SOME METHODS OF SPEEDING UP THE CONVERGENCE OF ITERATION METHODS*

B. T. POLYAK

(Moscow)

(*Received* 26 *November* 1962)

For the solution of the functional equation

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

(where P is an operator, usually linear, from B into B, and B is a Banach space) iteration methods are generally used. These consist of the construction of a series $x^0, \ldots, x^n, \ldots$, which converges to the solution (see, for example [1]). Continuous analogues of these methods are also known, in which a trajectory $x(t)$, $0 \leqslant t \leqslant \infty$ is constructed, which satisfies the ordinary differential equation in B and is such that $x(t)$ approaches the solution of (1) as $t \to \infty$ (see [2]). We shall call the method a k-step method if for the construction of each successive iteration x^{n+1} we use k previous iterations $x^n, \ldots, x^{n-k+1}$. The same term will also be used for continuous methods if $x(t)$ satisfies a differential equation of the k-th order or k-th degree. Iteration methods which are more widely used are one-step (e.g. methods of successive approximations). They are generally simple from the calculation point of view but often converge very slowly. This is confirmed both by the evaluation of the speed of convergence and by calculation in practice (for more details see below). Therefore the question of the rate of convergence is most important. Some multistep methods, which we shall consider further, which are only slightly more complicated than the corresponding one-step methods, make it possible to speed up the convergence substantially. Note that all the methods mentioned below are applicable also to the problem of minimizing the differentiable functional $f(x)$ in Hilbert space, so long as this problem reduces to the solution of the equation grad $f(x) = 0$.

---

* *Zh. Vych. Mat.*, 4, No. 5, 791-803, 1964.

1. The convergence of multistep methods

We shall begin with some supplementary statements, related to the spectral theory of operators in Banach space B. We shall denote by $\sigma(T)$ the spectrum of the linear operator T from B into B (all linear operators are assumed to be bounded).

Lemma 1. Let T_1 and T_2 be commutative operators. Then if $\lambda \in \sigma(T_1 + T_2)$, there exist $\lambda_1 \in \sigma(T_2)$ and $\lambda_2 \in \sigma(T_2)$ such that $\lambda = \lambda_1 + \lambda_2$. If $\lambda \in \sigma(T_1T_2)$, there exist $\lambda_1 \in \sigma(T_1)$ and $\lambda_2 \in \sigma(T_2)$ such that $\lambda = \lambda_1\lambda_2$.

Proof. Let us consider a commutative normalized ring, generated by the operators T_1, T_2 and I, their resolvents $(T_1 - \mu_1 I)^{-1}$, $(T_2 - \mu_2 I)^{-1}$ for all $\mu_1 \notin \sigma(T_1)$, $\mu_2 \notin \sigma(T_2)$ and the resolvents of their sum $(T_1 + T_2 - \mu I)^{-1}$ for all $\mu \notin \sigma(T_1 + T_2)$. Here and further on I is the unit operator. In such a ring the spectrum of T as an operator coincides with the spectrum of T as an element of the ring for T equal to T_1 or T_2 or $T_1 + T_2$. Applying to these operators the results [3] (viz. Theorem 4, p.32, and property (a) p.30) we obtain the first statement of the lemma. Similarly if we include in the ring the resolvents $(T_1T_2 - \mu I)^{-1}$, $\mu \notin \sigma(T_1T_2)$ and use property (b) p.30 (see [3]) the second statement of the lemma is obtained.

Lemma 2. Let $T_1, \ldots, T_k$ be mutually commutative operators, and $P(\lambda_1, \ldots, \lambda_k)$ polynomials in $\lambda_1, \ldots, \lambda_k$. Then if $\lambda \in \sigma(P(T_1, \ldots, T_k))$, there exist $\lambda_1 \in \sigma(T_1), \ldots, \lambda_k \in \sigma(T_k)$ such that $\lambda = P(\lambda_1, \ldots, \lambda_k)$.

This lemma follows directly from Lemma 1.

Let us consider the space $B^k = B \times B \times \ldots \times B$ with elements $X = (x_1, \ldots, x_k)$, $x_1 \in B, \ldots, x_k \in B$; B^k itself becomes a Banach space if we introduce the norm

$$\|X\| = \Big(\sum_{i=1}^{k} \|x_i\|^2 \Big)^{1/2}.$$

Let T be a linear operator from B^k into B^k, given by the operational matrix $((T_{ij}))$, $i, j = 1, \ldots, k$, where T_{ij} are linear operators from B into B. In other words

$$TX = Y, \qquad Y = (y_1, \ldots, y_k), \qquad y_i = \sum_{j=1}^{k} T_{ij}x_j.$$

Let us assume that all T_{ij} are mutually commutative. We shall denote $|T|$

the "determinant" of T – the linear operator from B into B, formed from T_{ij} by the same rule as in the determinant of a numerical matrix.

Lemma 3. If $\lambda \in \sigma(T)$, then $0 \in \sigma(|T - \lambda I|)$.

Proof. If $0 \notin \sigma(|T - \lambda I|)$, then $|T - \lambda I|^{-1}$ exists. Then the operational matrix $S = ((S_{ij}))$, $i, j = 1, \ldots, k$, where $S_{ij} = |T - \lambda I|^{-1} A_{ij}$, and A_{ij} is the cofactor of the element $T_{ij} - \delta_{ij}\lambda I$ (i.e. the linear operator from B into B, obtained from the elements $T - \lambda I$ by the same rule as the cofactor for a numerical matrix), is the inverse of the operator $T - \lambda I$. This is verified in the same way as for the case of numerical matrices. But this is impossible since, according to the assumption, $\lambda \in \sigma(T)$, i.e. $(T - \lambda I)^{-1}$ does not exist.

