Amer. Math. Soc. Transl.
(2) Vol. 178, 1996

On a Class of Polynomial Systems of Equations Following from the Formula for Total Probability and Possibilities for Eliminating Search in Solving Them

An. A. Muchnik

ABSTRACT. A generalization of linear systems of equations for stable states of discrete random processes is studied. The effectiveness of two classical methods of approximating a solution is examined.

1. Introduction

This paper is directed towards those readers who are meeting in practice with the problem considered below. The purpose is to give the most complete and independent exposition possible, requiring only minimal prior knowledge.

The mathematical model we shall be investigating was proposed by Petrovskiĭ. The result of Section 2 belongs to Vysotskiĭ. It is based on Brouwer's theorem, whose proof relies on Sperner's formula. The main arguments in Section 4 are similar to the approximation of continuous functions by polynomials using Bernstein's method. In Sections 5 and 6 we study conditions determining the convergence rate for the methods of iterations and descent. The computation of the number of ladders in Section 8 is an exercise in combinatorics.

Various information related to this subject can be found in [1].

New results contained in this article were presented at a seminar in the Center on the Problems of Development and Control in Humanities of the Ministry of Culture in 1982.

The author is grateful to S. F. Soprunov for stimulating discussions.

1. Statement of the problem

1.1. Preliminary remarks. Let P be a finite collection of finite sets; the sets of the collections will be called problems, and their elements will be called states. With each state we associate a variable which will be interpreted as the probability of the event that the problem containing this state lies precisely in this state. Our purpose is to find the probabilities of all states.

The algebra of events of the probability space is given by the collections of states, one from each problem. The event corresponding to a state consists of those collections that select this state from the corresponding problem.

Denote the event corresponding to the state S by $m(S)$ and the probability of the event μ by $w(\mu)$.

Let S be a state, $U \subset P$. Then it follows from a well-known theorem that

$$w(m(S)) = \sum_{v \in \otimes U} w(m(S) \mid \bigcap_{p \in U} m(\pi_p v)) \times (\bigcap_{p \in U} m(\pi_p v)),$$

where $\otimes U$ is the direct product of sets belonging to U and $\pi_p v$ is the projection of U on p. If we know the way in which problems in U influence the problem containing S, then we can assume that we are given conditional probabilities

$$w(m(s) | \bigcap_{p \in U} m(\pi_p v)).$$

If the problems in U are independent, then

$$w(\bigcap_{p \in U} m(\pi_p v)) = \prod_{p \in U} w(m(\pi_p v)).$$

Thus the two preceding conditions yield equations on the state probabilities. The verification of whether these conditions are satisfied must be carried out in each case arising in practice.

Formal description. Let P be a finite set of finite sets (problems) whose elements are variables with the domain $[0, 1]$. In addition, let $\sum_{x \in p} X = 1$ for each $p \in P$. Suppose we are given a function U associating with each problem a set of problems (influencing it). Suppose that a function is given associating with each variable X $(\in p \in P)$ and each collection $v \in \otimes U(p)$ the number $W_x(v) \in [0, 1]$. Here $\sum_{x \in p} W_x(v) = 1$ for all $p \in P$ and all $v \in \otimes U(p)$. We must solve the following system of equations:

$$\left\{ x = \sum_{v \in \otimes U(p)} W_x(v) \prod_{q \in U(p)} \pi_q(v) \mid x \in p \in P \right\}. \tag{1}$$

2. Existence of a solution

The system of equations (1) can be written in general form as $y = f(y)$, where y is a vector and f is an operator. Let us investigate the properties of f.

Domain. As we know, the variables of a problem take values in a simplex whose dimension is one less than the number of variables in the problem. Therefore, the domain of definition S for the operator f (the domain of definition of all variables) is the direct product of simplices.

PROPOSITION. *The simplex is homeomorphic to a cube of the same dimension.*

PROOF. Let n be the dimension. Embed the simplex and the cube in R^n so that the centers of both are at the origin $\vec{0}$. Denote the resulting figures by σ and κ respectively. Consider the functions $\alpha, \beta: R^n \setminus \{\vec{0}\} \to (0, \infty)$:

$$\alpha(y) = \sup_{ky \in \sigma} k, \quad \beta(y) = \sup_{ky \in \kappa} k.$$

Evidently, these functions are continuous and the ratios $\frac{\alpha}{\beta}, \frac{\beta}{\alpha}$ are bounded. The mapping G defined by

$$G(y) = \begin{cases} \frac{\beta(y)}{\alpha(y)} y & \text{for } y \neq \vec{0}, \\ \vec{0} & \text{for } y = \vec{0} \end{cases}$$

is a homeomorphism of σ on κ.

