Journal of Mathematical Sciences, Vol. 133, No. 4, 2006

NEWTON–KANTOROVICH METHOD AND ITS GLOBAL CONVERGENCE

B. T. Polyak*

UDC 519.62

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 historical perspective and focus on some recent applications and extensions of the method and some approaches to overcoming its local nature. Bibliography:* 56 *titles.*

1. Introduction

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

The paper is organized as follows. Section 2 provides the idea of the Newton method and the history of its development. Kantorovich's contribution is addressed in Sec. 3. The Newton method in its basic form possesses only local convergence; its global behavior and modifications intended for achieving global convergence are discussed in Secs. 4 and 5. The case of underdetermined systems is worth special consideration (Sec. 6). In its original form, the Newton–Kantorovich method is intended 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 the 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 the Newton method is very simple: it is linearization. Assume that $F : R^1 \to R^1$ is a differentiable function and we are solving the equation

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

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

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

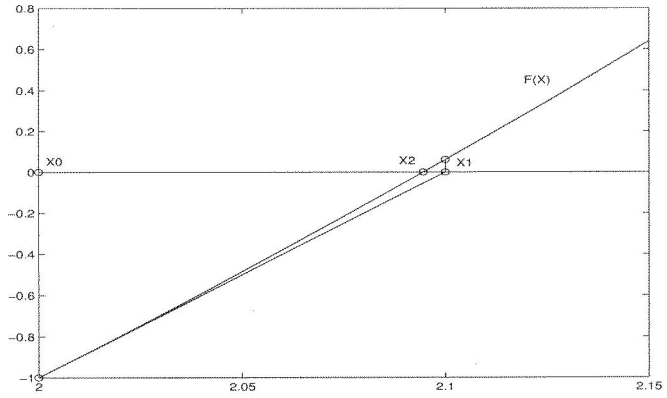

Fig. 1. Newton method.

*Institute of Control Sciences, Moscow, Russia, e-mail: boris@ipu.rssi.ru.

Published in *Zapiski Nauchnykh Seminarov POMI*, Vol. 312, 2004, pp. 256–274. Original article submitted July 28, 2004.

This is the method proposed by Newton in 1669. To be more precise, Newton dealt only with polynomials; in the expression for $F(x+h)$ he discarded higher order terms in h. The method was illustrated by the example of $F(x) = x^3 - 2x - 5 = 0$. The initial approximation for the root is $x = 2$. Then $F(2+h) = h^3 + 6h^2 + 10h - 1$; neglecting higher order terms, Newton obtains the 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.

It was J. Raphson who proposed in 1690 the general form of method (2) (not assuming $F(x)$ to be a polynomial and using the notion of derivative); that is why the method is often called the *Newton–Raphson method*.

The progress in development of the method is associated with famous mathematicians, such as Fourier, Cauchy, and others. For instance, Fourier in 1818 proved that the method converges quadratically in a neighborhood of a root, while Cauchy (1829, 1847) provided a 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 the same volume of *Proc. Nat. Acad. Sci. USA* in 1916. Fine proved the convergence in the n-dimensional case with no assumption on the existence of a solution. Bennet extended the result to the infinite-dimensional case; this was a surprising attempt, because at that time the foundations of functional analysis were not yet created. Basic results on the 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 at the special web site devoted to the 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 and Y are Banach spaces. The proposed method reads as (2):

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

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

Theorem 1. *Assume that 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 $x^* \in B$, the 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 is not technicalities, but the general formulation of the problem and the use of appropriate techniques of functional analysis. Until Kantorovich's papers [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 that it does not assume the existence of a solution, so that the theorem is not only a convergence result for a specific method, but simultaneously an 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 – can be put in the framework of (3), and the method (4) can be applied to them. Various examples of such applications are presented in the monographs [8, 9]. Moreover, the Newton–Kantorovich method, as a tool for obtaining existence results, was immediately used in the classical works by 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; a typical example is Lusternik's theorem on tangent spaces (see the original paper [19] or modern studies [20]).