Lemma 4. Let T be the same as above, $\lambda \in \sigma(T)$. Then we can find $\lambda_{ij} \in \sigma(T_{ij})$, $i, j = 1, \ldots, k$, such that λ is an eigenvalue of the numerical matrix $\Lambda = ((\lambda_{ij}))$, $i, j = 1, \ldots, k$.

Proof. Since $\lambda \in \sigma(T)$, according to Lemma 3, $0 \in \sigma(|T - \lambda I|)$. But $|T - \lambda I|$ is a polynomial in the operators T_{ij} and so, applying Lemma 2 and using the determinant $|T|$, we arrive at the conclusion that $\lambda_{ij} \in \sigma(T_{ij})$, $i, j = 1, \ldots, k$, can be found such that $|\Lambda - \lambda I| = 0$, where $\Lambda = ((\lambda_{ij}))$, $i, j = 1, \ldots, k$, and $|\Lambda - \lambda I|$ is the determinant of this numerical matrix. This proves the lemma.

In the case where all T_{ij} are functions of one operator Lemma 4 can be made more precise.

Lemma 5. Let A be a linear operator from B into B, $T_{ij} = f_{ij}(A)$ and $f_{ij}(\lambda)$ analytic functions in some neighbourhood of the spectrum of A, $T = ((T_{ij}))$, $i, j = 1, \ldots, k$ and $\mu \in \sigma(T)$. Then there exists $\lambda \in \sigma(A)$ such that μ is an eigenvalue of the numerical matrix $((f_{ij}(\lambda)))$, $i, j = 1, \ldots, k$.

The proof is exactly the same as for Lemma 4 but instead of Lemma 2 the theorem of spectral transformation must be used ([4], p. 607).

Using Lemmas 4 and 5 we can evaluate $\|T^n\|$ and $\|e^{tT}\|$, which we shall need later on.

Lemma 6. Let

$$T = ((T_{ij})), \quad i, j = 1, \ldots, k, \text{ and } \sup_{\{\lambda_{ij} \in \sigma(T_{ij})\}} \max_{1 \leqslant s \leqslant k} |\rho_s| \leqslant q,$$

where $\rho_1, \ldots, \rho_k$ are eigenvalues of the matrix $\Lambda = ((\lambda_{ij}))$, $i, j = 1, \ldots, k$.

Then $\|T^n\| \leqslant C(\varepsilon)(q+\varepsilon)^n$, $\varepsilon > 0$ is arbitrary. If T is the same as in Lemma 5 and

$$\sup_{\lambda \in \sigma(A)} \max_{1 \leqslant s \leqslant k} |\tau_s| \leqslant q,$$

where $\tau_1, \ldots, \tau_k$ are eigenvalues of the matrix $((f_{ij}(\lambda)))$, $i, j = 1, \ldots, k$ then the same evaluation holds.

Proof. From Lemma 4 (or 5) and the given assumptions it follows that $|\mu| \leqslant q$ for all $\mu \in \sigma(T)$. But

$$\lim_{n \to \infty} \|T^n\|^{1/n} = \sup_{\mu \in \sigma(T)} |\mu|$$

([4], p. 607), i.e.

$$\lim_{n \to \infty} \|T^n\|^{1/n} \leqslant q,$$

which proves the lemma.

Lemma 7. Let

$$T = ((T_{ij})), \qquad i, j = 1, \ldots, k, \quad \text{and} \quad \sup_{\{\lambda_{ij} \in \sigma(T_{ij})\}} \max_{1 \leqslant s \leqslant k} \operatorname{Re} \rho_s \leqslant r.$$

Then $\|e^{tT}\| \leqslant C(\varepsilon)\, e^{t(r+\varepsilon)}$. This evaluation is true if T is the same as in Lemma 5 and

$$\sup_{\lambda \in \sigma(A)} \max_{1 \leqslant s \leqslant k} \operatorname{Re} \tau_s \leqslant r.$$

Here ρ_s and τ_s are the same as in Lemma 6; $\varepsilon > 0$ is arbitrary.

Proof. From Lemma 4 (or 5) and the given assumptions it follows that $\operatorname{Re} \mu \leqslant r$ for all $\mu \in \sigma(T)$. But

$$\lim_{t \to \infty} \frac{1}{t} \log \|e^{tT}\| = \sup_{\mu \in \sigma(T)} \operatorname{Re} \mu$$

([4], p. 623), which proves the lemma.

The values obtained permit us to investigate the behaviour of the iteration sequence in B^k

$$X^{n+1} = TX^n \tag{2}$$

and the solution of the differential equation in B^k

$$\frac{dX}{dt} = TX. \tag{3}$$

Lemma 8. Let the conditions of Lemma 6 be satisfied. Then for any X^0 for the series (2)

$$\| X^n \| \leqslant C(\varepsilon) \| X^0 \| (q + \varepsilon)^n.$$

Actually, $\| X^n \| = \| T^n X^0 \| \leqslant \| T^n \| \| X^0 \|$.