Completion of the proof. The direct product of cubes is a cube whose dimension is equal to the sum of the dimensions of the factors. This and the proof of the statement imply that S is homeomorphic to a simplex of a certain dimension.

Range. Let us denote the xth coordinate of the vector $f(y)$ by $fx(y)$,

$$fx(y) = \sum_{v \in \otimes U(p)} W_x(v) \prod_{q \in U(p)} \pi_q(v),$$

where $x \in p \in P$. We have $fx(y) \geq 0$ for all y and for all x, since $fx(y)$ it is obtained from positive numbers by the operations of multiplication and addition. Also we have $\sum_{x \in p} fx(y) = 1$ for all y and all p. Indeed,

$$\begin{aligned} \sum_{x \in p} fx(y) &= \sum_{x \in p} \Big(\sum_{v \in \otimes U(p)} W_x(v) \prod_{q \in U(p)} \pi_q(v) \Big) \\ &= \sum_{v \in \otimes U(p)} \Big(\sum_{x \in p} W_x(v) \prod_{q \in U(p)} \pi_q(v) \Big) \\ &= \sum_{v \in \otimes U(p)} \Big(\Big(\sum_{x \in p} W_x(v) \Big) \prod_{q \in U(p)} \pi_q(v) \Big) \\ &= \sum_{v \in \otimes U(p)} \prod_{q \in U(p)} \pi_q(v) = \prod_{q \in U(p)} \sum_{z \in q} z = \prod_{q \in U(p)} 1 = 1. \end{aligned}$$

Therefore, the operator f maps the domain S into itself.

Continuity of the operator. The operator f is continuous since it is defined in terms of operations of multiplication and addition.

If ψ is a homeomorphism of the domain S onto the simplex, then finding a fixed point for the operator f is equivalent to finding a fixed point of the continuous operator $\psi^{-1} f \psi$ mapping the simplex onto itself. Let us first state the problem in discrete language. This is more than appropriate if we are looking for a computer-generated solution.

Partition of the simplex. We shall identify an n-simplex with the set of its vertices $(a_0, a_1, \ldots, a_{n-1}, a_n)$.

With each point b in the simplex $(a_0, a_1, \ldots, a_{n-1}, a_n)$ we associate *its proper* partition of the simplex into the simplices $(b, a_1, \ldots, a_{n-1}, a_n)$, $(a_0, b, \ldots, a_{n-1}, a_n)$, $\ldots$, $(a_0, a_1, \ldots, b, a_n)$, $(a_0, a_1, \ldots, a_{n-1}, b)$.

A face of a simplex is a simplex of dimension one less whose vertices are also vertices of the first simplex. By a partition we mean a partition of the simplex into simplices of the same dimension.

Given a partition of a face, it can be extended to the partition of the simplex in such a way that the vertices of each element in the partition are the vertices of some element of the partition of the face and a vertex of the simplex that does not belong to this face.

If a point is fixed in the interior of the simplex, and for each face a partition is selected, then we can extend a partition of each face to a partition of the corresponding element in the fixed point's proper partition. This results in a partition of the simplex that extends both its proper partition and given partitions of the faces.

If a point is fixed in the interior of each simplex of positive dimension, then we can fix a partition for each simplex by induction in the dimension. For the simplex of dimension 0 there is a unique partition. The induction step is performed according to the above procedure.

It is easy to prove by induction that each face of an element of the fixed partition either lies on the face of the simplex or is a face of another element of the partition. If each element of the partition is partitioned in some manner, this results in the partition of the initial simplex such that the *two-sided property* given in the preceding sentence is preserved. This partition process can be continued indefinitely. We want the size of the elements of the partition to tend to zero. In order to achieve this goal, fix the arithmetic mean of the vertices in the simplex $\frac{a_0+a_1+\ldots+a_{n-1}+a_n}{n+1}$. Let us estimate the distance from this center point to the vertices of the simplex in terms of the difference between vertices of the simplex. For example, the distance to a_0 is

$$\begin{aligned}
&\left|\frac{a_0 + a_1 + \ldots + a_{n-1} + a_n}{n+1} - a_0\right| \\
&\quad = \left|\frac{(a_1 - a_0) + (a_2 - a_0) + \ldots + (a_{n-1} - a_0) + (a_n - a_0)}{n+1}\right| \\
&\quad \le \frac{n \max |a_i - a_0|}{n+1} \le \left(1 - \frac{1}{n+1}\right) \max_i |a_i - a_0|.
\end{aligned}$$

