Ammosov North-Eastern Federal University

YAKUTIAN MATHEMATICAL JOURNAL

PUBLISHED SINCE 2014 SCIENTIFIC JOURNAL 4 ISSUES PER YEAR

Vol. 21, No. 3 (83) July—September, 2014

CONTENTS

Mathematics

Mathematical Modeling

Address for correspondence:
Ammosov North-Eastern Federal University, Belinskii St., 58, Yakutsk, 677000
Phone: 8(4112)32-14-99, Fax: 8(4112)36-43-47;
http://s-vfu.ru/universitet/rukovodstvo-i-struktura/instituty/niim/mzsvfu/
e-mail: prokopevav85@gmail.com; madu@ysu.ru: ivanegorov51@mail.ru

Yakutian Mathematical Journal
July—September, 2014. Vol. 21, No. 3

UDC 517.633

SMOOTH SOLUTIONS TO A BOUNDARY VALUE PROBLEM FOR AN OPERATOR–DIFFERENTIAL EQUATION OF MIXED TYPE

V. I. Antipin

Abstract. We study solvability of boundary value problems for the so-called kinetic operator-differential equations of the form $Bu_t - Lu = f$, with L and B families of linear operators in a complex Hilbert space E. We do not assume that B is invertible and that the spectrum of the pencil $L - \lambda B$ lies in one of the half-planes $\operatorname{Re}\lambda < a$ or $\operatorname{Re}\lambda > a$ ($a \in \mathbb{R}$). Under some conditions on these operators we study the question of smoothness of solutions in weighted Sobolev spaces.

Keywords: kinetic equation, operator-differential equation, weighted Sobolev spaces

The article is devoted to the study of the operator-differential equation

$$Au \equiv Bu_t - Lu = f(x,t), \tag{1}$$

where the linear operators B and L are defined in a given Hilbert space E, and B is selfadjoint. The boundary conditions are of the form

$$P^{+}u(0) = u_0, \quad P^{-}u(T) = u_T, \tag{2}$$

where P^{+} and P^{-} are the spectral projections of B corresponding to the positive and negative parts of the spectrum. We do not assume that B is invertible; in particular, B can have a nonzero kernel and its spectrum can contain infinite subsets of positive and negative semiaxes simultaneously.

Thus, we consider (1) which is not an equation of Sobolev type. As is known, similar equations arise in physics, in particular, in neutron transport, radiative transfer, and rarefied gas dynamics [1–13], as well as in geometry, population dynamics, hydrodynamics [14–17], and some other fields [5, 18].

In the case of a selfadjoint operator L, (1) is presented in [5]. In particular, some examples arising in applications can be found in [18]. An operator L is referred to as *Kato-sectorial* (see the definition in [19]), if $|(Lu, v)| \le c\|u\|_{H_1}\|v\|_{H_1}$ for $u, v \in D(L)$, with $\|u\|_{H_1} = \operatorname{Re}(-(Lu, u) + \|u\|)$. Some generalizations to the case of a dissipative operator obeying the Kato-sectoriality condition can be found in [20]. Boundary value problems for (1) under our constraints on L and B in the case when the Kato-sectoriality condition fails were probably not considered earlier. Note that the class of equations (1) includes many important partial differential equations; for example, we can take the odd order operators as L [20, 21] or the operators whose spectrum lies near the imaginary axis.

The author was supported by the Ministry of Education and Science of the Russian Federation Within the State Maintenance Program of Research Activities (Grant No. 3047).

1. Basic Assumptions

Let E and $H_1 \subset E$ be Hilbert spaces and let the last embedding be dense. The symbol $(\cdot,\cdot)$ designates the inner product on E. In this case the negative space H_1' constructed as the completion of E with respect to the norm

$$\|u\|_{H_1'} = \sup_{v\in H_1,\, v\neq 0} |(u,v)|/\|v\|_{H_1},$$

coincides with the space of continuous antilinear functionals over H_1 and the inner product on E admits extension to the duality between H_1 and H_1' [22].

If X and Y are Hilbert spaces then $L(X,Y)$ stands for the space of continuous linear operators on X with values in Y. If $X = Y$ then we write $L(X)$ rather than $L(X,X)$.

An operator $L : E \to E$ is called *dissipative* (*uniformly dissipative*) if $-\operatorname{Re}(Lu,u) \geq 0$ ($-\operatorname{Re}(Lu,u) \geq \delta\|u\|^2$, $\delta > 0$) for all $u \in D(L)$. Here $D(L)$ is the domain of L. An operator L is called *maximal dissipative* if L agrees with each of its dissipative extension.

Denote by $\rho(L)$ and $\sigma(L)$ the resolvent set and the spectrum of L. The basic assumptions on L and B are as follows:

(I) L is a maximal dissipative operator and there exists a Hilbert space F_1 densely embedded into E such that $D(L^*) \subset F_1 \subset E$ and there exists a constant $\delta_0 > 0$ such that $\operatorname{Re}(-L^*u,u) \geq \delta_0\|u\|^2_{F_1}$ for all $u \in D(L^*)$, with L^* the adjoint operator.

The condition (I) implies that L^* is a maximal dissipative operator as well and $0 \in \rho(L)\cap\rho(L^*)$ [19, Proposition C.7.2]; moreover, $\{\operatorname{Re}\lambda \geq 0\} \subset \rho(L)\cap\rho(L^*)$.

(II) B is selfadjoint in E and the embedding $F_1 \subset D(|B|^{1/2})$ is dense.

Lemma 1. *Under condition* (II), *B defines a continuous mapping of F_1 in F_1', where F_1' is the negative space constructed on the pair F_1, E.*

Proof. The operator $B : D(|B|) \to E$ (we introduce the graph norm on $D(|B|)$), being defined on $D(|B|)$, admits extension to a continuous mapping from $D(|B|^{1/2})$ in $(D(|B|^{1/2}))'$. Indeed,

$$\|Bu\|_{(D(|B|^{1/2}))'} = \sup_{v\in D(|B|^{1/2})} \frac{|(Bu,v)|}{\|v\|_{D(|B|^{1/2})}} = \sup_{v\in D(|B|^{1/2})} \frac{|(|B|^{1/2}u, |B|^{1/2}v)|}{\|v\|_{D(|B|^{1/2})}}$$
$$\leq \sup_{v\in D(|B|^{1/2})} \frac{\||B|^{1/2}u\|\cdot\||B|^{1/2}v\|}{\|v\|_{D(|B|^{1/2})}} \leq \||B|^{1/2}u\| \tag{3}$$

or

$$\|Bu\|_{(D(|B|^{1/2}))'} \leq \||B|^{1/2}u\|. \tag{4}$$

We have the natural embedding

$$F_1 \subset D(|B|^{1/2}) \subset E \subset (D(|B|^{1/2}))' \subset F_1' \tag{5}$$

(the dual of E is identified with E).