Lemma 9. Let the conditions of Lemma 7 be satisfied. Then with any $X(0)$ for the solution of (3) we have the evaluation

$$\| X(t) \| \leqslant C(\varepsilon) \| X(0) \| e^{t(r+\varepsilon)}.$$

Actually, $\| X(t) \| = \| e^{tT} X(0) \| \leqslant \| e^{tT} \| \| X(0) \|$.

These results may be used to study the convergence of the iteration series

$$x^{n+1} = T_1 x^n + \cdots + T_k x^{n-k+1} \tag{4}$$

or for the solution of the equation

$$\frac{d^k x}{dt^k} = T_1 \frac{d^{k-1} x}{dt^{k-1}} + \cdots + T_{k-1} \frac{dx}{dt} + T_k x \tag{5}$$

in the original space B.

Theorem 1. Let $T_1, \ldots, T_k$ be mutually commutative linear operators from B into B, and

$$\sup_{\{\rho_i \in \sigma(T_i)\}} \max_{1 \leqslant s \leqslant k} |\rho_s| \leqslant q,$$

where $\rho_1, \ldots, \rho_k$ are the roots of the equation $\rho^k = \lambda_1 \rho^{k-1} + \ldots + \lambda_k$. Then for any $x^0, \ldots, x^{k-1}$ for the series (4)

$$\| x^n \| \leqslant C(\varepsilon) \left(\sum_{i=0}^{k-1} \| x^i \|^2 \right)^{1/2} (q + \varepsilon)^n.$$

If $T_i = f_i(A)$, A is a linear operator from B into B, $f_i(\lambda)$ are functions which are analytic in some neighbourhood of the spectrum of A and

$$\sup_{\lambda \in \sigma(A)} \max_{1 \leqslant s \leqslant k} |\tau_s| \leqslant q,$$

where $\tau_1, \ldots, \tau_k$ are the roots of the equation $\tau^k = f_1(\lambda)\tau^{k-1} + \ldots + f_k(\lambda)$, the same evaluation holds for $\| x^n \|$.

Proof. Consider the series $X^k \in B^k$, where $X^n = (x^n, \ldots, x^{n+k-1})$, and $x^i \in B$ are terms of series (4).

As can be verified without difficulty, the relation $X^{n+1} = TX^n$, is true, where the linear operator T from B^k into B^k is given by the operational matrix

$$T = \begin{Vmatrix} 0 & I & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \ldots & I \\ T_k & T_{k-1} & \ldots & T_1 \end{Vmatrix}.$$

For this matrix the condition of Lemma 6 is satisfied so long as

$$\Lambda = \begin{Vmatrix} 0 & 1 & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \ldots & 1 \\ \lambda_k & \lambda_{k-1} & \ldots & \lambda_1 \end{Vmatrix}$$

Its eigenvalues are the roots of the equation $\rho^k = \lambda_1 \rho^{k-1} + \ldots + \lambda_k$, and therefore the fulfilment of the conditions of the lemma follows from the assumption of the first part of the theorem. The situation is also similar for the second part of the theorem. Now using Lemma 8 we obtain

$$\| X^n \| \leqslant C(\varepsilon) \| X^0 \| (q + \varepsilon)^n, \text{ but } \| X^n \| = \left(\sum_{i=0}^{k-1} \| x^{n+i} \|^2 \right)^{1/2} \geqslant \| x^n \|,$$

which proves the theorem.

Theorem 2. Let $T_1, \ldots, T_k$ be mutually commutative linear operators from B into B and

$$\sup_{\{\lambda_i \in \sigma(T_i)\}} \max_{1 \leqslant s \leqslant k} \operatorname{Re} \rho_s \leqslant r.$$

Then for any $x(0), \ldots, d^{k-1}x(0)/dt^{k-1}$ equation (5) has a solution and

$$\left\| \frac{d^i x}{dt^i} \right\| \leqslant C(\varepsilon) \left(\sum_{j=0}^{k-1} \left\| \frac{d^j x(0)}{dt^j} \right\|^2 \right)^{1/2} e^{t(r+\varepsilon)}, \qquad i = 0, \ldots, k-1.$$

This evaluation holds if

$$T_i = f_i(A) \text{ and } \sup_{\lambda \in \sigma(A)} \max_{1 \leqslant s \leqslant k} \operatorname{Re} \tau_s \leqslant r.$$

Here ρ_s, τ_s, f_i, A and ε are the same as in Theorem 1.

Proof. Let us introduce $X(t) \in B^k$, $X(t) = (x(t), \ldots, d^{k-1}x(t)/dt^{k-1})$, where $x(t)$ satisfies (5). Then, as we can easily verify, $dX(t)/dt = TX(t)$ where T is the same operational matrix as in Theorem 1. Further proof is carried out in the same way as in Theorem 1 but with reference to Lemmas 7 and 9.

