)$ is very fast.

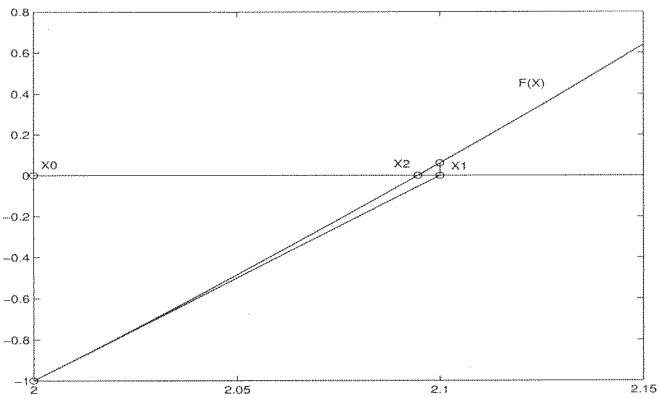

Fig. 1. Newton method.

It was J. Raphson, who proposed in 1690 the general form of method (2) (i.e., $F(x)$ was not assumed to be a polynomial only and the notion of a derivative was exploited), this is why the method is often called Newton–Raphson method.

The progress in development of the method is linked with such famous mathematicians as Fourier, Cauchy, and others. For instance, Fourier in 1818 proved that the method converges quadratically in the neighborhood of a root, while Cauchy (1829, 1847) provided multidimensional extension of (2) and used the method to prove the existence of a root of an equation. Important early contributions to the investigation of the method are due to Fine [10] and Bennet [11]; their papers are published in one volume of *Proceedings of Nat. Acad. of Sci. USA* in 1916. Fine proved the convergence in n-dimensional case with no assumption on the existence of a solution. Bennet extended the result for infinite-dimensional case, this is a surprising attempt because the foundations of functional analysis were not created at the moment. The basic results on Newton method and numerous references can be found in the books by Ostrowski [12] and Ortega and Rheinboldt [13]. More recent bibliography is available in the books [14, 15], survey paper [16] and on the special website, devoted to Newton method [17].

3. Kantorovich's contribution

Kantorovich [1], analyzes the same equation as (1):

$$F(x) = 0, \tag{3}$$

but now $F : X \to Y$, where X, Y are Banach spaces. The proposed method reads as (2)

$$x_{k+1} = x_k - F'(x_k)^{-1}F(x_k), \quad k = 0, 1, \ldots . \tag{4}$$

where $F'(x_k)$ is (Frechet) derivative of the nonlinear operator $F(x)$ at the point x_k and $F'(x_k)^{-1}$ – its inverse. The main convergence result from [1] looks as follows.

Theorem 1. *Suppose F is defined and twice continuously differentiable on a ball $B = \{x : ||x - x_0|| \le r\}$, the linear operator $F'(x_0)$ is invertible, $||F'(x_0)^{-1}F(x_0)|| \le \eta, ||F'(x_0)^{-1}F''(x)|| \le K, x \in B$ and*

$$h = K\eta < 1/2, \quad r \ge \frac{1 - \sqrt{1 - 2h}}{h}\eta. \tag{5}$$

Then Eq. (3) has a solution $x^* \in B$, process (4) is well defined and converges to x^* with quadratic rate:

$$||x_k - x^*|| \le \frac{\eta}{h2^n}(2h)^{2^n}. \tag{6}$$

The proof of the theorem is simple enough; the main novelty of Kantorovich's contribution were not technicalities, but general problem formulation and use of the appropriate techniques of functional analysis. Until Kantorovich's works [1–3] it was not understood that numerical analysis should be considered in the framework of functional analysis (note the title of [2]: "Functional analysis and applied mathematics.") Another peculiarity of Kantorovich's theorem is the lack of assumption on existence of a solution – the theorem is not only convergence result for a specific method, but simultaneously the existence theorem for nonlinear equations.

These properties of Kantorovich's approach ensured a wide range of applications. Numerous nonlinear problems – nonlinear integral equations, ordinary and partial differential equations, variational problems – could be casted in the framework of (3) and method (4) could be applied. Various examples of such applications are presented in monographs [8, 9]. Moreover, Newton–Kantorovich method as a tool for existence results was immediately used in classical works of Kolmogorov, Arnold and Moser (see, e.g., [18]) on "KAM-theory" in mechanics. Actually many classical results in functional analysis are proved by use of Newton-like methods; the typical example is Ljusternik theorem on tangent spaces, see the original paper [19] or modern research [20].

Later, Kantorovich [4, 5] obtained another proof of Theorem 1 and its versions, based on so called "method of majorants." The idea is to compare iterations (4) with scalar iterations, which in a sense majorize them and which possess convergence properties. This approach provides more flexibility; the original proof is a particular case with a quadratic majorant.

There exist numerous versions of Kantorovich's theorem, which differ in assumptions and results. We mention just one of them due to Mysovskikh [21].