Thus, we can assume that B, defined on $D(|B|^{1/2})$, is defined on F_1 as well and

$$\|Bu\|_{F_1'} \leq c\|u\|_{(D(|B|^{1/2}))'} \leq c_1\|u\|_{F_1}. \tag{6}$$

Lemma 2. *Let condition* (I) *holds. Then $D(L)$ is densely embedded into F_1 and the operators $L^{-1}, (L^*)^{-1}$ are extensible to the mappings of class $L(F_1', F_1)$.*

Proof. We have

$$\operatorname{Re}(-L^*u, u) \geq \delta_0 \|u\|^2_{F_1}, \tag{7}$$

$$|\operatorname{Re}(-L^*u, u)| \leq |(L^*u, u)| \leq \|L^*u\|_{F_1'} \|u\|_{F_1}, \quad u \in D(L^*). \tag{8}$$

Using (8) on the leftmost side of (3) and reducing $\|u\|_{F_1}$, we infer

$$\|L^*u\|_{F_1'} \geq \delta_0 \|u\|_{F_1}.$$

Since $0 \in \rho(L^*)$ [19], the inequality can be rewritten as

$$\|v\|_{F_1'} \geq \delta_0 \|(L^*)^{-1} v\|_{F_1}, \quad v \in E. \tag{9}$$

Since F_1 is densely embedded into E, E is densely embedded into F_1', i.e. $F_1 \subset E \subset F_1'$ and from (9) it follows that $(L^*)^{-1}$ is extensible to an operator of class $L(F_1', F_1)$. We have

$$(L^{-1}u, u) = (u, (L^*)^{-1}v), \quad u, v \in E. \tag{10}$$

Next, every $u \in E$ satisfies the estimates

$$\begin{aligned} \|L^{-1}u\|_{F_1} &= \sup_{v \in E} \frac{|(L^{-1}u, v)|}{\|v\|_{F_1'}} = \sup_{v \in E} \frac{|(u, (L^*)^{-1}v)|}{\|v\|_{F_1'}} \\ &\leq \sup_{v \in E} \frac{\|u\|_{F_1'} \|(L^*)^{-1}v)\|_{F_1}}{\|v\|_{F_1'}} \leq \sup_{v \in H} \frac{\|u\|_{F_1'}}{\delta_0} \frac{\|v\|_{F_1'}}{\|v\|_{F_1'}}, \end{aligned} \tag{11}$$

$$\|L^{-1}u\|_{F_1} \leq \frac{\|u\|_{F_1'}}{\delta_0}. \tag{12}$$

The estimate (12) implies that L^{-1} extends to an operator of class $L(F_1', F_1)$.

Consider the equality

$$(L^{-1}u, u) = (u, (L^*)^{-1}v), \quad u, v \in E. \tag{13}$$

We have

$$|(u, (L^*)^{-1}v)| \leq \|(L^*)^{-1}v\|_{F_1} \|u\|_{F_1'} \leq c\|v\|_{F_1'} \|u\|_{F_1'}. \tag{14}$$

If u is fixed then $(u, (L^*)^{-1}v)$ is a continuous antilinear functional over F_1' (on v). There exists $g \in F_1$ such that

$$(u, (L^*)^{-1}v) = (g, v) = (L^{-1}u, v), \quad u, v \in E.$$

Thus, $g = L^{-1}u$ and thereby $L^{-1}u \in F_1$, $u \in E$. Take $\psi \in D(L)$. In this case $u = L\psi \in E$ and $\psi = L^{-1}u$.

The facts proven yield $\psi \in F_1$. Therefore, $D(L) \subset F_1$. Assume to the contrary that $D(L)$ is not dense in F_1. There exists $v \in F_1'$, $v \neq 0$, such that

$$(L^{-1}u, v) = 0, \quad u \in E.$$

Passing to the limit we see that the equality $(u, (L^*)^{-1}v) = (L^{-1}u, v)$ is valid for $v \in F_1'$ and $u \in E$. Since the embedding of H in F_1' is dense and $(u, (L^*)^{-1}v) = 0$, we conclude that $(L^*)^{-1}v = 0$. Hence, $v = 0$; a contradiction. Thus, the embedding of $D(L)$ in F_1 is dense. The lemma is proven.

Denote by F_2 the class of functions $u \in F_1$ such that u is representable as $u = L^{-1}v$, where $v \in F_1'$ and

$$\|u\|_{F_2} = \|L^{-1}v\|_{F_1} + \|v\|_{F_1'} = \|u\|_{F_1} + \|Lu\|_{F_1'}.$$

In this case $D(L) \subset F_2 \subset F_1$.

Denote

$$H_1 = \{v \in L_2(0,T; D(L^*)),\ v_t \in L_2(0,T; F_1)\}.$$

Definition 1. A function $u \in L_2(0,T;F_1)$ is called a *generalized solution to the boundary value problem* (1), (2) if there exist $\tilde{u}_0, \tilde{u}_T \in H$ such that $P^-\tilde{u}_0 = \tilde{u}_0$, $P^+\tilde{u}_T = \tilde{u}_T$, and

$$\int_0^T (-(Bu, v_t) - (u, L^*v))\,dt + (Bu_T, v(T)) - (Bu_0, v(0))$$
$$+(B\tilde{u}_T, v(T)) - (B\tilde{u}_0, v(0)) = \int_0^T (f, v)\,dt \qquad (15)$$

for all $v \in L_2(0,T;D(L^*))$ and $v_t \in L_2(0,T;F_1)$.

If u is a generalized solution to (1), (2) in the sense of Definition 1 then u is a generalized solution to (1), (2) in the following more natural sense.

Definition 2. A function $u \in L_2(0,T;F_1)$ is called a *generalized solution to the boundary value problem* (1), (2) if

$$\int_0^T (-(Bu, v_t) - (u, L^*v))\,dt$$
$$+(BP^-u_T, v(T)) - (BP^+u_0, v(0)) = \int_0^T (f, v)\,dt \qquad (16)$$

for every $v \in L_2(0,T;D(L^*))$ such that $v_t \in L_2(0,T;F_1)$, $P^+v(T) = 0$, and $P^-v(0) = 0$.

We expose some consequences of the definition of generalized solution. Let $H_2 = D(L^*)$. In this case if $u \in L_2(0,T;F_1)$ then the expression Lu makes a sense and belongs to $L_2(0,T;H_2')$.