Theorems 1 and 2 permit us to prove the convergence of the linear multistep methods for the solution of the linear equation

$$Ax = b. \tag{6}$$

Theorem 3. Let $T_1, \ldots, T_k, A$ be mutually commutative operators from B into B and $\alpha_0, \ldots, \alpha_k$ numbers, such that $\sum_{i=0}^{k-1} \alpha_i = 1$. Let

$$\sup_{\lambda \in \sigma(A),\ \{\lambda_i \in \sigma(T_i)\}} \max_{1 \leqslant s \leqslant k} |\rho_s| \leqslant q < 1,$$

where $\rho_1, \ldots, \rho_k$ are the roots of the equation

$$\rho^k = \sum_{i=0}^{k-1} (\alpha_i + \lambda_{i+1}\lambda)\,\rho^i.$$

Then the solution x^* of equation (6) exists and is unique, and for any $x^0, \ldots, x^{k-1}$ the series

$$x^{n+1} = \sum_{i=0}^{k-1} \alpha_i x^{n-i} + \sum_{i=0}^{k-1} T_{i+1}\,(Ax^{n-i} - b)$$

converges to x^* at the rate of the geometric progression

$$\|x^n - x^*\| \leqslant C\,(\varepsilon)\left(\sum_{i=0}^{k-1} \|x^i - x^*\|^2\right)^{1/2} (q + \varepsilon)^n.$$

This evaluation is true if

$$T_i = f_i\,(A) \text{ and } \sup_{\lambda \in \sigma(A)} \max_{1 \leqslant s \leqslant k} |\tau_s| \leqslant q < 1,$$

where $\tau_1, \ldots, \tau_k$ are the roots of the equation

$$\tau^k = \sum_{i=0}^{k-1} (\alpha_i + f_{i+1}\,(\lambda)\,\lambda)\,\tau^i.$$

Proof. Since $\rho = 1$ is a root of the equation

$$\rho^k = \sum_{i=0}^{k-1} \alpha_i \rho^i,$$

the assumption that $0 \in \sigma(A)$ contradicts the condition $|\rho_s| < 1$ for all $1 \leqslant s \leqslant k$, $\lambda \in \sigma(A)$, and $\lambda_i \in \sigma(T_i)$. Therefore an operator A^{-1} exists and a solution of (6) exists and is unique, $x^* = A^{-1}b$. For $y^n = x^n - x^*$ the recurrence relation

$$y^{n+1} = \sum_{i=0}^{k-1} (\alpha_i + T_{i+1}A)\, y^{n-i}.$$

holds. For this sequence the conditions of Theorem 1 are satisfied and so $\|y^n\| \leqslant C(\varepsilon)\|y^0\|(q+\varepsilon)^n$, which proves the theorem.

Note. The statement, similar to the second part of the theorem, for the case of self-conjugate operators in Hilbert space is proved in [5].

Theorem 4. Let $T_1, \ldots, T_k, A$ be mutually commutative operators from B into B and

$$\sup_{\lambda \in \sigma(A),\ \{\lambda_i \in \sigma(T_i)\}} \max_{1 \leqslant s \leqslant k} \operatorname{Re} \rho_s \leqslant r < 0,$$

where $\rho_1, \ldots, \rho_k$ are the roots of the equation $\rho^k = \lambda_1 \rho^{k-1} + \ldots + \lambda_{k-1}\rho + \lambda_k \lambda$. Then the solution x^* of equation (6) exists and is unique, and for any $x(0), \ldots, d^{k-1}x(0)/dt^{k-1}$ the solution of the differential equation

$$\frac{d^k x}{dt^k} = T_1 \frac{d^{k-1}x}{dt^{k-1}} + \cdots + T_{k-1}\frac{dx}{dt} + T_k(Ax - b)$$

exists, is unique and converges to x^*, where

$$\|x(t) - x^*\| \leqslant C(\varepsilon)\, Re^{t(r+\varepsilon)}, \qquad \left\|\frac{d^i x(t)}{dt^i}\right\| \leqslant C(\varepsilon)\, Re^{t(r+\varepsilon)}, \qquad i = 1, \ldots, k-1,$$

$$R = \left(\|x(0) - x^*\|^2 + \sum_{i=1}^{k-1} \left\|\frac{d^i x(0)}{dt^i}\right\|^2\right)^{1/2}.$$

This same evaluation holds if

$$T_i = f_i(A) \text{ and } \sup_{\lambda \in \sigma(A)} \max_{1 \leqslant s \leqslant k} \operatorname{Re} \tau_s \leqslant r < 0,$$

where $\tau_1, \ldots, \tau_k$ are the roots of the equation $\tau^k = f_1(\lambda)\tau^{k-1} + \ldots + f_{k-1}(\lambda)\tau + f_k(\lambda)\lambda$.

Proof. The assumption that $0 \in \sigma(A)$ contradicts the conditions of the theorem, so long as, in this case, $\rho = 0$ (or $\tau = 0$) is a root of the corresponding equation. Therefore a unique solution $x^* = A^{-1}b$ exists. Using the notation $y(t) = x(t) - x^*$, we obtain for $y(t)$ the differential equation

$$\frac{d^k y}{dt^k} = T_1 \frac{d^{k-1}y}{dt^{k-1}} + \cdots + T_{k-1}\frac{dy}{dt} + T_k A y.$$

The application of Theorem 2 to this equation completes the proof.

Let us now turn to the investigation of the nonlinear case. The possibility of extending the results obtained to this case follows from the lemmas below.

Lemma 10. Suppose that the linear operator T satisfies the condition $\|T^n\| \leqslant C(\varepsilon)(q + \varepsilon)^n$, $q < 1$, and $\|Y(X)\| = o(\|X\|)$. Then the series

$$X^{n+1} = TX^n + Y(X^n)$$

approaches zero with sufficiently small X^0, where $\|X^n\| \leqslant C'(\varepsilon)\|X^0\| \times (q + \varepsilon)^n$.