Let us call the maximum of the distances between the vertices of the simplex its diameter. Thus, the distance from the center to a vertex (and, consequently, to any point of the simplex) does not exceed $\left(1 - \frac{1}{n+1}\right)$ times the diameter of the simplex. Hence it follows that the diameter of the elements of the fixed partition is bounded by the same number. So by partitioning the simplex in a fixed manner k times we obtain a partition with all elements having diameter not exceeding $\left(1 - \frac{1}{n+1}\right)^k$ times the diameter of the initial simplex.

Let φ be a continuous operator mapping the simplex into itself. With each point in the simplex we can associate the nondegenerate element of its proper partition which includes its image (an element can be degenerate if a point belongs to a face).

If there are several such elements, select one of them. An element of the proper partition is uniquely defined by that face of the partitioned simplex on which it is based. A face of the simplex is uniquely defined by the vertex that does not belong to it. Therefore, with each point of the simplex we have associated its vertex. This correspondence satisfies the following *boundary condition*: points lying on a face can only correspond to vertices belonging to the same face. It turns out that even this very "rough" function is sufficient to find the fixed point of the initial function (of the operator φ). If $\varphi(y) \neq y$, then the ray $\overrightarrow{y, \varphi(y)}$ intersects the face of the simplex corresponding to the point y. The intersection of all faces of the simplex is empty. Indeed, the simplex is defined by the equation $\sum_{i=0}^{u} x_i = 1$, where $x_i \geq 0$. A face is defined by the equation $x_i = 0$. Thus, if $\varphi(y) \neq y$, then the ray $\overrightarrow{y, \varphi(y)}$ does not intersect some face U. Since φ is continuous, we have $\varphi(Z) \neq Z$ for all Z sufficiently close to y, and the ray $\overrightarrow{Z, \varphi(Z)}$ does not intersect the same face. Thus, if arbitrarily close to point y there are points corresponding to all possible vertices, then $\varphi(y) = y$.

PROPOSITION. *Suppose there is a simplex and a function which associates to the vertices of the elements of the partition the vertices of the partitioned simplex. Suppose also that the two-sided property is satisfied for the partition, and the boundary property for the function.*

Then there exists an element in the partition in whose vertices the function positions all the vertices of the partitioned simplex.

PROOF. We shall prove by induction on the dimension that there exist an *odd number* of desired elements. The 0-simplex has exactly one desired element. To perform the step from n to $n+1$, let us assume that the vertices of the partitioned simplex are $a_0, a_1, \ldots, a_n, a_{n+1}$. A face in the element of the partition will be called regular, if the function positions $a_0, a_1, \ldots, a_n$ in its vertices. By the boundary condition regular faces occur only on the face $a_0, a_1, \ldots, a_n$ and inside the simplex. It is easy to check that the boundary condition implies the boundary condition for each of the faces of the partitioned simplex. The two-sided property is also satisfied for faces. Therefore, the inductive hypothesis implies that the number of regular faces on the face $a_0, a_1, \ldots, a_n$ is odd. Consider an element of the partition having a regular face. If the vertex of the element which does not belong to this face is mapped into a_{n+1}, then the element is the desired one, and there is one regular face in it. Otherwise the element is not the desired one and has two regular faces. Let us sum up all the regular faces over all elements of the partition. By the two-sided property, all internal faces will be counted twice, while the number of regular faces on the boundary, as was already proved, is odd; therefore the entire sum is odd. Hence the number of the desired element of the partition is odd.

If the diameter of the elements of the partition tends to 0, then (since the simplex is compact) there exists a limit point for the sequence of contracting desired elements. This point is a fixed point for the operator φ.

3. Existence of the algorithm

Given a partition and a function, one can use a search process to find the desired element of the partition. As a rule, we can determine how small a partition

should be in order for the operator φ to move the points of the desired element of the partition by less than ε. Given $\varepsilon > 0$, we must find $\delta > 0$ such that for any $|y_1 - y_1| < \delta$ the relation $|\varphi(y_2) - \varphi(y_1)| < \varepsilon$ is satisfied. Such a δ exists since any continuous operator on a compact is uniformly continuous. On the other hand, it is usually impossible to specify how small a partition should be in order that the distance between the desired element of the partition and the fixed point be less than ε (the work of V. P. Orevkov [2]).