Theorem 2. *Suppose F is defined and twice continuously differentiable on a ball $B = \{x : ||x - x_0|| \leq r\}$, the linear operator $F'(x)$ is invertible on B and $||F'(x)^{-1}|| \leq \beta, ||F''(x)|| \leq K, x \in B, ||F(x_0)|| \leq \eta$ and*

$$h = K\beta^2\eta < 2, r \geq \beta\eta \sum_{n=0}^{\infty}(h/2)^{2^n-1}. \tag{7}$$

Then Eq. (3) has a solution $x^* \in B$, process (4) converges to x^* with *quadratic rate:*

$$||x_k - x^*|| \le \frac{\beta\eta(h/2)^{2^k-1}}{1-(h/2)^{2^k}}. \tag{8}$$

The difference with Theorem 1 is assumption on invertibility of $F'(x)$ on B (while in Theorem 1 it was assumed to be invertible in initial point x_0 only) and slighter assumption on h ($h < 2$ instead of $h < 1/2$). Other versions can be found in the books [8, 9, 12, 13, 22, 23].

4. Global behavior

The critical condition for convergence is (5) or (7). They mean that at the initial approximation x_0 the function $||F(x_0)||$ should be small enough, that is x_0 should be close to the solution. Thus Newton–Kantorovich method is locally convergent. Very simple one-dimensional examples demonstrate the lack of global convergence even for smooth monotone $F(x)$. There are many ways to modify the method to achieve its global convergence (we will discuss them later), but the problem of interest is the global behavior of iterations. Obviously there are many simple situations – say, there is a neighborhood S of a solution such that $x_0 \in S$ implies convergence to the solution (such set is called *basin of attraction*) while trajectories starting outside Q do not converge (e.g., tend to infinity). However in the case of non-uniqueness of a solution the structure of basins of attractions may be very complicated and exhibit fractal nature. It was Cayley who formulated this problem as early as 1879.

Let us consider Cayley's example: solve equation $z^3 = 1$ by Newton method. Thus we take $F(z) = z^3 - 1$ and apply method (2):

$$z_{k+1} = z_k - \frac{z_k^3 - 1}{3z_k^2} = \frac{2z_k}{3} + \frac{1}{3z_k^2}.$$

It is worth to mention that we formulated Newton method in real spaces but it is as well applicable in complex spaces; in the above example we take $z_k \in \mathbf{C}$. The equation has three roots

$$z_1^* = 1, z_{2,3}^* = -1/2 \pm i\sqrt{3}/2$$

and it is natural to expect that the entire plane $\mathbf{C}$ is partitioned into three basins of attraction

$$S_m = \{z_0 : z_k \to z_m^*\}, m = 1, 2, 3$$

located around the corresponding roots. However, the true picture is much more involved. First, there is a single point $z = 0$ where the method is not defined. It has three preimages – points z_0 such that $z_1 = 0$, they are $-\rho, \rho(1/2 \pm i\sqrt{3}/2)$, where $\rho = 1/\sqrt[3]{2}$. Again, each of them has three preimages and so on. Thus there are 3^k points z_0 which are mapped after k iterations into the point $z_k = 0$, and they generate a countable set of such initial points, that the method is not applicable at some iteration

$$S_0 = \{z_0 : z_k = 0 \text{ for some } k\}.$$

It can be proved that for all points $z_0 \notin S_0$ the method converges to one of the solutions (note that if $|z_k| > 1$ then $|z_{k+1}| < |z_k|$) and we have

$$\mathbf{C} = \bigcup_{m=0}^{3} S_m.$$

The sets S_m have a fractal structure: S_0 is the boundary of each S_m, $m = 1, 2, 3$ and in any neighborhood of any point $z \in S_0$ there exist points from S_m, $m = 1, 2, 3$. The set S_1 is displayed at Fig. 2, the set S_0 – at Fig. 3.

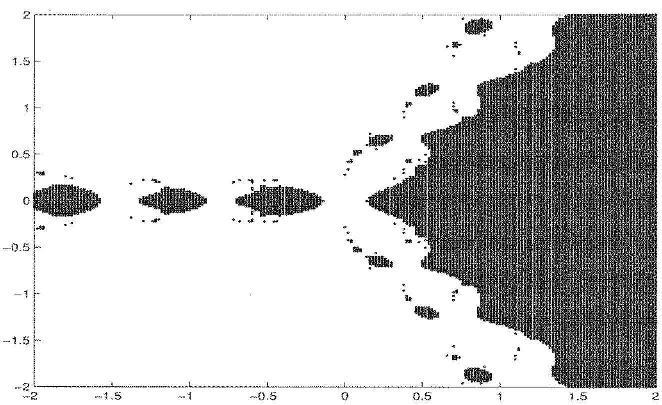

Fig. 2. Basin of attraction for $x^* = 1$.