Later, Kantorovich [4, 5] obtained another proof of Theorem 1 and its versions, based on the so-called "method of majorants." The idea is to compare iterations (4) with scalar iterations that in a sense majorize them and that possess convergence properties. This approach provides more flexibility; the original proof is a particular case corresponding to 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. *Assume that F is defined and twice continuously differentiable on a ball $B = \{x : ||x - x_0|| \le r\}$, the linear operator $F'(x)$ is invertible on B, $||F'(x)^{-1}|| \le \beta$, $||F''(x)|| \le K, x \in B$, $||F(x_0)|| \le \eta$, and*

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

Then Eq. (3) *has a solution $x^* \in B$ and the 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 the assumption that $F'(x)$ is invertible on B (while in Theorem 1 it was assumed to be invertible only at the initial point x_0) and a weaker 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

Conditions (5) or (7) are critical for convergence. 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 a solution. Thus the Newton–Kantorovich method is locally convergent. Very simple one-dimensional examples demonstrate the absense of global convergence even for smooth monotone functions $F(x)$. There are many ways to modify the method so as to achieve global convergence (we will discuss them later), but a 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 a set is called a *basin of attraction*) while trajectories starting outside S do not converge (e.g., tend to infinity). However, in the case of nonunique solution the structure of basins of attraction may be very complicated and exhibit fractal nature. It was Cayley who formulated this problem as early as in 1879.

Let us consider Cayley's example: let us solve the equation $z^3 = 1$ by the 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 mentioning that we formulated the 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:

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

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

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

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, i.e., points z_0 such that $z_1 = 0$; namely, $-\rho$, $\rho(1/2 \pm i\sqrt{3}/2)$, where $\rho = 1/\sqrt[3]{2}$. Each of them again has three preimages, and so on. Thus there are 3^k points z_0 that are mapped to the point $z_k = 0$ after k iterations, and they generate a countable set S_0 of initial points such that the method started from them fails at some iteration:

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

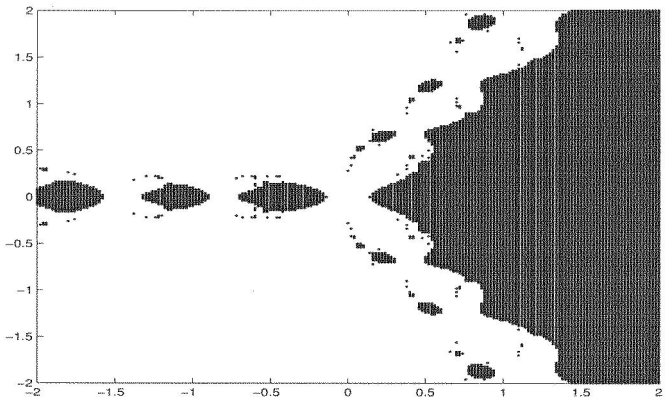

FIG. 2. Basin of attraction for $x^* = 1$.

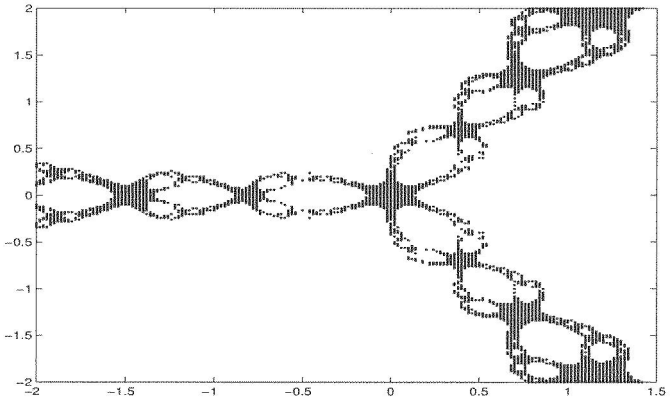

FIG. 3. Points with no convergence.

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 shown in Fig. 2, and the set S_0, in Fig. 3.

The sets of initial points that possess no convergence (like the set S_0) for iterations of general rational maps were studied by Julia [24] and are now called *Julia sets*. A lot of examples related to the Newton method can be found in numerous books on fractals [25, 26], papers [27, 28], and web materials [29, 30] (we mention just few of the available sources). Some of these examples exhibit much more complicated behavior than Cayley's one. For instance, one can encounter periodic or chaotic trajectories of iterations.

5. Overcoming the local nature of the method

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

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

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

where the step size α_k, $0 < \alpha_k \le 1$, is chosen so that $||F(x)||$ decrease monotonically, that is, $||F(x_{k+1})|| < ||F(x_k)||$. Below we will discuss particular algorithms for choosing α_k for minimization problems. The main goal in constructing such algorithms is to maintain a balance between convergence and the rate of convergence; that is, we should take $\alpha_k < 1$ when x_k is outside a basin of attraction of the "pure Newton method" and switch to $\alpha_k = 1$ inside this basin.