4. Problem complexity

We will show that our main problem is as hard as the general problem of finding the fixed point of a continuous mapping of a simplex into itself.

Relation between probability and relative frequency. Suppose we have a random variable assuming values 0 and 1. Perform n independent trials of the variable. One can prove that the probability that the relative frequency of the occurrence of 1 and the probability of the occurrence of 1 differ by more than ε, is less than $1/c \cdot \varepsilon e^{2\varepsilon^2 n}$, where c can be taken just above $4\sqrt{2\pi}$ ("just above" $\to 0$ as $n \to \infty$). This estimate tends to 0 as $n \to \infty$ and does not include the probabilities of values for the random variable. Now let ψ be a random variable assuming k values $a_1, \ldots, a_k$. For each i we can consider the random variable taking value 1 if $\psi = a_i$ and 0 otherwise. Applying the above estimate to these variables, we see that the probability that the relative frequency of the event that at least one value of ψ in n independent trials differs from the probability of the same value by no more than ε, does not exceed $k/c \cdot \varepsilon \cdot e^{2\varepsilon^2 n}$.

Let us proceed with the construction of system (1).

Construction. The set P will consist of the problem A and the problems $B_1, \ldots, B_n$. Each of the problems will have k variables. The variables of the problem B_j will be denoted by $(b_{j_1}, \ldots, b_{j_k})$. The vector of the variables for a problem will be denoted in the same way as the problem itself.

The variables of the problem A are interpreted as the probabilities of values of the variable Ψ. Suppose that a sequence of n independent trials of ψ is given. Then b_{ij} is the probability of the event that $\psi = a_i$ in trial j.

Let us proceed with the formal construction. Each B-problem is influenced by one problem A, while A is influenced by all B_j.

The comparisons for B-problems are of the form $B_j = A$. The interpretation of this is that the same random variable is realized in all trials.

In vector notation, equations for the problem A are of the form:

$$A = \sum_{v \in \otimes w(v)} \prod_{j=1}^{n} \pi_j(v),$$

where $\otimes B$ is the direct product of $B_1, \ldots, B_n$.

Substituting in the right-hand side for the variables of the B-problem the corresponding variables of the problem A, we obtain a fixed point problem for a $(k-1)$-simplex.

Let f be a continuous mapping of the $(k-1)$-simplex into itself. An appropriate choice of the coefficients $W(v)$ makes the operator $\sum_{v \in \otimes B} w(v) \prod_{j=1}^{n} \pi_j(v)$, where

for the variables of the B-problems are substituted the corresponding variables of problem A, uniformly close to f. The larger n, the better the approximation.

To each element $v \in \otimes B$ there corresponds a sequence of values of the variable ψ. Denote the vector of frequencies of occurrences of $a_1, \ldots, a_k$ by $h(v)$. Evidently, $h(v)$ lies in the domain of definition of f. Set $w(v) = f(h(v))$. Define the norm of the vector as the maximum of absolute values of its coordinates. A continuous mapping of a compact is uniformly continuous. Therefore, there exists ε such that for any points y_1, y_2 of the simplex the relation $|y_1 - y_2| \leq \varepsilon$ implies

$$|f(y_1) - f(y_2)| \leq \delta.$$

Note that $\prod_{j=1}^{n} \pi_j(v)$ is the probability that the sequence of values ψ obtained in n independent trials coincides with v.

Substitute for the value of the variables in the sum

$$\sum_{v \in \otimes B} f(h(v)) \prod_{j=1}^{n} \pi_j(v)$$

a point of the simplex y. Then the sum splits into two parts: over such v for which $|h(v) - y| > \varepsilon$, and over such v for which $|h(v) - y| \leq \varepsilon$. The first part of the sum is small because

$$\sum_{\substack{v \in \otimes B \\ |h(v)-y|>\varepsilon}} \prod_{j=1}^{n} \pi_j(v)$$

after the substitution for b_{j_i}, expresses the probability of the event that the vector of frequencies differs from the vector of probabilities by more than ε (for at least one value of ψ). The coefficients of $f(h(v))$ are bounded. In the second part of the sum we have $|f(h(v)) - f(y)| \leq \delta$ (uniform continuity). The expression

$$\sum_{\substack{v \in \otimes B \\ |h(v)-y|\leq\varepsilon}} \prod_{j=1}^{n} \pi_j(v)$$

after the substitution of $\pi_i(y)$ for b_{j_i} gives the probability of the event that the frequencies are close to the probability of y, which is thus close to 1. Hence the entire sum is close to $f(y)$, as required.