The sets of initial points which possess no convergence (like the set S_0) for iterations of general rational maps were studied by Julia [24] and are called now *Julia sets*. A lot of examples linked with Newton method

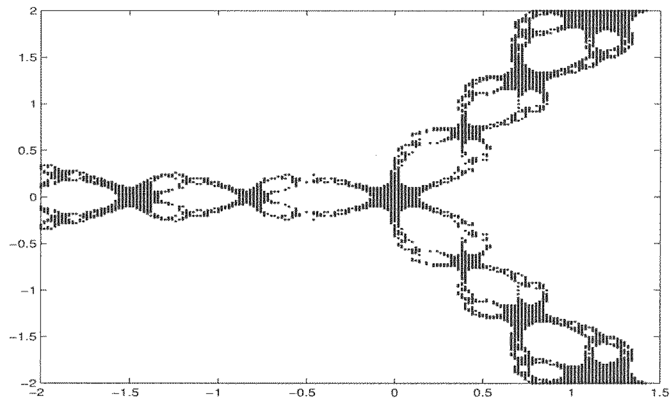

Fig. 3. Points with no convergence.

can be found in numerous books on fractals [25, 26], papers [27, 28] and web materials [29, 30] (we mention just few of the sources available). Some of these examples exhibit much more complicated behavior than for Cayley's one. For instance, one can meet periodic or chaotic trajectories of iterations.

5. Overcoming local nature of the method

In Cayley's example Newton method converged for almost all initial points (exceptions were the countable number of points S_0); the complex structure of basins of attraction was caused by the existence of several roots. However if (1) has a single root, the method usually has local convergence only. For instance, take $F(x) = \arctan x$, such function is smooth, monotone and has a single root $x^* = 0$. It is easy to check that (2) converges if and only if $|x_0| < 1$, for $|x_0| = 1$ we have periodic behavior of iterations $x_0 = -x_1 = x_2 = -x_3 \dots$, while for $|x_0| > 1$ the iterations diverge: $|x_k| \to \infty$.

There are several ways to modify basic Newton method to achieve global convergence. The first one is to introduce a regulated step-size to avoid too large steps; this is so called *damped Newton method*:

$$x_{k+1} = x_k - \alpha_k F'(x_k)^{-1} F(x_k), \quad k = 0, 1, \dots . \tag{9}$$

where step-size $0 < \alpha_k \le 1$ is chosen to achieve monotone decrease of $||F(x)||$, that is condition $||F(x_{k+1})|| < ||F(x_k)||$ holds. We will discuss the particular algorithms for choosing α_k later for minimization problems. The main goal in constructing such algorithms is to preserve a balance between convergence and rate of convergence, i.e., one should take $\alpha_k < 1$ when x_k is beyond a basin of attraction of "pure Newton method" and switch to $\alpha_k = 1$ inside this basin.

The second approach is *Levenberg–Marquardt method* [31, 32]:

$$x_{k+1} = x_k - (\alpha_k I + F'(x_k))^{-1} F(x_k), \quad k = 0, 1, \dots . \tag{10}$$

For $\alpha_k = 0$ the method converts into pure Newton, while for $\alpha_k \gg 1$ it is close to the gradient method, which usually converges globally. There are various strategies for adjusting parameters α_k, they are described, for instance, in [13]. Method (10) works even when pure Newton does not – in situation with degenerate operator $F'(x_k)$. As we shall see, for minimization problems method (10) is very promising, because it does not assume the matrix $F'(x_k)$ to be positive definite.

One more strategy is to modify Newton method to prevent large steps; it can be done not by choosing step-sizes as in damped Newton method (10), but introducing *trust region*, where linear approximation of $F(x)$ is valid. Such approach (originated in [33] and widely developed in [34]) will be discussed later in connection with optimization problems.

6. Underdetermined systems

In the above analysis of the Newton method it was assumed that the linear operator $F'(x_k)$ is invertible. However there are situations when this is definitely wrong. For instance suppose that $F : \mathbf{R}^n \to \mathbf{R}^m$, where $m < n$. That is we solve an underdetermined system of m equations with $n > m$ variables; of course a rectangular matrix $F'(x_k)$ has no inverse. Nevertheless an extension of Newton method for this case can be provided, this is due to Graves [35], who exploited the method to prove existence of solutions of nonlinear mappings. We present the result from [36] where emphasis on the method is done with more accurate estimates.