Lemma 11. Let $\|e^{tT}\| \leqslant C(\varepsilon)e^{t(r+\varepsilon)}$, $r < 0$, and $\|Y(X)\| = o(\|X\|)$, where $Y(X)$ satisfies a Lipschitz condition in the neighbourhood of zero. Then the solution of the equation

$$\frac{dX}{dt} = TX + Y(X)$$

exists for sufficiently small $X(0)$ and $\|X(t)\| \leqslant C'(\varepsilon)\|X(0)\|e^{t(r+\varepsilon)}$.

The proof of these lemmas is carried out in exactly the same way as one of the proofs of the known theorems of stability for the finite-dimensional case (see, for instance, [6], Chap. 4).

Note. In Lemma 11 a Lipschitz condition for $Y(X)$ is required only in order to stipulate the existence of the solution of the differential equation. In the finite-dimensional case a sufficient requirement is the continuity of $Y(x)$.

We now quote theorems on the convergence of multistep methods for the nonlinear case. These theorems are proved in exactly the same way as the corresponding Theorems 1-4 for the linear case taking Lemmas 10 and 11 into account.

Theorem 5. Let the conditions of Theorem 1 be satisfied, while $q < 1$ and $\|y_i(x)\| = o(\|x\|)$, $i = 1, \ldots, k$. Then for any sufficiently small $x^0, \ldots, x^{k-1}$ the series

$$x^{n+1} = \sum_{i=0}^{k-1} (T_{i+1}x^{n+i} + y_{i+1}(x^{n+i}))$$

converges to zero, and

$$\|x^n\| \leqslant C(\varepsilon)\left(\sum_{i=0}^{k-1}\|x^i\|^2\right)^{1/2}(q+\varepsilon)^n.$$

Theorem 6. Let the conditions of Theorem 2 be satisfied where $r < 0$, $\|y(x)\| = o(\|x\|)$ and $y(x)$ satisfies a Lipschitz condition in the neighbourhood of zero. Then the equation

$$\frac{d^k x}{dt^k} = T_1\frac{d^{k-1}x}{dt^{k-1}} + \cdots + T_{k-1}\frac{dx}{dt} + T_k x + y(x)$$

has a solution for sufficiently small $x(0), \ldots, d^{k-1}x(0)/dt^{k-1}$, where

$$\left\|\frac{d^i x(t)}{dt^i}\right\| \leqslant C(\varepsilon)\left(\sum_{j=0}^{k-1}\left\|\frac{d^j x(0)}{dt^j}\right\|^2\right)^{1/2} e^{t(r+\varepsilon)}, \qquad i = 0, \ldots, k-1.$$

Theorem 7. Let x^* be a solution of the equation $P(x) = 0$, where P is a nonlinear operator from B into B, which has a derivative $P'(x^*)$ at the point x^*. Let $T_1, \ldots, T_k$ be linear operators which are commutative among themselves and with $P'(x^*)$, and

$$\sup_{\substack{\lambda \in \sigma(P'(x^*)) \\ \{\lambda_i \in \sigma(T_i)\}}} \max_{1 \leqslant s \leqslant k} |\rho_s| \leqslant q < 1,$$

where $\rho_1, \ldots, \rho_k$ are the roots of the equation

$$\rho^k = \sum_{i=0}^{k-1}(\alpha_i + \lambda_{i+1}\lambda)\rho^i, \qquad \sum_{i=0}^{k-1}\alpha_i = 1.$$

Then for any $x^0, \ldots, x^{k-1}$, sufficiently close to x^*, the series

$$x^{n+1} = \sum_{i=0}^{k-1}\alpha_i x^{n-i} + \sum_{i=0}^{k-1} T_{i+1}P(x^{n-i}) \tag{7}$$

converges to x^*, where

$$\|x^n - x^*\| \leqslant C(\varepsilon)\left(\sum_{i=0}^{k-1}\|x^i - x^*\|^2\right)^{1/2}(q+\varepsilon)^n.$$

This is true if

$$T_i = f_i(P'(x^*)) \text{ and } \sup_{\lambda \in \sigma(P'(x^*))} \max_{1 \leqslant s \leqslant k} |\tau_s| \leqslant q < 1,$$

where $\tau_1, \ldots, \tau_k$ are the roots of the equation

$$\tau^k = \sum_{i=0}^{k-1} (\alpha_i + f_{i+1}(\lambda)\lambda)\tau^i.$$

Proof. Since $P(x)$ is differentiable at x^* and $P(x^*) = 0$, then $P(x) = P'(x^*)(x - x^*) + y(x - x^*)$ where $\|y(x - x^*)\| = o(\|x - x^*\|)$. Therefor (7) can be written as

$$x^{n+1} = \sum_{i=0}^{k-1} \alpha_i x^{n-i} + \sum_{i=0}^{k-1} T_{i+1}P'(x^*)(x^{n-i} - x^*) + \sum_{i=0}^{k-1} T_{i+1}y(x^{n-i} - x^*).$$

Introducing $z^n = x^n - x^*$ and using the fact that $\sum_{i=0}^{k-1} \alpha_i = 1$, we obtain

$$z^{n+1} = \sum_{i=0}^{k-1} (\alpha_i + T_{i+1}P'(x^*))z^{n-i} + \sum_{i=0}^{k-1} T_{i+1}y(z^{n-i}).$$

Now we only have to use Theorem 5 to complete the proof.