Reduction to the quadratic case. In the preceding construction the degree of equations was equal to n (the number of trials) and increased with the growth in the accuracy of the approximation of the mapping f. The number of variables in the problem was the same number k. It turns out that the construction can be modified in such a way that the degree of equations is equal to 1 or 2, although at the expense of having more variables.

Namely, starting with system (1) we can construct an equivalent system of the same type in which each equation has degree 1 or 2. For each problem p, introduce a new problem p' whose variables are indexed by elements of $\otimes U(p)$. To each equation of the old system there corresponds a linear equation of the new system

$$x = \sum_{v \in p'} W_x(v) \cdot v, \quad x \in p \in P.$$

It remains to add the condition that $v = \prod_{q \in U(p)} \pi_q(v)$ for $v \in p'$. If A and B are problems, then we can construct a problem whose variables are denoted by the element $A\emptyset B$, with the equations $\langle a, b \rangle = a \cdot b$, $b \in B$. The degree of these equations is 2. By applying these equalities (i.e., introducing new problems with corresponding equations) as many times as there are problems in $U(p)$ (to be more precise, one time less), we obtain the required equalities for $v \in p'$.

5. Iterative method

If μ is a metric space and f a continuous mapping of μ into itself, then the iterative method for finding a fixed point of f is an attempt to find the limit of $f^n(x_0)$ as $n \to \infty$, where $x_0 \in \mu$. If this limit exists, then it is the fixed point of f. If f is a contraction mapping with coefficient α (i.e., $0 \leq \alpha < 1$ and $\rho(f(x), f(y)) \leq \alpha\rho(x, y)$ for all $x, y \in \mu$, where ρ is the distance in μ), then $\rho(f^{n+1}(x_0), f^n(x_0)) \leq \alpha^n(\rho(f(x_0), x_0)$ for $n \geq 0$. Therefore,

$$\rho(f^{n+k}(x_0), f^n(x_0)) \leq \alpha^n \rho(f(x_0), x_0) \cdot \sum_{i=0}^{k-1} \alpha^i$$

$$= \alpha^n \rho(f(x_0), x_0) \cdot \frac{1 - \alpha^k}{1 - \alpha} \leq \frac{\alpha^n}{1 - \alpha} \rho(f(x_0), x_0)$$

for all $n \geq 0$, $k > 0$. Thus, $\rho(f^{n+k}(x_0), f^n(x_0))$ tends to 0 as $n \to \infty$, and the smaller α is, the faster. If μ is a complete space, then the limit of $f^n(x_0)$ as $n \to \infty$ exists and $\rho(f^n(x_0), x) \leq \frac{\alpha^n}{1-\alpha} \rho(f(x_0), x_0)$.

Differentiable case. In our case μ is the product of simplices (which is a complete space), f a polynomial (and, consequently, differentiable) mapping.

In general, if f is a continuous differential mapping from a normed vector space A into a normed vector space B, then f is a contraction map with the coefficient α if and only if the derivative of f at each point is a contracting linear operator with the coefficient α. The "only if" part is proved by differentiating f, the "if" part by integrating f'.

In our case the derivative acts in the linear space D tangent to the domain of definition of f (i.e., to S).

Let us represent vectors in D in the basis of initial variables. Denote the projection of the vector y on the direction corresponding to the variable x by y_x.

$$D = \left\{ y | \forall p \in P \sum_{x \in p} y_x = 0 \right\}.$$

Norm. Introduce a norm on vectors $\|y\| = \max_{p \in P} \sum_{x \in p} |y_x|$. (Prove that it is a norm!) Our aim is to estimate the norm of f' (uniformly over the points at which the derivative is taken) with respect to the norm on vectors. If the norm of f' does not exceed α (< 1) at all points, then f is a contracting mapping with coefficient α.

Bound for the norm of f'. Consider a point x in S, Its coordinates are values of variables X. Let us estimate the norm of the derivative of f at this point applied to the vector y ($\in D$) in terms of the norm of y. Since $\|f'(y)\| =$

$\max_{p\in P}\sum_{x\in p}|f'_x(y)|$, it is sufficient (and necessary) to estimate $\sum_{x\in p}|f'_x(y)|$ for a single p.