Theorem 3. *Suppose $F : X \to Y$ is defined and differentiable on a ball $B = \{x : ||x - x_0|| \le r\}$, its derivative satisfies Lipschitz condition on B*

$$||F'(x) - F'(z)|| \le L||x - z||, x, z \in B,$$

$F'(x)$ maps X onto Y and the following estimate holds:

$$||F'(x)^* y|| \ge \mu ||y|| \quad \text{for any} \quad x \in B, \quad y \in Y^*, \tag{11}$$

with $\mu > 0$ (star denotes conjugation). Introduce function

$$H_n(t) = \sum_{k=n}^{\infty} t^{2^k}$$

and suppose that

$$h = \frac{L\mu^2||F(x_0)||}{2} < 1, \quad \rho = \frac{2H_0(h)}{L\mu} \leq r. \tag{12}$$

Then the method

$$\begin{aligned} &x_{k+1} = x_k - y_k, \quad F'(x_k)y_k = F(x_k), \\ &||y_k|| \leq \mu||F(x_k)||, \quad k = 0, 1, \dots \end{aligned} \tag{13}$$

is well defined and converges to a solution x^* of the equation $F(x) = 0, ||x^* - x_0|| \leq \rho$ with the rate

$$||x_k - x^*|| \leq \frac{2H_k(h)}{L\mu}. \tag{14}$$

Thus at each iteration of the method one should solve a linear equation $F'(x_k)y = F(x_k)$, where linear operator $F'(x_k)$ in general does not have the inverse; however it maps X onto Y and this equation has a solution (may be not a single one). Among the solutions there exists the solution y_k with the property $||y_k|| \leq \mu||F(x_k)||$; this one is exploited in the method. In finite dimensional case ($X = \mathbf{R}^n, Y = \mathbf{R}^m, n > m$) such solution is provided by the formula $y_k = F'(x_k)^+ F(x_k)$, where A^+ denotes the pseudoinverse of the matrix A.

We consider an application of Theorem 3 to the convex analysis result on convexity of nonlinear image of a small ball in Hilbert spaces [37].

Theorem 4. *Suppose X, Y are Hilbert spaces, $F : X \to Y$ is defined and differentiable on a ball $B = \{x : ||x - a|| \leq r\}$, its derivative satisfies Lipschitz condition on B*

$$||F'(x) - F'(z)|| \leq L||x - z||, x, z \in B,$$

$F'(a)$ maps X onto Y and the following estimate holds:

$$||F'(a)^* y|| \geq \mu||y|| \quad \text{for any} \quad y \in Y, \tag{15}$$

with $\mu > 0$ and $r < \mu/(2L)$. Then the image of the ball B under the map F is convex, i.e., the set $S = \{F(x) : x \in B\}$ is a convex set in Y.

This theorem has numerous applications in optimization [37], linear algebra [38], optimal control [39]. For instance, the pseudospectrum of a

$n \times n$ matrix (a set of all eigenvalues of perturbed matrices for perturbations bounded in Frobenius norm) happens to be the union of n convex sets on the complex plane provided that the nominal matrix has all distinct eigenvalues and perturbations are small enough. Another result is the convexity of the reachable set of a nonlinear system for L_2 bounded controls.

7. Unconstrained optimization

Consider the simplest unconstrained minimization problem in a Hilbert space H:

$$\min f(x), \quad x \in H. \tag{16}$$

Assuming that f is twice differentiable, we can get Newton method for minimization by two different approaches.

First, necessary (and sufficient for f convex) condition for minimization is Fermat's condition

$$\nabla f(x) = 0,$$

that is, we should solve equation (3) with $F(x) = \nabla f(x)$. Applying Newton method for this equation we arrive to Newton method for minimization:

$$x_{k+1} = x_k - (\nabla^2 f(x_k))^{-1} \nabla f(x_k), \quad k = 0, 1, \dots, \tag{17}$$

where $\nabla^2 f$ denotes the second Frechet derivative (the Hessian matrix in finite dimensional case).

Second, we can approximate $f(x)$ in the neighborhood of a point x_k by three terms of its Taylor series:

$$f(x_k + h) \approx f_k(h) = f(x_k) + (\nabla f(x_k), h) + 1/2(\nabla^2 f(x_k)h, h).$$

Then, minimizing the quadratic function $f_k(h)$ we get the same method (17). Both these interpretations would be exploited for constructing Newton methods for more general optimization problems.

The theorems on convergence of Newton method for equations can be immediately adopted to unconstrained minimization case (just replace $F(x)$ by $\nabla f(x)$ and $F'(x)$ by $\nabla^2 f(x)$). The main feature of the method – fast local convergence – remains unchanged. However, there are some specific properties. The most important: Newton method in its pure form does not distinguish minima, maxima and saddle points – when started from a neighborhood of a nonsingular critical point (i.e., a point x^* with $\nabla f(x^*) = 0, \nabla^2 f(x^*)$ invertible) the method converges to it, making no difference between minimum, maximum or saddle points.