Theorem 8. Let $P(x^*) = 0$ and the operator $P(x)$ have a bounded derivative $P'(x)$ in some neighbourhood of x^*. Let $T_1, \ldots, T_k$ be linear operators, commutative among themselves and with $P'(x^*)$ and

$$\sup_{\substack{\lambda \in \sigma(P'(x^*)) \\ \{\lambda_i \in \sigma(T_i)\}}} \max_{1 \leqslant s \leqslant k} \operatorname{Re} \rho_s \leqslant r < 0,$$

where $\rho_1, \ldots, \rho_k$ are the roots of the equation $\rho^k = \lambda_1\rho^{k-1} + \ldots + \lambda_{k-1}\rho + \lambda_k\lambda$. Then for any sufficiently small $x(0) - x^*$, $dx(0)/dt$, $\ldots$, $d^{k-1}x(0)/dt^{k-1}$ the differential equation

$$\frac{d^k x}{dt^k} = T_1 \frac{d^{k-1}x}{dt^{k-1}} + \cdots + T_{k-1}\frac{dx}{dt} + T_k P(x) \tag{8}$$

has a solution, where

$$\|x(t) - x^*\| \leqslant C(\varepsilon) Re^{t(r+\varepsilon)}, \qquad \left\|\frac{d^i x(t)}{dt^i}\right\| \leqslant C(\varepsilon) Re^{t(r+\varepsilon)}, \qquad i = 1, \ldots, k-1,$$

$$R = \left(\|x(0) - x^*\|^2 + \sum_{i=1}^{k-1} \left\|\frac{d^i x(0)}{dt^i}\right\|^2\right)^{1/2}.$$

This is true if

$$T_i = f_i(P'(x^*)) \text{ and } \sup_{\lambda \in \sigma(P'(x^*))} \max_{1 \leqslant s \leqslant k} \operatorname{Re} \tau_s \leqslant r < 0,$$

where $\tau_1, \ldots, \tau_k$ are the roots of the equation $\tau^k = f_1(\lambda)\tau^{k-1} + \cdots + f_{k-1}(\lambda)\tau + f_k(\lambda)\lambda$.

The proof is carried out as for the above with reference to Theorem 7. The boundedness of the derivative $P'(x)$ is necessary in order to satisfy the Lipschitz condition for $y(x)$.

Note that for nonlinear problems we have obtained only local theorems on convergence. Nonlocal results of this type would be very interesting.

2. Examples, the numerical aspect

Let us now consider in more detail some two-step methods and show that they actually speed up of the convergence in comparison with the corresponding one-step methods. We shall study the method

$$x^{n+1} = x^n - \alpha P(x^n) + \beta(x^n - x^{n-1}) \tag{9}$$

and its continuous analogue

$$\frac{d^2x}{dt^2} = \alpha_1 \frac{dx}{dt} + \alpha_2 P(x). \tag{10}$$

Note that methods of the type (9) are widely used in linear algebra for speeding up the convergence of iterative methods of solving linear equations. Methods such as Lysternik's, conjugate gradients, Abramov's, Faddeyev's "general three-term iteration process", etc. all have the same form as method (9), but with variables α and β (see [7], Chap.9). Method (9) with constant α and β for problems of linear algebra was put forward by Frankel ([8], see also [9], pp.255-261).

We now mention a theorem on the convergence of methods (9) and (10). Here we limit ourselves to the case where $P(x)$ is a potential operator (i.e. the gradient of some functional) in Hilbert space. Then the one-step methods

$$x^{n+1} = x^n - \alpha P(x^n), \tag{11}$$

$$dx/dt = \alpha P(x) \tag{12}$$

are simply the gradient methods of minimizing this functional (see [10]).

Theorem 9. Let the functional $f(x)$, given in some region S of Hilbert space H, be differentiable in some neighbourhood of the point $x^* \in S$, i.e.

$$f(x + y) = f(x) + (P(x), y) + o(\| y \|),$$

where $P(x)$ is the gradient of this functional at the point x. Let $P(x^*) = 0$. Let us assume that $f(x)$ is twice differentiable at x^*, i.e.

$$f(x^* + y) = f(x) + (P(x^*), y) + \tfrac{1}{2}(Ay, y) + o(\| y \|^2),$$

where A is a self-conjugate operator from H into H. Suppose that for all $y \in H$

$$m(y, y) \leqslant (Ay, y) \leqslant M(y,y), \qquad m > 0. \tag{13}$$

Then:

(1) x^* is a local minimum point of $f(x)$;

(2) for $0 < \alpha < 2/M$, for all x^0 sufficiently near to x^*, the series (11) converges to x^*, while $\| x^n - x^* \| \leqslant C_1(\varepsilon) \| x^0 - x^* \| (q_1 + \varepsilon)^n$, $0 \leqslant q_1 < 1$, $q_1 = \max\{|1 - \alpha m|, |1 - \alpha M|\}$. The quantity q_1 is a minimum and is equal to

$$\bar{q}_1 = \frac{M - m}{M + m} \quad \text{for} \quad \alpha = \frac{2}{M + m};$$

(3) for $0 \leqslant \beta < 1$, $0 < \alpha < 2(1 + \beta)/M$ for any x^0, x^1, sufficiently near to x^*, the sequence (9) converges to x^*, where $\| x^n - x^* \| \leqslant C_2(\varepsilon) (\| x^0 - x^* \|^2 + \| x^1 - x^* \|^2)^{1/2} (q_2 + \varepsilon)^n$, $0 \leqslant q_2 < 1$. The quantity q_2 is a minimum and is given by

$$\bar{q}_2 = \frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}} \quad \text{for} \quad \alpha = \frac{4}{(\sqrt{M} + \sqrt{m})^2},$$

$$\beta = \left(\frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}} \right)^2.$$