If $v \in C_0^\infty(0,T;H_2)$ then (15) can be rewritten as

$$\int_0^T (-Bu, v_t)\,dt = \int_0^T (Lu, v)\,dt + \int_0^T (f, v)\,dt, \quad v \in C_0^\infty(0,T;H_2). \qquad (17)$$

We have $Lu+f \in L_2(0,T;H_2')$ which fact and the definition of generalized derivative imply that $Bu \in L_2(0,T;F_1')$ (by Lemma 1) has the generalized derivative $(Bu)_t \in L_2(0,T;H_2')$. Since $F_1' \subset H_2'$, we derive that $Bu \in C([0,T];H_2')$ after a possible modification on a set of measure zero. The relation (17) validates the equality $Bu_t - Lu = f$ in $L_2(0,T;H_2')$.

Consider $v \in C^\infty([0,T];H_2)$ in (15). Integrating by parts and involving the equality $Bu_t - Lu = f$, we see that

$$(B(u(T) - u_T - \tilde{u}_T), v(T)) - (B(u(0) - u_0 - \tilde{u}_T), v(0)) = 0.$$

Since the functions $v(T)$ and $v(0)$ are arbitrary, we conclude that

$$Bu(T) = B(u_T + \tilde{u}_T), \quad Bu(0) = B(u_0 + \tilde{u}_0).$$

The equality holds in H_2', since $Bu \in C([0,T];H_2')$; so, the traces $Bu(0)$ and $Bu(T)$ exist and belong to $(D(|B|^{1/2}))'$, while $BP^- u(T) = Bu_T$, and $BP^+ u(0) = Bu_0$.

The following theorem is proven in [23].

Theorem 1. *Let* (I) *and* (II) *hold. Then, for all* $f \in L_2(0,T;F_1')$ *and* $u_0, u_T \in H$, *there exists a generalized solution* $u \in L_2(0,T;F_1)$ *to the boundary value problem* (1), (2) *in the sense of Definition* 1.

Introduce the following additional conditions:

(III) $\operatorname{Re}(-Lu,u) \geq \delta_0 \|u\|_{F_1}^2$, $u \in D(L)$;

(IV) there exist constants $c > 0$ and $\theta \in (0,1)$ such that

$$|(Bu,u)| \leq c\|u\|_{F_1}^{2\theta} \|L^{-1}Bu\|_{F_1}^{2(1-\theta)}, \quad u \in F_1.$$

(V) $B|_{F_1} \in L(F_1, E)$.

Note that $\|L^{-1}Bu\|_{F_1} \leq c\|Bu\|_{F_1'} \leq c_1\|u\|_{F_1}$ (see Lemma 1).

If L obeys the Kato-sectoriality condition and the conditions of Theorem 1 then we can say that (III) and (IV) are excessive, and they always hold.

Let $g(x)$ be a function positive almost everywhere in a domain G. Define the space $L_{2,g}(G;H)$ (with H a Banach space) as the space of strongly measurable functions on G with values in H such that

$$\|u\|_{L_{2,g}(G;H)} = \left(\int\limits_G g(x)\|u(x)\|_H^2 \, dx \right)^{1/2} < \infty.$$

Put $\varphi_i(t) = t^{2i\alpha}(T-t)^{2i\alpha}$, where $\alpha = \frac{1}{2(1-\theta)}$.

Theorem 2. *If* (I)–(IV) *hold and* $\partial_t^i f \in L_{2,\varphi_i}(0,T;F_1')$, $i = 0,1,\dots,m$ *then a generalized solution that of Theorem* 1 *has the generalized derivatives* $\partial_t^i u \in L_{2,\varphi_i}(0,T;F_1)$, $i = 0,1,\dots,m$.

If, moreover, (V) *holds and* $\partial_t^i f \in L_{2,\varphi_{i+1}}(0,T;E)$, $i = 0,1,\dots,$ $m-1$, *then* $u \in L_{2,\varphi_{i+1}}(0,T;D(L))$.

2. Proof of the Main Result

Proof of Theorem 2. Fix numbers $T_1 < T/2 < T_2$. Consider the case of $m = 1$. We have (in $L_2(0,T;H_2')$) that

$$Bu_t - Lu = f. \tag{18}$$

Take

$$\varphi_{0i}(t) = \begin{cases} t^{2i\alpha}(T_2 - t)^{2i\alpha}, & t \in [0;T_2), \\ 0, & t \in [T_2;T], \end{cases} \tag{19}$$

$$\psi_{0i}(t) = \begin{cases} 0, & t \in [0;T_1], \\ (t-T_1)^{2i\alpha}(T-t)^{2i\alpha}, & t \in (T_1;T], \end{cases} \tag{20}$$

where $T_1 < T_2$.

Let $w_1(\eta)$ and $w_2(\eta)$ be averaging kernels with the properties

$$w_i \geq 0, \quad w_i \in C_0^\infty(\mathbb{R}), \quad i = 1, 2,$$

$$\operatorname{supp}(w_1) \subset (1/2, 1), \quad \operatorname{supp}(w_2) \subset (-1, -1/2), \quad \int\limits_{\mathbb{R}} w_i(\eta)\, d\eta = 1, \quad i = 1, 2.$$

Put

$$u_\rho^1 = P_\rho^1 u = \frac{1}{\rho} \int\limits_0^T w_1\left(\frac{\eta - t}{\rho}\right) u(\eta)\, d\eta = \int\limits_0^1 w_1(\xi) u(t + \rho\xi)\, d\xi, \quad t \in [0, T_2], \tag{21}$$

$$u_\rho^2 = P_\rho^2 u = \frac{1}{\rho} \int\limits_0^T w_2\left(\frac{\eta - t}{\rho}\right) u(\eta)\, d\eta = \int\limits_{-1}^0 w_2(\xi) u(t + \rho\xi)\, d\xi, \quad t \in [T_1, T], \tag{22}$$

$$\rho < \min((T - T_2), T_1)/2,$$

$$Bu_{\rho t}^1 = \int\limits_0^1 w_1(\xi) Bu_t(t + \rho\xi)\, d\xi$$

$$= \int\limits_0^1 w_1(\xi) Lu(t + \rho\xi)\, d\xi + \int\limits_0^1 w_1(\xi) f(t + \rho\xi)\, d\xi = Lu_\rho^1 + P_\rho^1 f, \quad t \leq T_2. \tag{23}$$

A similar equality holds for u_ρ^2. Thus, we have on the corresponding segments that

$$Bu_{\rho t}^1 = Lu_\rho^1 + P_\rho^1 f, \quad Bu_{\rho t}^2 = Lu_\rho^2 + P_\rho^2 f. \tag{24}$$