There are numerous ways to convert the method into globally convergent one, they are similar to modifications discussed in Sec. 5 (damped Newton, Levenberg–Marquardt, trust region). We consider an important version due to Nesterov and Nemirovski [40], in this version its *complexity* (the number of iterations to achieve the desired accuracy) can be estimated. Nesterov and Nemirovski introduce the class of *self-concordant* functions. These are three times differentiable convex functions defined on a convex set $D \subset \mathbf{R}^n$, satisfying the property

$$|\nabla^3 f(x)(h,h,h)| \le 2(\nabla^2 f(x)h,h)^{3/2} \quad \forall x \in D, h \in \mathbf{R}^n.$$

The above formula includes the third and second derivatives of f and their action on a vector $h \in \mathbf{R}^n$, in simpler form it can be presented as the relation between the third and second derivatives of a scalar function $\varphi(t) = f(x+th)$:

$$|\varphi'''(0)| \le 2(\varphi''(0))^{3/2}$$

for all $x \in D, h \in \mathbf{R}^n$. For instance, the function $f(x) = -\ln x, x > 0, x \in \mathbf{R}^1$ is self-concordant, while the function $f(x) = 1/x, x > 0$ is not. For $x_k \in D$ define *Newton's decrement*

$$\delta_k = \sqrt{\nabla f(x_k)^T (\nabla^2 f(x_k))^{-1} \nabla f(x_k)}.$$

Now *Nesterov–Nemirovski* version of the damped Newton method reads

$$x_{k+1} = x_k - \alpha_k (\nabla^2 f(x_k))^{-1} \nabla f(x_k), \quad k = 0, 1, \dots \tag{18}$$

$$\alpha_k = 1 \ \text{ if } \ \delta_k \le 1/4, \tag{19}$$

$$\alpha_k = 1/(1+\delta_k) \ \text{ if } \ \delta_k > 1/4. \tag{20}$$

Theorem 5. *If $f(x)$ is self-concordant and $f(x) \ge f^*$ for any $x \in D$ then for method* (18)–(20) *with $x_0 \in D, \varepsilon > 0$ one has*

$$f(x_k) - f^* \le \varepsilon$$

for

$$k = c_1 + c_2 \log\log(1/\varepsilon) + c_3(f(x_0) - f^*). \tag{21}$$

Here c_1, c_2, and c_3 are some absolute constants.

The idea of the proof is simple enough. Minimization procedure consists of two stages. At Stage 1, method (18), (20) is applied, and the function monotonically decreases: $f(x_{k+1}) \leq f(x_k) - \gamma$, where γ is a positive constant. Obviously this stage terminates after a finite number of iterations (because $f(x)$ is bounded from below). At Stage 2, pure Newton method (18), (19) works and it converges with quadratic rate: $2\delta_{k+1} \leq (2\delta_k)^2$. Note that Newton decrement provides convenient tools to express the rate of convergence: there are no constants like L, K, η in Theorems 1–3.

For the inequality (21) simple values of the constants can be determined. Note that $\log\log(1/\varepsilon)$ is not large even if ε is small enough; for all reasonable ε the following estimate holds

$$k = 5 + 11(f(x_0) - f^*).$$

However, numerous simulation results for various types of optimization problems of different dimensions [41] validate the following empirical formula for the number of steps for method (18)–(20):

$$k = 5 + 0.6(f(x_0) - f^*).$$

The serious restriction of the above analysis was the convexity assumption. Recently [42] another version of Newton method has been proposed, where global convergence and complexity results were obtained for nonnecessary convex functions. Suppose that $f \in C^{2,1}$, that is f is twice differentiable on $\mathbf{R}^n$ and its second derivative satisfies Lipschitz condition with constant L. The following version of Newton method is considered in [42]:

$$x_{k+1} = \arg\min_x f_k(x), \quad h = x - x_k, \tag{22}$$

$$f_k(x) = f(x_k) + (\nabla f(x_k), h) + \frac{1}{2}(\nabla^2 f(x_k)h, h) + \frac{L}{6}||h||^3. \tag{23}$$

Thus at each iteration we solve unconstrained minimization problem with the same quadratic term as in pure Newton, but regularized via a cubic term. This problem looks hard and nonconvex, however it can be reduced to one-dimensional convex optimization (see details in [42]). Surprisingly, the proposed method has many advantages if compared with pure Newton. First, it converges globally, for arbitrary initial point. Second, it does not converge to maximum points or saddle points in contrast with Newton method. Third, its complexity for various classes of functions can be estimated.

8. Constrained optimization

We start with some particular cases of constrained optimization problems.

The first is optimization subject to *simple constraints*:

$$\min f(x), \quad x \in Q, \tag{24}$$

where the set Q in a Hilbert space H is "simple" in the sense that (24) with quadratic function $f(x)$ can be solved explicitly. For instance, Q may be a ball, a linear subspace etc. The extension of Newton method for this case is based on the second interpretation of the method for unconstrained minimization:

$$x_{k+1} = \arg\min_{x\in Q} f_k(x), \tag{25}$$
$$f_k(x) = (\nabla f(x_k), x - x_k) + (1/2)(\nabla^2 f(x_k)(x - x_k), (x - x_k)).$$

The method converges under the same assumptions as for unconstrained case: if $f(x)$ is convex and twice differentiable with Lipschitz second derivatives on Q, Q is closed convex, $f(x)$ attains its minimum on Q in a point $x^*, \nabla^2 f(x^*) > 0$, then sequence (25) converges locally to x^* with quadratic rate. This result has been obtained in [43], see more details and examples in [44, 45].