This is what we shall do now.

By definition

$$f_x=\sum_{v\in\otimes U(p)}W_x(v)\prod_{q\in U(p)}\pi_q(v).$$

An application of the chain rule formula yields

$$f'_x(y)=\sum_{v\in\otimes U(p)}W_x(v)\sum_{r\in U(p)}y_{\pi_r}(v)\prod_{\substack{q\in U(p)\\ q\neq r}}\pi_q(v).$$

Using the fact that $\sum_{z\in r}z=1$, we can multiply each $y\pi_r(v)$ by $\sum_{z\in r}$. Thus,

$$f'_x(y)=\sum_{v\in\otimes U(p)}W_x(v)\sum_{r\in U(p)}y_{\pi_r}(v)\Big(\sum_{z\in r}z\Big)\prod_{\substack{q\in U(p)\\ q\neq r}}\pi_q(v).$$

Making use of the distributive property in the right-hand side, we obtain a linear combinations of monomials $\prod_{q\in U(p)}\pi_q(v)$, where $V\in\otimes U(p)$. Namely,

$$f'_x(y)=\sum_{v\in\otimes U(p)}\Big(\sum_{q\in U(p)}\sum_{z\in q}w_x(v_q^z)y_z\Big)\prod_{q\in u(p)}\pi_q(v),$$

where v_q^z denotes the result of substituting z for the qth coordinate of v. Since the variables x are nonnegative, we have

$$|f'_x(y)|\le\sum_{v\in\otimes U(p)}\Big(\sum_{q\in U(p)}\Big|\sum_{z\in q}w_x(v_q^z)yz\Big|\Big)\prod_{q\in u(p)}\pi_q(v).$$

Define $m_x(v_q)=\min_z w_x(v_q^z)$. Since $y\in D$, we have $\sum_{z\in q}y_z=0$, which implies that

$$\Big|\sum_{z\in q}w_x(v_q^z)y_z\Big|=\Big|\sum_{z\in q}(w_x(v_q^z)-m_x(v_q))y_z\Big|.$$

Since the values $(w_x(v_q^z)-m_x(v_q))$ are nonnegative, we have

$$|f'_x(y)|\le\sum_{v\in\otimes U(p)}\Big(\sum_{q\in U(p)}\Big(\sum_{z\in q}w_x(v_q^z)|y_z|-m_x(v_q)\sum_{z\in q}|y_z|\Big)\Big)\prod_{q\in U(p)}\pi_q(v).$$

Making use of the last inequality to estimate $\sum_{x\in p}|f'_x(y)|$, move the sign of summation over X inside. This gives us

$$\sum_{x\in p}|f'_x(y)|\le\sum_{v\in\otimes U(p)}\Big(\sum_{q\in U(p)}\Big(\Big(1-\sum_{x\in p}m_x(v_q)\Big)\sum_{z\in q}|y_z|\Big)\Big)\prod_{q\in U(p)}\pi_q(v).$$

Define $\nu(p)$ as the cardinal number of the set $U(p)$, and

$$\mu(p)=\min_{v\in\otimes U(p)}\sum_{q\in U(p)}\sum_{x\in p}m_x(v_q).$$

Since $\sum_{z\in q}|z_q| \le \|y\|$, we have

$$\sum_{x\in p}|f'_x(y)| \le \sum_{v\in\otimes U(p)}\prod_{q\in U(p)}\pi_q(V)$$
$$= (\nu(p)-\mu(p))\|y\|\sum_{v\in\otimes U(p)}\prod_{q\in U(p)}\pi_q(v) = (\nu(p)-\mu(p))\|y\|.$$

Thus, $\|f'\| \le \max_{p\in P}(\nu(p)-\mu(p))$.

6. Method of descent

Suppose that $\varphi: S \to R$ is a differentiable functional and that our task is to find its minimal value. In order to apply the known descent methods, it is useful to estimate $\min_{y\in S}(\varphi(z))'(y-z)$ in terms of $\varphi(z)$, where $z \in S$ and $(\varphi(z))'$ is the derivative of φ at the point z. The idea of the descent method is that a translation along the vector $(y-z)$ reduces the value of φ.