The properties of averaged functions yield $u_\rho^1 \in C^\infty([0, T_2]; F_1)$ and $u_\rho^2 \in C^\infty([T_1, T]; F_1)$. Lemma 1 justifies that $B\partial_t^i u_\rho^1 \in L_2(0, T_2; F_1')$ and $B\partial_t^i u_\rho^2 \in L_2(T_1, T; F_1')$ for every i. From (24) it follows that

$$\partial_t^i u_\rho^j = L^{-1} v, \quad v = B\partial_t^{i+1} u_\rho^j - \partial_t^i P_\rho^j f \in L_2(0, T; F_1'), \quad j = 1, 2, \ i = 1, 2, \ldots. \tag{25}$$

Thus, $\partial_t^i u_\rho^1 \in L_2(0, T_2; F_2)$ and $\partial_t^i u_\rho^2 \in L_2(T_1, T; F_2)$ for every i and the following hold on the corresponding segments:

$$B\partial_t^{i+1} u_\rho^j - L\partial_t^i u_\rho^j = \partial_t^i f_\rho^j, \quad j = 1, 2. \tag{26}$$

We see that

$$\operatorname{Re}(-Lu, u) \geq \delta_0 \|u\|_{F_1}^2, \quad u \in D(L). \tag{27}$$

In view of the density of $D(L)$ in F_2, passing to the limit we arrive at the inequality

$$\operatorname{Re}(-Lu, u) \geq \delta_0 \|u\|_{F_1}^2 \tag{28}$$

valid for all $u \in F_2$.

Since $\sqrt{\varphi_i}\partial_t^i f \in L_2(0, T_2; F_1')$, $i = 0, 1, \ldots, m$, the conventional properties of averaged functions easily imply that

$$P_\rho^1 f \in C^\infty([0, T_2]; F_1), \quad P_\rho^2 f \in C^\infty([T_1, T]; F_1) \tag{29}$$

and, moreover,

$$\sum_{i=0}^{m} \big\|\partial_t^i\big(P_\rho^1 f - f\big)\big\|_{L_{2,\varphi_i}(0,T_2;F_1')} \to 0, \quad \sum_{i=0}^{m} \big\|\partial_t^i\big(P_\rho^2 f - f\big)\big\|_{L_{2,\varphi_i}(T_1,T;F_1')} \to 0 \tag{30}$$

as $\rho \to 0$.

Compose the inner product of (26) for $j = 1$ and $\varphi_{0i}(t)$, take the real part and integrate the result over $(0, T)$. Integrating by parts, we infer

$$\begin{aligned}&\int_0^T \operatorname{Re}\big(-L\partial_t^i u_\rho^1, \partial_t^i u_\rho^1\big)\varphi_{0i}\,dt\\ &= \int_0^T \frac{1}{2}\varphi_{0it} \operatorname{Re}\big(B\partial_t^i u_\rho^1, \partial_t^i u_\rho^1\big)\,dt + \int_0^T \operatorname{Re}\big(\partial_t^i P_\rho^1 f \varphi_{0i}, \partial_t^i u_\rho^1\big)\,dt.\end{aligned} \tag{31}$$

Estimating the right-hand side on using condition (IV), we derive that

$$\begin{aligned}&\int_0^{T_2} \frac{1}{2}|\varphi_{0it}|\big|\big(B\partial_t^i u_\rho^1, \partial_t^i u_\rho^1\big)\big|\,dt \le \int_0^{T_2} |\varphi_{0it}|\big\|\partial_t^i u_\rho^1\big\|_{F_1}^{2\theta}\big\|L^{-1}B\partial_t^i u_\rho^1\big\|_{F_1}^{2(1-\theta)}\,dt\\ &\le c_0\Big(\int_0^{T_2} \varphi_{0i}\big\|\partial_t^i u_\rho^1\big\|_{F_1}^2\,dt\Big)^{\theta}\Big(\int_0^{T_2} \varphi_{0(i-1)}\big\|L^{-1}B\partial_t^i u_\rho^1\big\|_{F_1}^2\,dt\Big)^{1-\theta};\end{aligned} \tag{32}$$

here we have applied the inequality

$$|\varphi_{0it}| \le c_0\varphi_{0i}^{\theta}\varphi_{0(i-1)}^{1-\theta}, \quad t \in (0, T_2).$$

The relation (26) yields

$$\begin{aligned}&\big\|L^{-1}B\partial_t^i u_\rho^1\big\|_{L_{2,\varphi_{0(i-1)}}(0,T_2;F_1)}^2 \le 2\big(\big\|L^{-1}\partial_t^{i-1}P_\rho^1 f\big\|_{L_{2,\varphi_{0(i-1)}}(0,T_2;F_1)}^2\\ &+\big\|\partial_t^{i-1}P_\rho^1 u\big\|_{L_{2,\varphi_{0(i-1)}}(0,T_2;F_1)}^2\big) \le c\big(\big\|\partial_t^{i-1}P_\rho^1 f\big\|_{L_{2,\varphi_{0(i-1)}}(0,T_2;F_1')}^2\\ &+\big\|\partial_t^{i-1}P_\rho^1 u\big\|_{L_{2,\varphi_{0(i-1)}}(0,T_2;F_1)}^2\big).\end{aligned} \tag{33}$$

The Cauchy inequality with a small parameter ε, (32), and (33) imply that

$$\begin{aligned}&\int_0^{T_2} \frac{1}{2}|\varphi_{0it}|\big|\big(B\partial_t^i u_\rho^1, \partial_t^i u_\rho^1\big)\big|\,dt \le \varepsilon\big\|\partial_t^i u_\rho^1\big\|_{L_{2,\varphi_{0i}}(0,T_2;F_1)}^2\\ &+c(\varepsilon)\big(\big\|\partial_t^{i-1}P_\rho^1 f\big\|_{L_{2,\varphi_{0(i-1)}}(0,T_2;F_1')}^2 + \big\|\partial_t^{i-1}u_\rho^1\big\|_{L_{2,\varphi_{0(i-1)}}(0,T_2;F_1)}^2\big).\end{aligned} \tag{34}$$

Similarly,

$$\Big|\int_0^{T_2} \big(\partial_t^i P_\rho^1 f, \varphi_{0i}\partial_t^i u_\rho^1\big)\,dt\Big| \le \varepsilon\big\|\partial_t^i u_\rho^1\big\|_{L_{2,\varphi_{0i}}(0,T_2;F_1)}^2 + c(\varepsilon)\big\|\partial_t^i P_\rho^1 f\big\|_{L_{2,\varphi_{0i}}(0,T_2;F_1')}^2. \tag{35}$$