Another simple situation is *equality constrained* optimization:

$$\min f(x), \quad g(x) = 0, \tag{26}$$

where $f : X \to \mathbf{R}^1$, $g : X \to Y$; X and Y are Hilbert spaces. If a solution x^* of the problem exists and is a regular point ($g'(x^*)$ maps X onto Y), then there exists a Lagrange multiplier y^* such that the pair x^*, y^* is a stationary point of the Lagrangian

$$L(x, y) = f(x) + (y, g(x))$$

that is the solution of the nonlinear equation

$$L'_x(x, y) = 0, \quad L'_y(x, y) = 0. \tag{27}$$

Hence we can apply Newton method for solving this equation. Under natural assumptions it converges locally to x^*, y^*, see rigorous results in [46, 45, 47]. Various implementations of the method can be also found in these references.

There are other versions of Newton method for solving (26) which do not involve dual variables y. Historically, the first application of the Newton-like method for finding the largest eigenvalue of a matrix $A = A^T$ by reducing it to constrained optimization problem:

$$\max(Ax, x), \quad ||x||^2 = 1$$

is due to Rayleigh (1899). Later, Kantorovich [3], has proposed pure Newton method for the above problem. Another simple situation, where calculations can be simplified, is the case of linear constraints $g(x) = Cx - d$. Then the method is equivalent to Newton method for unconstrained minimization of a restriction of f on the affine subspace $Cx = d$.

Now we proceed to applications of Newton method for convex constrained optimization problems – the area, where the method plays key role in constructing the most effective optimization algorithms. The basic scheme of *interior-point methods* looks as follows [40]. For the *convex optimization problem*

$$\min f(x), \quad x \in Q \tag{28}$$

with convex self-concordant f and convex $Q \in \mathbf{R}^n$ we construct a self-concordant *barrier* $F(x)$, defined on the interior of Q and growing to infinity when a point approaches the boundary of Q:

$$F : \operatorname{int} Q \to \mathbf{R}^1, \quad F(x) \to \infty \text{ for } x \to \partial Q.$$

Such barriers exist for numerous examples of the constraints, for instance, if $Q = \{x : x \geq 0\}$, then logarithmic barrier has the desired properties:

$$F(x) = \sum_{i=1}^{n} - \ln x_i.$$

Using the barriers, we take a function

$$f_k(x) = t_k f(x) + F(x)$$

depending on a parameter $t_k > 0$. It can be proved under natural assumptions that $f_k(x)$ has a minimum point x_k^* on int$Q$ and $f(x_k^*) \to f^*$ (the minimal value in (28)) when $t_k \to \infty$ (this is so called *central path*). However, there is no need to obtain precise values of x_k^*, it suffices to make one step of the damped Newton method (18)-(20) and then to vary t_k. There exists such way to adjust parameters α_k, t_k that the method yields polynomial-time complexity, see details and rigorous results in [40, 49, 48, 41]. The theoretical result on polynomial-time complexity is very important, however the practical simulation results on implementation of interior-point methods are not less important. They demonstrate very high effectiveness of the method for broad spectrum of convex optimization problems – from *linear programming* to *semidefinite programming*. For instance, the method is the successful competitor to the classical simplex method for linear programs.

The similar ideas are exploited for nonconvex constrained optimization problems (see, e.g., [34] or [50]).

9. Some extensions

Newton method has numerous extensions, we consider below just few of them.

- **Relaxed smoothness assumptions.** Newton method can be applied in many situations where equations (or functions to be minimized) are not smooth enough. Paper [51] provides typical results in this direction. The simplest example is solving equation (1) with piece-wise linear nonsmooth convex function F. Then, if a solution exists, method (2) (naturally extended) finds it after a finite number of iterations.
- **Multiple roots.** We analyzed Newton method in a neighborhood of a simple root x^*. In the case of a multiple root Newton method either remains convergent, but looses its fast rate of convergence, or diverges. There is a modification of the method, which preserves quadratic convergence; it is due to Schroder (1870). We provide it for one-dimensional case (1):

  $$x_{k+1} = x_k - pF'(x_k)^{-1}F(x_k), \quad k = 0, 1, \dots, \tag{29}$$

  here p is the multiplicity of the root. Unfortunately, we should know in advance this multiplicity, moreover situation in multidimensional case can be much more complicated.
- **Higher-order methods**. Local rate of convergence of Newton method is fast enough; however some researchers construct methods with still higher rate of convergence. This problem looks a bit artificial (actually very few iterations of Newton method are required to obtain high precision, when we achieve its convergence domain), so there is no need to accelerate the method.
- **Continuous version.** Instead of discrete iterations in Newton

method (4) one can apply its continuous analog

$$\dot{x}(t) = -F'(x(t))^{-1}F(x(t)). \tag{30}$$

The simplest difference approximation of (30) leads to damped Newton method (9) with constant step-size $\alpha_k = \alpha$. As we know, damped Newton method can exhibit global convergence, thus one can expect the same property for the continuous version. Indeed, the global convergence for (30) has been validated by S. Smale [52], and its rate of convergence is analyzed in [53]. Of course, the numerical value of continuous methods is arguable, because their implementation demands their discretization.