In addition let $f(x)$ be twice differentiable in the neighbourhood of x^* and the second derivative be bounded there. Then:

(4) for any $x(0)$, sufficiently near to x^*, and any $\alpha < 0$ equation (12) has a solution, where $\| x(t) - x^* \| \leqslant C_3(\varepsilon) \| x(0) - x^* \| e^{t(r_1 + \varepsilon)}$, $r_1 = \alpha m < 0$;

(5) for any sufficiently small $x(0) - x^*$, $dx(0)/dt$ and any $\alpha_1 < 0$, $\alpha_2 < 0$, equation (10) has a solution, where $\|x(t) - x^*\| \leqslant C_4(\varepsilon) Re^{t(r_2+\varepsilon)}$,

$$R = \left(\|x(0) - x^*\|^2 + \sum_{i=1}^{k-1} \left\| \frac{d^i x(0)}{dt^i} \right\|^2 \right)^{1/2},$$

$$r_2 = \frac{\alpha_1 + \sqrt{g}}{2}, \qquad g = \max\{0, \alpha_1^2 + 4\alpha_2 m\}.$$

Proof. The first statement of the theorem is obvious. The second and fourth statements may be obtained without difficulty from the results given above, but we shall not dwell on this because these statements were known previously [10]. We turn now to the proof of the third statement. Obviously method (9) is a particular case of method (7) with $k = 2$, $\alpha_0 = 1 + \beta$, $\alpha_1 = -\beta$, $(\alpha_0 + \alpha_1) = 1$, $T_1 = -\alpha I$, $T_2 = 0$. To apply Theorem 7 it is sufficient to verify that

$$\sup_{\lambda \in \sigma(A)} \max \{|\rho_1|, |\rho_2|\} \leqslant q_2 < 1,$$

where ρ_1, ρ_2 are the roots of the equation $\rho^2 = (\alpha_0 - \alpha\lambda) + \alpha_1$, i.e. the equation $\rho^2 - (1 + \beta - \alpha\lambda)\rho - \beta = 0$ (so long as $A = P'(x^*)$). From the assumption (13) it follows that if $\lambda \in \sigma(A)$, then $m \leqslant \lambda \leqslant M$. Therefore it is sufficient to prove that

$$\sup_{m \leqslant \lambda \leqslant M} \max \{|\rho_1|, |\rho_2|\} \leqslant q_2 < 1.$$

But for the proof of convergence of method (9) for a finite-dimensional quadratic functional we must consider the same equation (see [9], pp. 255-261). It was shown there that with the assumption made about α and β in fact

$$\sup_{m \leqslant \lambda \leqslant M} \max \{|\rho_1| |\rho_2|\} \leqslant q_2 < 1.$$

There the quantity q_2 was given depending on α and β and, in particular, the minimum value $\bar{q}_2$. Thus Theorem 7 is applicable which proves the third statement of the theorem.

We shall now prove the last statement. Method (10) is a particular case of method (8) with $k = 2$, $T_1 = \alpha_1 I$ and $T_2 = \alpha_2 I$. All the conditions of Theorem 8 are satisfied and it is only necessary to prove that

$$\sup_{m \leqslant \lambda \leqslant M} \max \{\mathrm{Re}\,\rho_1, \mathrm{Re}\,\rho_2\} \leqslant r_2 < 0,$$

where ρ_1 and ρ_2 are the roots of the equation $\rho^2 = \alpha_1\rho + \alpha_2\lambda$. But

$$\rho_{1,2}=\frac{\alpha_1\pm\sqrt{\alpha_1^2+4\alpha_2\lambda}}{2}, \qquad \operatorname{Re}\rho_{1,2}=\frac{\alpha_1\pm\sqrt{g(\lambda)}}{2},$$

$$g(\lambda)=\max\{0,\ \alpha_1^2+4\alpha_2\lambda\},$$

$$\sup_{m\leqslant\lambda\leqslant M}\max\{\operatorname{Re}\rho_1,\ \operatorname{Re}\rho_2\}=\frac{\alpha_1+\sqrt{g(m)}}{2}=r_2<0.$$

Thus Theorem 8 is applicable and this completes the proof.

Note. If $f(x)$ is a quadratic functional the given methods converge with any initial values (instead of Theorems 7 and 8 we must use Theorems 3 and 4). In the case of a non-quadratic functional we can obtain nonlocal theorems for one-step methods [10]; for two-step methods we have succeeded in obtaining only the most particular results of a nonlocal type.