In view of (31) and (34), (35), choosing a sufficiently small number ε, we find that

$$\delta_0\big\|\partial_t^i u_\rho^1\big\|^2_{L_{2,\varphi_{0i}}(0,T_2;F_1)} \le c\big[\big\|\partial_t^i P_\rho^1 f\big\|^2_{L_{2,\varphi_{0i}}(0,T_2;F_1')} + \big\|\partial_t^{i-1} P_\rho^1 f\big\|^2_{L_{2,\varphi_{0(i-1)}}(0,T_2;F_1')} + \big\|\partial_t^{i-1} u_\rho^1\big\|^2_{L_{2,\varphi_{0(i-1)}}(0,T_2;F_1)}\big] \tag{36}$$

or

$$\big\|\partial_t^i u_\rho^1\big\|^2_{L_{2,\varphi_{0i}}(0,T_2;F_1)} \le c_1\big[\big\|\partial_t^i P_\rho^1 f\big\|^2_{L_{2,\varphi_{0i}}(0,T_2;F_1')} + \big\|\partial_t^{i-1} P_\rho^1 f\big\|^2_{L_{2,\varphi_{0(i-1)}}(0,T_2;F_1')} + \big\|\partial_t^{i-1} u_\rho^1\big\|^2_{L_{2,\varphi_{0(i-1)}}(0,T_2;F_1)}\big], \tag{37}$$

where c_1 is a constant independent of u and ρ. Multiply (36) by q^i, $q \in (0,1)$, and sum the result over i from 1 to m. We see that

$$\sum_{i=1}^m q^i\big\|\partial_t^i u_\rho^1\big\|^2_{L_{2,\varphi_{0i}}(0,T_2;F_1)} \le \sum_{i=1}^m q^i c_1\bigg(\big\|\partial_t^i P_\rho^1 f\big\|^2_{L_{2,\varphi_{0i}}(0,T_2;F_1')} + \big\|\partial_t^{i-1} P_\rho^1 f\big\|^2_{L_{2,\varphi_{0(i-1)}}(0,T_2;F_1')} + \sum_{i=1}^m q^i\big\|\partial_t^{i-1} u_\rho^1\big\|^2_{L_{2,\varphi_{0(i-1)}}(0,T_2;F_1)}\bigg). \tag{38}$$

This inequality can be rewritten as

$$\sum_{i=1}^m (q^i - c_1 q^{i+1})\big\|\partial_t^i u_\rho^1\big\|^2_{L_{2,\varphi_{0i}}(0,T_2;F_1)} \le c_1\sum_{i=1}^m q^i\big(\big\|\partial_t^i P_\rho^1 f\big\|^2_{L_{2,\varphi_{0i}}(0,T_2;F_1')} + \big\|\partial_t^{i-1} P_\rho^1 f\big\|^2_{L_{2,\varphi_{0(i-1)}}(0,T_2;F_1')}\big) + qc_1\|u_\rho^1\|^2_{L_2(0,T_2;F_1)}. \tag{39}$$

Take $q = 1/(2c_1)$. In this case (39) is representable as

$$\sum_{i=1}^m \big\|\partial_t^i u_\rho^1\big\|^2_{L_{2,\varphi_{0i}}(0,T_2;F_1)} \le c_2\sum_{i=0}^m \big\|\partial_t^i P_\rho^1 f\big\|^2_{L_{2,\varphi_{0i}}(0,T_2;F_1')} + c_3\|u_\rho^1\|^2_{L_2(0,T_2;F_1)}. \tag{40}$$

The estimate

$$\|u_\rho^1\|^2_{L_2(0,T_2;F_1)} \le c_4\|u\|^2_{L_2(0,T_2;F_1)}$$

is obvious. From this estimate, (40), and (30) it follows that

$$\sum_{i=1}^m \|\partial_t^i u_\rho^1\|^2_{L_{2,\varphi_{0i}}(0,T_2;F_1)} \le c_4\sum_{i=0}^m \|\partial_t^i f\|^2_{L_{2,\varphi_i}(0,T;F_1')} + c_5\|u\|^2_{L_2(0,T_2;F_1)}. \tag{41}$$

Repeating the arguments for u_ρ^2, we arrive at the estimate

$$\sum_{i=1}^m \big\|\partial_t^i u_\rho^2\big\|^2_{L_{2,\psi_{0i}}(T_1,T;F_1)} \le c_6\sum_{i=0}^m \big\|\partial_t^i f\big\|^2_{L_{2,\varphi_i}(T_1,T;F_1')} + c_7\|u\|^2_{L_2(T_1,T;F_1)}. \tag{42}$$

All constants in (42) and (41) are independent of ρ. Hence, there exists a sequence $\rho_k \to 0$ as $k \to \infty$ such that

$$\partial_t^i u_{\rho_k}^1 \to v_i^1 \in L_{2,\varphi_{0i}}(0,T_2;F_1), \quad \partial_t^i u_{\rho_k}^2 \to v_i^2 \in L_{2,\psi_{0i}}(0,T_2;F_1), \tag{43}$$

where we have the weak convergence in the corresponding spaces. Since $u_{\rho_k}^1 \to u$ in $L_2(0,T_2;F_1)$ and $u_{\rho_k}^2 \to u$ in $L_2(T_1,T;F_1)$, it is easy to justify that there exist the generalized derivatives $\partial_t^i u \in L_{2,\varphi_{0i}}(0,T_2;F_1)$ and $\partial_t^i u \in L_{2,\psi_{0i}}(T_1,T;F_1)$, $i = 1,2,\dots,m$. The last statements easily validate the existence of the generalized derivatives $\partial_t^i u \in L_{2,\varphi_i}(0,T_2;F_1)$. Theorem 2 is proven.

REFERENCES

1. *Case K. M. and Zweifel P. F.* Linear Transport Theory. Reading, MA: Addison-Wesley, 1969.

2. *Cercignani C.* Mathematical Methods in Kinetic Theory. New York: Pergamon Press, 1969.

3. *Cercignani C.* Theory and Applications of the Boltzmann Equation. New York: Elsevier, 1975.

4. *Greenberg W., Van der Mee C. V. M., and Zweifel P. F.* Generalized kinetic equations // Integral Equations Oper. Theory. 1984. V. 7, N 1. P. 60–95.

5. *Van der Mee C. V. M.,* Semigroup and Factorization Methods in Transport Theory, Math. Centrum, Amsterdam (1981) (Math. Centre Tracts, 146).

6. *Van der Mee C. V. M.* Exponentially dichotomous operators and applications. Basel; Boston; Berlin: Birkhäuser-Verlag, 2008. (Oper. Theory Adv. Appl.; V. 182).

7. *Hangelbroek R. J.* Linear analysis and solution of neutron transport problem // Transp. Theory Stat. Phys. 1976. N 5. P. 1–85.

8. *Kaper H. G., Lekkerkerker C. G., and Hejtmanek J.* Spectral Methods in Linear Transport Theory. Basel; Boston; Stuttgart: Birkhäuser Verlag, 1982.