The second approach is the *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 $\alpha_k = 0$ it turns into the pure Newton method, 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]. The method (10) works even when the pure Newton method does not: if the operator $F'(x_k)$ is degenerate. As we will see, the method (10) is very promising for minimization problems, because it does not assume the matrix $F'(x_k)$ to be positive definite.

One more strategy is to modify the Newton method so as to prevent large steps; this can be done not by choosing step sizes as in the damped Newton method (9), but by introducing a *trust region*, where the linear approximation of $F(x)$ is valid. This 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, the rectangular matrix $F'(x_k)$ has no inverse. Nevertheless, an extension of the Newton method to this case can be provided; it is due to Graves [35], who used the method to prove the existence of solutions of nonlinear mappings. We present a result from the paper [36], where special emphasis is placed on the method itself and more accurate estimates are given.

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

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

$F'(x)$ 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 $\mu > 0$ (the star denotes conjugation). Introduce the function

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

and suppose that

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

Then the method

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

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

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

Thus at each iteration of the method one should solve the linear equation $F'(x_k)y = F(x_k)$, where the linear operator $F'(x_k)$ in general has no inverse; however, it maps X onto Y, and this equation has a solution (maybe not unique). Among the solutions there exists a solution y_k with the property $||y_k|| \le \mu ||F(x_k)||$; this one is used in the method. In the finite-dimensional case ($X = \mathbf{R}^n$, $Y = \mathbf{R}^m$, $n > m$), such a solution is provided by the formula $y_k = F'(x_k)^+ F(x_k)$, where A^+ denotes the pseudoinverse of a matrix A.

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

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

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

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

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

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 convex in Y.

This theorem has numerous applications in optimization [37], linear algebra [38], optimal control [39]. For instance, the pseudospectrum of an $n \times n$ matrix (the set of all eigenvalues of perturbed matrices for perturbations bounded in the Frobenius norm) turns out to be the union of n convex sets on the complex plane provided that all eigenvalues of the nominal matrix are distinct 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 obtain the Newton method for minimization by two different approaches.

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

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

that is, we should solve Eq. (3) with $F(x) = \nabla f(x)$. Applying the Newton method to this equation, we arrive at the 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 Fréchet derivative (the Hessian matrix in the finite-dimensional case).

Second, we can approximate $f(x)$ in a 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 obtain the same method (17). Both these interpretations can be used to construct Newton methods for more general optimization problems.

The theorems on convergence of the Newton method for equations can immediately be adjusted to the case of unconstrained minimization (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 one is as follows: the Newton method in its pure form does not distinguish minima, maxima, and saddle points: having started from a neighborhood of a nonsingular critical point (i.e., a point x^* with $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ invertible), the method converges to it, making no difference between minima, maxima, or saddle points.

There are numerous ways to convert the method into a globally convergent one; they are similar to the 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, the *complexity* of the method (the number of iterations to achieve the desired accuracy) can be estimated. Nesterov and Nemirovski introduce the class of *self-concordant* functions. These are thrice differentiable convex functions defined on a convex set $D \subset \mathbf{R}^n$ that satisfy the property

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

The above formula involves the third and second derivatives of f and their action on a vector $h \in \mathbf{R}^n$; in a simpler form, it can be presented as a relation between the third and second derivatives of the 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 $f(x) = -\log x$, $x > 0$, $x \in \mathbf{R}^1$, is self-concordant, while the function $f(x) = 1/x$, $x > 0$, is not. For $x_k \in D$ define the *Newton decrement*

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

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

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

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

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

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

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

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. The minimization procedure consists of two stages. At Stage 1, one applies the method (18), (20), and the function monotonically decreases: $f(x_{k+1}) \le f(x_k) - \gamma$, 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, one applies the pure Newton method (18), (19), and it converges with quadratic rate: $2\delta_{k+1} \le (2\delta_k)^2$. Note that the Newton decrement provides a convenient tool for expressing the rate of convergence: the formulation involves no constants like L, K, η in Theorems 1–3.