We shall now compare methods (9) and (11) from the point of view of calculation. Note first of all that near the minimum point the functionals, as a rule, satisfy the conditions of Theorem 9, so that the equations given by this theorem are generally applicable. As was shown above, the gradient method (11) with optimum choice of α converges as a geometric progression with common ratio $\bar{q}_1 = (M - m)/(M + m)$, and the two-step method (9) with optimum choice of parameters as a geometric progression with common ratio $\bar{q}_2 = (\sqrt{M} - \sqrt{m})/(\sqrt{M} + \sqrt{m})$. The case where $p = M/m$ is large is of considerable interest, so long as for the majority of practical functions, near the minimum, $f(x)$ varies weakly in some direction (i.e. m is small) and strongly in another (M is large). Note that in the terminology of Gelfand and Tsetlin [11] such functions are said to be well organized, the first directions essential and the second nonessential. For a quadratic functional, corresponding to this there is the case of an ill conditioned matrix (p here is the same as P, the number of the conditionality, see [7]). For large p we have $\bar{q}_1 \approx 1 - 2/p$, i.e. $\bar{q}_1$ is near unity and the rate of convergence of the gradient method is low. For the two-step method with large p we have $\bar{q}_2 \approx 1 - 2/\sqrt{p}$. Therefore, in order to decrease the deviation from the minimum point e^k times, in the gradient method we require a number of iterations of the order $n_1 = k/\ln \bar{q}_1 \approx kp/2$ but in method (9) of the order $n_2 = k/\ln \bar{q}_2 \approx k\sqrt{p}/2$. For large p we shall have $n_1 \gg n_2$, so that the speeding up in this case is very effective.

This is confirmed by qualitative considerations. Note that method (10) can be described by the same equation as the motion of a body in

a potential field (so that method (9) – a discrete analogue of (10) – can be called "the method of a small heavy sphere"). The motion proceeds not in the direction of the force (i.e. antigradient) because of the presence of inertia. The term $\beta(x^n - x^{n-1})$, giving inertia to the motion, will lead to motion along the "essential" direction, i.e. along "the bottom of the trough" [11].

From another point of view method (9) requires only slightly larger calculations at one iteration than (11).

A few words on the choice of the parameters α and β. As long as the values of m and M are generally unknown the parameters have to be chosen empirically. It is convenient to do this as follows. Initially we do the computation with $\beta = 0$, choosing the best value of α. When the rate of convergence slows down we add the term with β. Here the slower the rate of convergence the nearer to unity we must take β. Generally $\beta = 0.8 - 0.99$. Simultaneously we can increase α (since the optimum α in (9) is approximately twice as large as in (11) for large p). We must bear in mind that $f(x^n)$ in method (9) does not necessarily vary monotonically from iteration to iteration and sections of increase may occur. Only if the increase in $f(x^n)$ is stable must we decrease α and β.

We note also one merit of method (9). It will "bring in" small minima on account of the inertia. Actually, if $P(x^n)$ is small (x^n is near a local minimum), but $x^n - x^{n-1}$ is large, then $x^{n+1} - x^n$ will also be large and x^{n+1}, possibly, turn out to be inside the region of attraction of this local minimum. For this purpose (and not for speeding up convergence) a method, similar to (9), was used to minimize some concrete functions in [12].

Method (9) was tested in the solution of a number of problems. The test showed that in the majority of cases it actually gives a marked speeding up (up to tenfold) in comparison with the gradient method. At the same time method (9) frequently gives too slow a convergence and we have to use more powerful (but of course more laborious) methods, for example Newton's method.

We might similarly using the theorems of the first section also investigate other k-step methods with $k > 2$, but they would scarcely turn out to be suitable for practical calculations, so long as the speeding up which they give is low compared with the two-step method, and the difficulties of choice of parameters in these methods sharply increase. Much more important (but also more difficult) for investigating classes of methods are nonlinear and nonstationary multistep methods (i.e. methods of the form $x^{n+1} = Q_n(x^n, \ldots, x^{n-k+1})$, where Q_n does

not necessarily depend linearly on x^i and $P(x^i)$). In this area only various disconnected results are known.

Translated by H.F. Cleaves

REFERENCES

1. Kantorovich, L.V., Approximate solution of functional equations. *Uspekhi Mat. Nauk*, 11, 6, 99-116, 1956.

2. Gavurin, M.K., Nonlinear functional equations and continuous analogues of iterative methods. *Izv. VUZov. Ser. Mat.*, 5, 18-31, 1958.

3. Gel'fand, I.M., Raikov, D.A. and Shilov, G.E., *Commutative Normalized Rings* (Kommutativnye normirovannye kol'tsa). Fizmatgiz, Moscow, 1960.

4. Dunford, N. and Schwartz, C., *Linear Operators*. Interscience, 1958.

5. Buledz, A.S., The problem of speeding up the convergence of iteration processes in the approximate solution of linear operational equations. *Dop. Akad. Nauk SSSR*, 3, 265-269, 1961.

6. Bellman, R., *Stability theory of differential equations*. McGraw-Hill, 1954.

7. Faddeev, D.K. and Faddeeva, V.N., *Numerical Methods of Linear Algebra* (Vychislitelnye metodi lineinoi algebry). Fizmatgiz, Moscow, 1960.

8. Frankel, S., Convergence rates of iterative treatments of partial differential equations. Math. Tables and Other Aids Comput., 4, 65-75, 1950.

9. Sayl'ev, V.K., *Integration of Parabolic-Type Equations by the Network Method* (Integrirovanie yravnenni parabolicheskogo tipe metodom setok). Fizmatgiz, Moscow, 1960.

10. Polyak, B.T., Gradient methods for minimizing functionals. *Zh. vych. mat.*, 3, 4, 643-654, 1963.

11. Gel'fand, I.M. and Tsetlin, M.L., Principles of nonlocal search in systems of automatic optimization. *Dokl. Akad. Nauk SSSR*, 137, 2, 295-298, 1961.

12. Inomata and Cumada. On the golf method. *Bu l. Electrotec. Lab.*, 25, 7, 495-512, 1961.