9. *Stephan H.* Nichtgleichgewichtsprozesse: direkte und inverse Probleme. Aachen: Shaker, 1996.

10. *Mokhtar-Kharroubi M.* Mathematical Topics in Neutron Transport Theory. New Approach. Singapore: World Sci. Publ. Co. Pte. Ltd, 1997. (Ser. Adv. Math. Appl. Sci.; V. 46).

11. *Beals R.* Indefinite Sturm–Liouville problems and half-range completeness // J. Differ. Equations. 1985. V. 56, N 3. P. 391–408.

12. *Beals R.* An abstract treatment of some forward-backward problems of transport and scattering // J. Funct. Anal. 1979. V. 34, N 1. P. 1–20.

13. *Beals R. and Protopopescu V.* Half-range completeness for the Fokker–Planck equation // J. Stat. Phys. 1983. V. 32, N 3. P. 391–408.

14. *Latrach K.* Compactness properties for linear transport operator with abstract boundary conditions in slab geometry // Transp. Theory Stat. Phys. 1993. V. 22. P. 39–65.

15. *Latrach K. and Mokhtar-Kharroubi M.* On an unbounded linear operator arising in theory of growing cell population // J. Math. Anal. Appl. 1997. V. 211. P. 273–294.

16. *Webb G.* A model of proliferating cell population with inherited cycle length // J. Math. Biol. 1986. N 23. P. 269–282.

17. *Lar′kin N. A., Novikov V. A., and Yanenko N. N.* Nonlinear Equations of Variable Type [in Russian]. Novosibirsk: Nauka, 1983.

18. *Karabash I. M.* Abstract kinetic equations with positive collision operators // Oper. Theory Adv. Appl. 2008. V. 188. P. 175–195.

19. *Haase M.* The functional calculus for sectorial operators. Basel; Boston; Berlin: Birkhäuser-Verlag, 2006. (Oper. Theory Adv. Appl.; V. 169).

20. *Pyatkov S. G. and Abasheeva N. L.* Solvability of boundary value problems for operator-differential equations of mixed type // Siberian Math. J. 2000. V. 41, N 6. P. 1174–1187.

21. *Dzhuraev T. D.* Boundary Value Problems for Mixed and Mixed-Composite Type Equations [in Russian]. Tashkent: Fan, 1979.

22. Berezanskiĭ Yu. M. Expansions in Eigenfunctions of Selfadjoint Operators. Providence: Amer. Math. Soc., 1968.

23. *Antipin V. I.* Solvability of a boundary value problem for operator-differential equations of mixed type // Siberian Math. J. 2013. V. 54, N 2. P. 185–195.

November 18, 2014

V. I. Antipin
North-Eastern Federal University, Yakutsk, Russia
antvasiv@mail.ru

Yakutian Mathematical Journal
July—September, 2014. Vol. 21, No. 3

UDC 512.643.5

ALMOST ORTHOGONALITY OF INVARIANT SUBSPACES

V. M. Gordienko

Abstract. For an arbitrary matrix it is proven that the invariant subspace corresponding to a localized group of eigenvalues whose modules are close to the norm of this matrix consists of almost eigenvectors and this space is almost orthogonal to other invariant subspaces of the matrix.

Similarly, for a dissipative matrix the invariant subspace corresponding to a localized group of eigenvalues close to the real axis consists of almost eigenvectors and this subspace is almost orthogonal other invariant subspaces of the matrix.

Keywords: eigenvalue, subspace, close to orthogonal, dissipative matrix

This article consists of the two sections independent from each other. In Section 1 we prove that, for an arbitrary matrix, the invariant subspace corresponding to a localized group of eigenvalues whose modules are close to the norm of this matrix consists of almost eigenvectors and is almost orthogonal to other invariant subspaces of the matrix.

In Section 2 we prove that, for a dissipative matrix, the invariant subspace corresponding to a localized group of eigenvalues close to the real axis consists of almost eigenvectors and this subspace is almost orthogonal other invariant subspaces of the matrix.

§ 1. Almost Orthogonality of Invariant Subspaces of an Arbitrary Matrix

Assume that B is a matrix of order n and $\lambda_1, \lambda_2, \dots, \lambda_k$ is a group of eigenvalues (enumerated with multiplicity taken into account) close to each other in the sense that

$$|\lambda_i - \lambda_j| \leq \delta \|B\|$$

whose modules are close to the maximal value; i.e.,

$$\|B\| - |\lambda_j| \leq \varrho \|B\|,$$

with $0 < \delta < \varrho$. Let L be an invariant subspace corresponding to the eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_k$. We demonstrate that for small ϱ the action of B on L obeys the following three conditions:

1) $\|B\|^{-1} B$ is almost unitary on L, i.e.

$$\|(\|B\|^2 I_n - B^* B)\varphi\| \leq 2k\varrho \|B\|^2 \|\varphi\| \quad \text{for } \varphi \in L; \tag{1}$$

The author was supported by the Interdisciplinary Project of the Siberian Division of the Russian Academy of Sciences (Grant 80).

2) every vector $\varphi \in L$ is almost eigenvector, i.e.

$$\|(B-\lambda_j I_n)\varphi\| \le \sqrt{3(k-1)}\sqrt{\varrho}\|B\|\cdot\|\varphi\|; \tag{2}$$

3) the subspace L is almost orthogonal to other invariant subspaces; namely, if M is an invariant subspace of B, while $\mu_1, \mu_2, \dots, \mu_m$ are the corresponding eigenvalues, $|\lambda_i - \mu_j| \ge d\|B\|$ $(d \ge \varrho)$, and $\varphi \in L$, $\psi \in M$ then

$$|\langle\varphi,\psi\rangle| \le \frac{8k\sqrt{\varrho}}{d}\sum_{j=0}^{m-1}\left(\frac{2}{d}\right)^j\|\varphi\|\cdot\|\psi\|. \tag{3}$$

Proceed with the proof of (1). Reduce B to triangular Schur form with the use of a unitary matrix U: $U^*U = I_n$. Let the first k columns of U form a basis for the invariant subspace L. In this case

$$B = U^*\begin{pmatrix} B_1 & B_{12} \\ 0 & B_2 \end{pmatrix}U,$$

where B_1 and B_2 are triangular matrices and

$$B_1 = \begin{pmatrix} \lambda_1 & b_{12} & b_{13} & \cdots & b_{1k} \\ & \lambda_2 & b_{23} & \cdots & b_{2k} \\ & & \ddots & \dots & \vdots \\ & & & \lambda_{k-1} & b_{k-1,k} \\ & & & & \lambda_k \end{pmatrix}.$$