- **Data at one point.** All results on the convergence of Newton method include assumptions which are valid in some neighborhood of the solution or in some prescribed ball. In contrast, Smale [54] provides convergence theorem based on data available at the single initial point. However these data include bounds for all derivatives.
- **Solving complementarity and equilibrium problems**. There are many works, where Newton method is applied to problems which can not be casted into equation solving, the typical examples are complementarity, variational inequalities and equilibrium-type problems.
- **Implementation issues**. We are unable to discuss implementation issues of various versions of Newton method, which are indeed important for their practical application. Many details can be found in the recent book [15], while codes of corresponding algorithms can be downloaded from [55]. For convex optimization problems such issues are discussed in [41].
- **Complexity**. There exist very deep results on the complexity of basic problems of numerical analysis (e.g., finding all roots of a polynomial), closely related to Newton method (some modification of the method is usually proved to achieve the best possible result). The interested reader can consult [56].

References

1. L. V. Kantorovich, *On Newton's method for functional equations.* — Dokl. Akad. Nauk SSSR **59**, No. 7 (1948), 1237–1240.
2. L. V. Kantorovich, *Functional analysis and applied mathematics.* — Uspekhi Mat. Nauk **3**, No. 1 (1948), 89–185.

3. L. V. Kantorovich, *On Newton method.* — Trudy Steklov Math. Inst. **28** (1949), 104–144.
4. L. V. Kantorovich, *Principle of majorants and Newton's method.* — Dokl. Akad. Nauk SSSR **76**, No. 1 (1951), 17–20.
5. L. V. Kantorovich, *Some further applications of principle of majorants.* — Dokl. Akad. Nauk SSSR **80**, No. 6 (1951), 849–852.
6. L. V. Kantorovich, *On approximate solution of functional equations.* — Uspekhi Mat. Nauk **11**, No. 6 (1956), 99–116.
7. L. V. Kantorovich, *Some further applications of Newton method.* — Vestnik LGU, Ser. Math. Mech., No. 7 (1957), 68–103.
8. L. V. Kantorovich, G. P. Akilov, *Functional Analysis in Normed Spaces.* Fizmatgiz, Moscow (1959).
9. L. V. Kantorovich, G. P. Akilov, *Functional Analysis.* Nauka, Moscow (1977).
10. H. Fine, *On Newton's method of approximation.* — Proc. Nat. Acad. Sci. USA **2**, No. 9 (1916), 546–552.
11. A. A. Bennet, *Newton's method in general analysis.* — Proc. Nat. Acad. Sci. USA **2**, No. 10 (1916), 592–598.
12. A. M. Ostrowski, *Solution of Equations and Systems of Equations.* Academic Press, Basel (1960).
13. J. M. Ortega, W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables.* Academic Press, New York–London (1970).
14. W. C. Rheinboldt, *Methods for Solving Systems of Nonlinear Equations.* SIAM, Philadelphia (1998).
15. P. Deulfhard, *Newton Methods for Nonlinear Problems: Affine Invariant and Adaptive Algorithms.* Springer, Berlin (2004).
16. T. J. Ypma, *Historical development of the Newton–Raphson method.* — SIAM Review **37**, No. 4 (1995), 531–551.
17. J. H. Mathews, *Bibliography for Newton's method*; `math.fullerton.edu/mathews/newtonsmethod/Newton'sMethodBib/Links/Newton'sMethodBib_lnk_3.html`.
18. V. I. Arnold, *Small denominators and problem of stability in classical and celestial mechanics.* — Uspekhi Mat. Nauk **18**, No. 6 (1963), 91–192.
19. L. A. Ljusternik, *On conditional extrema of functionals.* — Math. Sb. **41**, No. 3 (1934), 390–401.
20. A. D. Ioffe, *On the local surjection property.* — Nonlinear Analysis: Theory, Methods, and Appl. **11**, No. 5 (1987), 565–592.
21. I. P. Mysovskikh, *On convergence of L. V. Kantorovich's method for functional equations and its applications.* — Dokl. Akad. Nauk SSSR **70**, No. 4 (1950), 565–568.
22. M. A. Krasnoselski, G. M. Vainikko, P. P. Zabreiko, Ya. B. Rutitski, V. Ya. Stetsenko, *Approximate Solution of Operator Equations.* Nauka, Moscow (1969).
23. L. Collatz, *Functional Analysis and Numerical Mathematics.* Academic Press, New York (1966).
24. G. Julia, *Sur l'iteration des fonctions rationelles.* — J. Math. Pure Appl. **8** (1918), 47–245.
25. M. Barnsley, *Fractals Everywhere.* Academic Press, London (1993).
26. B. Mandelbrot, *The Fractal Geometry of Nature.* W. H. Freeman, New York (1983).