In our case $\varphi(z) = (f(z)-z, f(z)-z)$, where $(\cdot,\cdot)$ is the usual scalar product. Clearly, $\varphi(z) \ge 0$, and $\varphi(z) = 0 \leftrightarrow f(z) = z$. We have

$$\begin{aligned}(\varphi(z))'(y-z) &= 2(f(z)-z, f(z)-z)'(y-z))\\ &= 2(f(z)-z, (f(x))'(y-z)+f(z)-y)\\ &= -2(f(z)-z, f(z)-z)+2(f(z)-z, (f(z))'(y-z)+f(z)-y)\\ &= -2\varphi(z)+(f(z)-z, (f(z))'(y-z)+f(z)-y).\end{aligned}$$

Now our goal is to find $y \in S$ such that

$$(f(z))'(y-z)+f(z)-y = 0.$$

For this value of y we have $(\varphi(z))'(y-z) = -2\varphi(z)$. We want to prove that there exists $y \in S$ such that $(f(z))'(y-z)+f(z) = y$, i.e., the fixed point of the affine (with respect to y) operator $(f(z))'(y-z)+f(z)$ in the domain S. Since an affine operator is continuous, it is sufficient to prove that it maps the domain S into itself. Note that since $y \in S, z \in S$, we have $(y-z) \in D$ and, consequently, also $(f(z))(y-z) \in D$. Since $f(z) \in S$, for each $p \in P$ we have

$$\sum_{x\in p}((f(z))'(y-z)+f(z))_x = 1.$$

Therefore, it remains to prove that

$$((f(z))'(y-z)+f(z)+f(z))_x \ge 0$$

for each $x \in p \in P$. Of course, this inequality does not always hold. We find a simple necessary and sufficient condition for it to be true.

Let $p \in P$ and $x \in p$. We need

$$(f(z))'_x(y) \ge (f(z)-f_x(z))$$

for all $y, z \in S$. Since f_x is a homogeneous polynomial of degree $\nu(p)$, we have $(f(z))'_x(z) = \nu(p)f_x(z)$. This can also be easily established by a direct application

of the derivation formula. In the process of estimating the norm of f', we proved that

$$(f(z))'_x(y) = \sum_{v\in\otimes U(p)} \left(\sum_{q\in U(p)} \sum_{t\in q} W_x(v_q^t)y_t \right) \prod_{q\in U(p)} \pi_q(v).$$

Thus, we need that the equality

$$\sum_{v\in\otimes U(p)} \left(\sum_{q\in U(p)} \sum_{t\in q} W_x(vt_q)y_t \right) \prod_{q\in U(p)} \pi_q(v) \geq (\nu(p)-1)f_x(z)$$

holds for all $y, z \in S$. Since

$$f_x(z) = \sum_{v\in\otimes U(p)} w_x(v) \prod_{q\in U(p)} \pi_q(v),$$

we have an inequality for linear combinations of $\prod_{q\in U(p)} \pi_q(v)$. Substituting for each $v \in \otimes U(p)$ that value of z for which $\prod_{q\in U(p)} \pi_q(v) = 1$ and the remaining products are equal to 0, we see that the inequality must be true coefficientwise (evidently, it is also sufficient). Thus, for all $y \in S$ and all $v \in \otimes U(p)$ we have

$$\sum_{q\in U(p)} \sum_{t\in q} W_x(v_q^t)y_t \geq (\nu(p)-1)w_x(v).$$

Substituting those y for which in every q there exists t such that $y_t = 1$ and the remaining coordinates are equal to 0, we see that

$$\sum_{q\in U(p)} \min_{t\in q} w_x(v_q^t) \geq (\nu(p)-1)W_x(v).$$

Clearly, this relation is also sufficient. Thus, a necessary and sufficient condition is

$$\sum_{q\in U(p)} m_x(v_q) \geq (\nu(p)-1)W_x(v).$$

Norm of f'. If we sum up the latter inequalities over all $x \in p$, we obtain $\sum_{x\in p} \sum_{q\in U(p)} m_x(v_q) \geq \nu(p) - 1$, and $\mu(p) - \nu(p) - 1$. In this case $\|f'\| \leq 1$. This (and the next example) illustrates the comparative efficiency of the iterative and the descent methods.

Example. Let $P = \{\{x_1, x_2\}\}$. Consider the system

$$\left\{ \begin{matrix} x_1 = x_2 \\ x_2 = x_1 \end{matrix} \right\}.$$

It is equivalent to the equation $x = 1 - x$. The iterative method does not work in this case, but the method of descent gives fast convergence to the solution $x = 0.5$.