Assume that $\varphi \in L$, in this case $U\varphi = \begin{pmatrix} u \\ 0 \end{pmatrix}$, $u \in \mathbb{C}^k$,

$$\begin{aligned} \langle(\|B\|^2 I_n - B^*B)\varphi,\varphi\rangle &= \langle(\|B\|^2 I_n - B_1^*B_1)u,u\rangle \\ \le \|(\|B\|^2 I_k - B_1^*B_1)\|\cdot\|u\|^2 &= \|(\|B\|^2 I_k - B_1^*B_1)\|\cdot\|\varphi\|^2. \end{aligned}$$

Since $\|B\|^2 I_k - B_1^*B_1 \ge 0$, we infer

$$\begin{aligned} \|(\|B\|^2 I_n - B_1^*B_1)\| &\le \operatorname{tr}(\|B\|^2 I_n - B_1^*B_1) \\ &= \bigl(\|B\|^2 - \bigl|\lambda_1^2\bigr|\bigr) + \bigl(\|B\|^2 - \bigl|\lambda_2^2\bigr| - |b_{12}|^2\bigr) \\ &+ \cdots + \bigl(\|B\|^2 - \bigl|\lambda_k^2\bigr| - |b_{k-1,k}|^2 - \cdots - |b_{1,k}|^2\bigr) \\ \le \sum_{j=1}^{k}\bigl(\|B\|^2 - \bigl|\lambda_j^2\bigr|\bigr) &= \sum_{j=1}^{k}(\|B\| + |\lambda_j|)(\|B\| - |\lambda_j|) \le 2k\|B\|^2\varrho. \end{aligned}$$

Thus,

$$\langle(\|B\|^2 I_n - B^*B)\varphi,\varphi\rangle \le 2k\varrho\|B\|^2\|\varphi\|^2 \quad \text{for } \varphi \in L$$

and so

$$\|(\|B\|^2 I_n - B^*B)\varphi\| \le 2k\varrho\|B\|^2\|\varphi\| \quad \text{for } \varphi \in L.$$

Prove (2). Obviously,

$$\|(B - \lambda_j I_n)\varphi\| \le \|B_1 - \lambda_j I_k\|\cdot\|\varphi\|$$

for $\varphi \in L$. We can estimate the norm of $B_1 - \lambda_j I_k$ by the square root of the sum of squares of modules of all its entries. Since $\|B_1\|^2 I_k - B_1^*B_1 \ge 0$, the diagonal entries of $\|B_1\|^2 I_k - B_1^*B_1$ are nonnegative and

$$\|B_1\|^2 - (|b_{1,j}|^2 + |b_{2,j}|^2 + \cdots + |b_{j-1,j}|^2 + |\lambda_j|^2) \ge 0.$$

Hence,

$$|b_{1,j}|^2 + |b_{2,j}|^2 + \cdots + |b_{j-1,j}|^2 \le \|B_1\|^2 - |\lambda_j|^2 \le \|B\|^2 - |\lambda_j|^2 \le 2\varrho\|B\|^2.$$

Therefore, the sum of squares of all entries beyond the diagonal of B_1 (and $B_1 - \lambda_j I_k$) does not exceed $2(k-1)\varrho\|B\|^2$. Hence,

$$\|B - \lambda_j I_k\| \le \sqrt{(k-1)\delta^2\|B\|^2 + 2(k-1)\varrho\|B\|^2} \le \sqrt{3(k-1)\varrho} \cdot \|B\|.$$

Thus,

$$\|(B - \lambda_j I_n)\varphi\| \le \sqrt{3(k-1)}\sqrt{\varrho}\|B\| \cdot \|\varphi\| \quad \text{for } \varphi \in L.$$

Proceed with the proof of (3). Take $\varphi \in L$ and let $\psi \in M$. Put

$$\varphi_1 = (B - \lambda_1 I_n)\varphi,$$
$$\psi_1 = (B - \mu_1 I_n)\psi, \quad \psi_2 = (B - \mu_2 I_n)\psi_1, \quad \ldots, \quad \psi_j = (B - \mu_j I_n)\psi_{j-1},$$
$$\ldots, \quad \psi_m = (B - \mu_m I_n)\psi_{m-1} = 0.$$

It is easy to check that

$$\langle (\|B\|^2 I_n - B^* B)\varphi, \psi_{j-1}\rangle$$
$$= (\|B\|^2 - \lambda_1\overline{\mu_j})\langle\varphi, \psi_{j-1}\rangle - \lambda_1\langle\varphi, \psi_j\rangle - \overline{\mu_j}\langle\varphi_1, \psi_{j-1}\rangle - \langle\varphi_1, \psi_j\rangle.$$

Here $j = 1, 2, \ldots, m$; by convention $\psi_0 = \psi$. Therefore,

$$|\|B\|^2 - \lambda_1\overline{\mu_j}| \cdot |\langle\varphi, \psi_{j-1}\rangle|$$
$$\le |\langle(\|B\|^2 I_n - B^* B)\varphi, \psi_{j-1}\rangle| + |\lambda_1| \cdot |\langle\varphi, \psi_j\rangle| + |\mu_j| \cdot |\langle\varphi_1, \psi_{j-1}\rangle| + |\langle\varphi_1, \psi_j\rangle|.$$

Using the inequalities $|\lambda_j| \le \|B\|$ and $|\mu_j| \le \|B\|$ together with (1) and (2), we conclude that

$$|\|B\|^2 - \lambda_1\overline{\mu_j}| \cdot |\langle\varphi, \psi_{j-1}\rangle|$$
$$\le 2k\varrho\|B\|^2\|\varphi\| \cdot \|\psi_{j-1}\| + \|B\| \cdot |\langle\varphi, \psi_j\rangle| + 3\sqrt{3(k-1)}\sqrt{\varrho}\|B\|^2\,\|\varphi\| \cdot \|\psi_{j-1}\|$$
$$\le 8k\sqrt{\varrho}\|B\|^2\|\varphi\| \cdot \|\psi_{j-1}\| + \|B\| \cdot |\langle\varphi, \psi_j\rangle|.$$

It is immediate that

$$|\|B\|^2 - \lambda_1\overline{\mu_j}|^2 = \|B\|^2|\lambda_1 - \mu_j|^2 + (\|B\|^2 - |\lambda_1|^2)(\|B\|^2 - |\mu_j|^2)$$

and so

$$|\|B\| - \lambda_1\overline{\mu_j}| \ge \|B\| \cdot |\lambda_1 - \mu_j| \ge d\|B\|^2.$$

Thus,

$$|\langle\varphi, \psi_{j-1}\rangle| \le \frac{8k\sqrt{\varrho}}{d}\|\varphi\| \cdot \|\psi_{j-1}\| + \frac{1}{\|B\|d}|\langle\varphi, \psi_j\rangle|.$$