One can determine simple values of the constants for (21). 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 in the method (18)–(20):

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

A serious restriction in the above analysis is the convexity assumption. Recently ([42]), another version of the Newton method was proposed, where global convergence and complexity results were obtained for functions that are not necessarily convex. Assume that $f \in C^{2,1}$, that is, f is twice differentiable on $\mathbf{R}^n$ and its second derivative satisfies the Lipschitz condition with constant L. In [42], the following version of the Newton method is considered:

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

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

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

8. Constrained optimization

We begin with some particular cases of constrained optimization problems.

The first one is optimization subject to *simple constraints*:

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

where Q is a set in a Hilbert space H that 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 the Newton method to 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 in the unconstrained case: if $f(x)$ is convex and twice differentiable with Lipschitz second derivatives on Q, Q is a closed convex set, $f(x)$ attains its minimum on Q at a point $x^*, \nabla^2 f(x^*) > 0$, then the sequence (25) converges locally to x^* with quadratic rate. This result was obtained in [43]; for more details and examples, see [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, a solution of the nonlinear equation

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

Hence we can apply the Newton method to 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 also be found in these papers.

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

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

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

Now we proceed to applications of the Newton method to convex constrained optimization problems – the area where the method plays a 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 as the point approaches the boundary of Q:

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

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

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

Using the barrier, we take the function

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

dependent 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)) as $t_k \to \infty$ (this is the so-called *central path*). However, there is no need to obtain precise values of x_k^*, it suffices to perform one step of the damped Newton method (18)–(20) and then to vary t_k. There exists a way of adjusting the parameters α_k, t_k so that the method has polynomial-time complexity; for details and rigorous results, see [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 no less important. They demonstrate very high efficiency of the method for broad spectrum of convex optimization problems – from *linear programming* to *semidefinite programming*. For instance, the method is a successful competitor to the classical simplex method for linear programs.

Similar ideas are used for nonconvex constrained optimization problems (see, e.g., [34] or [50]).

9. Some extensions

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

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

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

  where p is the multiplicity of the root. Unfortunately, we must know this multiplicity in advance; moreover, the situation in the multidimensional case can be much more complicated.
- **Higher-order methods**. The local rate of convergence of the 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 the Newton method are required to obtain high precision when we reach its convergence domain, so there is no need to accelerate the method).
- **Continuous version.** Instead of discrete iterations in the Newton method (4), we 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 the damped Newton method (9) with constant step size $\alpha_k = \alpha$. As we know, the damped Newton method can exhibit global convergence, thus we can expect the same property for the continuous version. Indeed, the global convergence for (30) was proved by S. Smale [52], and its rate of convergence was analyzed in [53]. Of course, the value of continuous methods for numerical calculations is arguable, because an implementation of such methods requires their discretization.
- **Data at one point.** All results on the convergence of the Newton method include assumptions that are valid in some neighborhood of a solution or in some prescribed ball. Contrary to this, Smale [54] provides a convergence theorem based on data available at the single initial point. However, these data involve bounds on all derivatives.
- **Solving complementarity and equilibrium problems**. There are many papers where the Newton method is applied to problems that cannot be reduced to solving equations; typical examples are complementarity, variational inequalities, and equilibrium-type problems.
- **Implementation issues**. We are unable to discuss implementation issues of various versions of the Newton method, which are indeed important for their practical application. Many details can be found in the recent book [15], while the codes of the 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 the Newton method (some modification of the method is usually proved to yield 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**, 1237–1240 (1948).
2. L. V. Kantorovich, "Functional analysis and applied mathematics," *Uspekhi Mat. Nauk*, **3**, 89–185 (1948).
3. L. V. Kantorovich, "On Newton method," *Trudy Steklov Math. Inst.*, **28**, 104–144 (1949).
4. L. V. Kantorovich, "Principle of majorants and Newton's method," *Dokl. Akad. Nauk SSSR*, **76**, 17–20 (1951).
5. L. V. Kantorovich, "Some further applications of principle of majorants," *Dokl. Akad. Nauk SSSR*, **80**, 849–852 (1951).
6. L. V. Kantorovich, "On approximate solution of functional equations," *Uspekhi Mat. Nauk*, **11**, 99–116 (1956).
7. L. V. Kantorovich, "Some further applications of Newton method," *Vestn. LGU, Ser. Math. Mech.*, No. 7, 68–103 (1957).
8. L. V. Kantorovich and G. P. Akilov, *Functional Analysis in Normed Spaces* [in Russian], Fizmatgiz, Moscow (1959).
9. L. V. Kantorovich and G. P. Akilov, *Functional Analysis* [in Russian], Nauka, Moscow (1977).
10. H. Fine, "On Newton's method of approximation," *Proc. Nat. Acad. Sci. USA*, **2**, 546–552 (1916).
11. A. A. Bennet, "Newton's method in general analysis," *Proc. Nat. Acad. Sci. USA*, **2**, 592–598 (1916).
12. A. M. Ostrowski, *Solution of Equations and Systems of Equations*, Academic Press, Basel (1960).
13. J. M. Ortega and 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**, 531–551 (1995).
17. J. H. Mathews, *Bibliography for Newton's method*; `http://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**, 91–192 (1963).
19. L. A. Lusternik, "On conditional extrema of functionals," *Mat. Sb.*, **41**, 390–401 (1934).
20. A. D. Ioffe, "On the local surjection property," *Nonlinear Anal.*, **11**, 565–592 (1987).
21. I. P. Mysovskikh, "On convergence of L. V. Kantorovich's method for functional equations and its applications," *Dokl. Akad. Nauk SSSR*, **70**, 565–568 (1950).
22. M. A. Krasnoselski, G. M. Vainikko, P. P. Zabreiko, Ya. B. Rutitski, and V. Ya. Stetsenko, *Approximate Solution of Operator Equations* [in Russian], 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**, 47–245 (1918).
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, and D. Sullivan, "On the iteration of a rational function: computer experiments with Newton's method," *Comm. Math. Phys.*, **91**, 267–277 (1983).
28. H. O. Peitgen, D. Saupe, and F. Haeseler, "Cayley's problem and Julia sets," *Math. Intelligencer*, **6**, 11–20 (1984).
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**, 164–168 (1944).
32. D. Marquardt, "An algorithm for least squares estimation of nonlinear parameters," *SIAM J. Appl. Math.*, **11**, 431–441 (1963).
33. S. Goldfeld, R. Quandt, and H. Trotter, "Maximization by quadratic hill climbing," *Econometrica*, **34**, 541–551 (1966).