27. J. H. Curry, L. Garnett, D. Sullivan, *On the iteration of a rational function: computer experiments with Newton's method.* — Comm. Math. Phys. **91** (1983), 267–277.
28. H. O. Peitgen, D. Saupe, F. Haeseler, *Cayley's problem and Julia sets.* — Math. Intellig. **6**, No. 2 (1984), 11–20.
29. D. E. Joyce, *Newton basins*; `http//aleph0.clarku.edu/djoyce/newton/newton.html`.
30. R. M. Dickau, *Newton's method*; `http//mathforum.org/advanced/robertd/newnewton.html`.
31. K. Levenberg, *A method for the solution of certain nonlinear problems in least squares.* — Quart. Appl. Math. **2** (1944), 164–168.
32. D. Marquardt, *An algorithm for least squares estimation of nonlinear parameters.* — SIAM J. Appl. Math. **11** (1963), 431–441.
33. S. Goldfeld, R. Quandt, H. Trotter, *Maximization by quadratic hill climbing.* — Econometrica **34** (1966), 541–551.
34. A. B. Conn, N. I. M. Gould, Ph. L. Toint, *Trust Region Methods.* SIAM, Philadelphia (2000).
35. L. M. Graves, *Some mapping theorems.* — Duke Math. J. **17** (1950), 111–114.
36. B. T. Polyak, *Gradient methods for solving equations and inequalities.* — USSR Comp. Math. and Math. Phys. **4**, No. 6 (1964), 17–32.
37. B. T. Polyak, *Convexity of nonlinear image of a small ball with applications to optimization.* — Set-Valued Analysis **9**, No. 1/2 (2001), 159–168.
38. B. T. Polyak, *The convexity principle and its applications.* — Bull. Braz. Math. Soc. **34**, No. 1 (2003), 59–75.
39. B. T. Polyak, *Convexity of the reachable set of nonlinear systems under L_2 bounded controls.* — Dynamics Continuous Discrete Impulsive Systems **11**, Nos. 2–3 (2004), 255–268.
40. Yu. Nesterov, A. Nemirovski, *Interior-Point Polynomial Algorithms in Convex Programming.* SIAM, Philadelphia (1994).
41. S. Boyd, L. Vandenberghe, *Convex Optimization.* Cambridge Univ. Press, Cambridge (2004).
42. Yu. Nesterov, B. Polyak, *Cubic regularization of a Newton scheme and its global performance.* — CORE Discussion Papers, No. 41 (2003), submitted to Math. Progr.
43. E. S. Levitin, B. T. Polyak, *Constrained minimization methods.* — USSR Comp. Math. and Math. Phys. **6**, No. 5 (1966), 1–50.
44. B. T. Polyak, *Introduction to Optimization.* Optimization Software, New York (1987).
45. D. P. Bertsekas, *Nonlinear Programming.* Athena Scientific, Belmont (1999).
46. B. T. Polyak, *Iterative methods using Lagrange multipliers for solving extremal problems with equality-type constraints.* — USSR Comp. Math. and Math. Phys. **10**, No. 5 (1970), 42–52.
47. D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Method.* Academic Press, New York (1982).
48. A. Ben-Tal, A. Nemirovski, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications.* SIAM, Philadelphia (2001).

49. Yu. Nesterov, *Introductory Lectures on Convex Programming*. Kluwer, Boston (2004).
50. L. T. Biegler, I. E. Grossmann, Part I: *Retrospective on Optimization*, Part II: *Future Perspective on Optimization*, Preprints, Carnegie-Mellon Univ., Chem. Eng. Dept. (2004).
51. X. Chen, L. Qi, D. Sun, *Global and superlinear convergence of the smoothing Newton method and its application to general box constrained variational inequalities*. — Math. Comput. **67**, No. 222 (1998), 519–540.
52. S. Smale, *A convergent process of price adjustment and global Newton methods*. — J. Math. Econom. **3** (1976), 107–120.
53. A. G. Ramm, *Acceleration of convergence: a continuous analog of the Newton method*. — Applic. Anal. **81**, No. 4 (2002), 1001–1004.
54. S. Smale, *Newton's method estimates from data at one point*.— In: The Merging of Disciplines: New Directions in Pure, Applied and Computational Mathematics, R. Ewing, K. Gross, C. Martin (eds.), Springer (1986).
55. *Electronic library ZIB*; `http://www.zib.de/SciSoft/NewtonLib`.
56. S. Smale, *Complexity theory and numerical analysis*. — Acta Numerica **6** (1997), 523–551.

Institute for Control Science
Moscow

E-mail: boris@ipu.rssi.ru

Поступило 28 июля 2004 г.