Finding $m_x(v_q)$. When computing $m_x(v_q)$ we have to bear in mind that $m_x(v_q)$ does not depend on the qth coordinate of v. Therefore, one can organize the table of "truncated" vectors $\bar{v}$ with one coordinate dropped. First we perform a search process in the lexicographic order among the vectors without the first coordinate, then without the second, and so on. Each vector v is "connected" with $\nu(p)$ "truncated" vectors. Performing the search in the lexicographic order,

we shall replace every number located in the squares of the table connected with v by the minimum of its $uw_x(v)$. This procedure reduces the search.

7. Main ideas

Two main ideas were used to derive the results in this article: transformation of the formula for $f'_x(y)$ in order to find the degree of the polynomial, and implicit determination of the desired value for y in the justification of the descent method.

8. Multiple influence

Both main ideas are also applicable if we assume that one problem can influence another n times (i.e., its variables occur n times in the vector v).

When vectors v are stored in the computer memory, there is the problem of saving memory, since the order of coordinates is not always essential. Consider the following problem. Suppose that in the vector v, all copies of the variable with a smaller number precede all copies of the variable with larger number, i.e., the vector starts with the copies of the first variable, then the copies of the second variable, and so on. (The variables in the problem under consideration are assumed to be ordered in some way.) This creates the property of uniqueness for the representation. We shall call such vectors canonical.

Let the variables belonging to the problem under consideration be $x_1, \ldots, x_k$. Then the number of all sequences of these variables of length n is equal to k^n. A sequence in which the numbers of variables are nondecreasing will be called a *ladder*. Let us compute the number of all ladders. Consider n cells, in which the variables can be positioned. Distribute $k-1$ walls among the cells in such a way that the ith wall separates the group of variables x_i from the group of variables x_{i+1} (two walls can to be adjacent to each other if the corresponding group of variables in empty). Evidently, the ladders and the placements of walls are in one-to-one correspondence. Consider $(n+k-1)$ positions such that each of them can be occupied by either a cell or a wall. By placing cells in n positions and walls in $(k-1)$ positions we obtain a placement of walls. It is now evident that the desired number of ladders is equal to

$$\binom{n}{n+k-1} = \frac{(n+k-1)!}{n!(k-1)!}$$

(the binomial coefficient).

The number of canonical vectors is evidently equal to the product of the number of ladders for all problems of influence.

It will be convenient to store ladders in the computer memory in reverse lexicographic order. Then the address of a ladder is equal to the number of ladders (it is now more convenient to speak about $(n-k)$-ladders) greater than the given one. The set of $(n-k)$-ladders that are greater that the given one is decomposed into n sets: the ith set contains the ladders that are equal to the positions with numbers less than i, and is greater than the given ladder at the position i. If the ith position in the ladder is occupied by the variable x_j, then it is not hard to see that the number of ladders in the ith set corresponding to this ladder is equal to the number of all $(n-i+1)(k-j)$-ladders. (For $\chi = 0$ the number ν of χ-ladders is equal to 0.)

If a canonical vector v contains variables of the problems from the set U, then the address of v is equal to

$$\sum_{q\in U} \alpha(v,q) \prod_{\substack{r\in U\\ r<q}} \beta(r),$$

where $\alpha(v,q)$ is the address of the ladder of variables of the problem q in the vector v, $\beta(r)$ is the number of all ladders of variables of the problem r; we assume that an ordering $r < q$ is defined on the problems; and the product of the empty set of factors is equal to 1.

For a fast computation of addresses of canonical vectors, two tables should be created in the computer memory: the table of the numbers $\binom{\kappa}{\nu}$, where $0 \le \kappa\nu <$ max(multiplicity of the influence of q) + (the number of variables in q); and the table of numbers $\prod_{\substack{r\in U\\ r\le q}} \beta(r)$, where $q \in U$.

References

1. V. S. Vysotskiĭ and S. A. Petrovskiĭ, *A study of some questions of existence and uniqueness of solutions for systems of equations in the nonlinear analysis model on problem nets*, Analysis on Problem Nets, vol. 1, Moscow, 1980, pp. 83–101. (Russian)
2. V. P. Orevkov, *A uniformly continuous constructive mapping of a square into itself with no fixed points*, Appendix 6 to the Russian transl. of J. Barwise (ed.), *Handbook of mathematical logic*. Part IV: *Proof theory and constructive mathematics*, "Nauka", Moscow, 1982. (Russian)