34. A. B. Conn, N. I. M. Gould, and Ph. L. Toint, *Trust Region Methods*, SIAM, Philadelphia (2000).
35. L. M. Graves, "Some mapping theorems," *Duke Math. J.*, **17**, 111–114 (1950).
36. B. T. Polyak, "Gradient methods for solving equations and inequalities," *USSR Comp. Math. Math. Phys.*, **4**, 17–32 (1964).
37. B. T. Polyak, "Convexity of nonlinear image of a small ball with applications to optimization," *Set-Valued Anal.*, **9**, 159–168 (2001).
38. B. T. Polyak, "The convexity principle and its applications," *Bull. Braz. Math. Soc.*, **34**, 59–75 (2003).
39. B. T. Polyak, "Convexity of the reachable set of nonlinear systems under L_2 bounded controls," *Dyn. Contin. Discrete Impuls. Syst.*, **11**, 255–268 (2004).
40. Yu. Nesterov and A. Nemirovski, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM, Philadelphia (1994).
41. S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge Univ. Press, Cambridge (2004).
42. Yu. Nesterov and 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 and B. T. Polyak, "Constrained minimization methods," *USSR Comp. Math. Math. Phys.*, **6**, 1–50 (1966).
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. Math. Phys.*, **10**, 42–52 (1970).
47. D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Method*, Academic Press, New York (1982).
48. A. Ben-Tal and 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 and I. E. Grossmann, Part I: "Retrospective on Optimization," Part II: "Future Perspective on Optimization," Preprints, Carnegie-Mellon Univ. (2004).
51. X. Chen, L. Qi, and D. Sun, "Global and superlinear convergence of the smoothing Newton method and its application to general box constrained variational inequalities," *Math. Comp.*, **67**, 519–540 (1998).
52. S. Smale, "A convergent process of price adjustment and global Newton methods," *J. Math. Econom.*, **3**, 107–120 (1976).
53. A. G. Ramm, "Acceleration of convergence: a continuous analog of the Newton method," *Appl. Anal.*, **81**, 1001–1004 (2002).
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, and C. Martin (eds.), Springer (1986).
55. Electronic library ZIB; `http://www.zib.de/SciSoft/NewtonLib`.
56. S. Smale, "Complexity theory and numerical analysis," *Acta Numer.*, **6**, 523–551 (1997).