Write out the last inequality for $j = 1, 2, \ldots, m-1, m$; recall that $\psi_0 = \psi$, $\psi_m = 0$.

$$|\langle\varphi, \psi\rangle| \le \frac{8k\sqrt{\varrho}}{d}\|\varphi\| \cdot \|\psi\| + \frac{1}{\|B\|d}|\langle\varphi, \psi_1\rangle|,$$

$$|\langle\varphi, \psi_1\rangle| \le \frac{8k\sqrt{\varrho}}{d}\|\varphi\| \cdot \|\psi_1\| + \frac{1}{\|B\|d}|\langle\varphi, \psi_2\rangle|,$$

$$\cdots\cdots\cdots\cdots$$

$$|\langle\varphi, \psi_{m-2}\rangle| \le \frac{8k\sqrt{\varrho}}{d}\|\varphi\| \cdot \|\psi_{m-2}\| + \frac{1}{\|B\|d}|\langle\varphi, \psi_{m-1}\rangle|,$$

$$|\langle\varphi, \psi_{m-1}\rangle| \le \frac{8k\sqrt{\varrho}}{d}\|\varphi\| \cdot \|\psi_{m-1}\|.$$

Excluding the expressions $|\langle\varphi, \psi_j\rangle|$ subsequently and using the estimates $\|\psi_j\| \le \|B\|^j\|\psi\|$, we arrive at (3).

§ 2. Almost Orthogonality of Invariant Subspaces of a Dissipative Matrix

Let A be a dissipative matrix of order n, i.e. $\operatorname{Im} A \equiv \frac{1}{2i}[A - A^*] \geq 0$. As is known, all eigenvalues of such matrix lie in the upper half-plane $\operatorname{Im}\lambda \geq 0$. Let $\lambda_1, \lambda_2, \ldots, \lambda_k$ be a group of eigenvalues (enumerated with multiplicity counted) close to each other in the sense that

$$|\lambda_i - \lambda_j| \leq \delta\|A\|$$

and close to the real axis, i.e.,

$$\operatorname{Im}\lambda_j \leq h\|A\|,$$

where $0 < \delta < h$.

Let L be an invariant subspace corresponding to the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$. Prove that for h small the action of the operator A on L obeys the conditions:

1) the matrix A is almost selfadjoint on L, i.e.

$$\frac{1}{2}\|(A - A^*)\varphi\| = \|\operatorname{Im} A\varphi\| \leq kh\|A\| \cdot \|\varphi\| \quad \text{for } \varphi \in L; \tag{4}$$

2) every vector $\varphi \in L$ is almost eigenvector, i.e.,

$$\|(A - \lambda_j I_n)\varphi\| \leq \sqrt{2}kh\|A\| \cdot \|\varphi\|; \tag{5}$$

3) L is almost orthogonal to other invariant subspaces; namely, if M is an invariant subspace of A, $\mu_1, \mu_2, \ldots, \mu_m$ are the eigenvalues, $|\lambda_i - \mu_j| \geq d\|A\|$ $(d \geq h)$, and $\varphi \in L$, $\psi \in M$ then

$$|\langle\varphi, \psi\rangle| \leq \frac{(2+\sqrt{2})kh}{d}\sum_{j=0}^{m-1}\left(\frac{2}{d}\right)^j \|\varphi\| \cdot \|\psi\|. \tag{6}$$

Proceed with the proof of (4). Reduce A to triangular Schur form with the help of a unitary matrix U: $U^*U = I_n$. Let the first k columns of U constitute a basis for L. In this case

$$A = U^*\begin{pmatrix} A_1 & A_{12} \\ 0 & A_2 \end{pmatrix} U,$$

where A_1 and A_2 are triangular matrices and

$$A_1 = \begin{pmatrix} \lambda_1 & a_{12} & a_{13} & \cdots & a_{1k} \\ & \lambda_2 & a_{23} & \cdots & a_{2k} \\ & & \ddots & \ldots & \vdots \\ & & & \lambda_{k-1} & a_{k-1,k} \\ & & & & \lambda_k \end{pmatrix}.$$

Let $\varphi \in L$. In this case $U\varphi = \begin{pmatrix} u \\ 0 \end{pmatrix}$, $u \in \mathbb{C}^k$,

$$\langle \operatorname{Im} A\varphi, \varphi\rangle = \langle \operatorname{Im} A_1 u, u\rangle \leq \|\operatorname{Im} A\varphi_1\| \cdot \|u\|^2 = \|\operatorname{Im} A\varphi_1\| \cdot \|\varphi\|^2.$$

Since $\operatorname{Im} A \geq 0$, we have $\operatorname{Im} A_1 \geq 0$, and so

$$\|\operatorname{Im} A_1\| \leq \operatorname{tr}\operatorname{Im} A_1 = \sum_{j=1}^{k}\operatorname{Im}\lambda_j \leq kh\|A\|.$$

Thus,

$$\langle \operatorname{Im} A\varphi, \varphi\rangle \le kh\|A\|\|\varphi\|^2 \quad \text{for } \varphi \in L$$

implying that

$$\|\operatorname{Im} A\varphi\| \le kh\|A\|\|\varphi\| \quad \text{for } \varphi \in L.$$

Prove (5). Obviously, for $\varphi \in L$,

$$\|(A - \lambda_j I_n)\varphi\| \le \|A_1 - \lambda_j I_k\| \cdot \|\varphi\|.$$

We can estimate the norm of $A_1 - \lambda_j I_k$ by the square root of the sum of squares of modules of all its entries. The inequality $\operatorname{Im} A_1 \ge 0$ ensures that all central minors of second order of $\operatorname{Im} A_1$ are nonnegative, i.e., $|a_{ij}|^2 \le 4 \operatorname{Im}\lambda_i \operatorname{Im}\lambda_j$. Hence,

$$\begin{aligned}\|A_1 - \lambda_j I_k\| &\le \sqrt{(k-1)\delta^2\|A\|^2 + \frac{(k-1)k}{2}4h^2\|A\|^2}\\ &\le \sqrt{(k-1)(2k+1)}h\|A\| \le \sqrt{2}kh\|A\|.\end{aligned}$$

Therefore, we infer

$$\|(A - \lambda_j I_n)\varphi\| \le \sqrt{2}kh\|A\| \cdot \|\varphi\| \quad \text{for } \varphi \in L.$$

Proceed with the proof of (6). Let $\varphi \in L$ and $\psi \in M$. Put

$$\begin{gathered}\varphi_1 = (A - \lambda_1 I_n)\varphi,\\ \psi_1 = (A - \mu_1 I_n)\psi, \quad \psi_2 = (A - \mu_2 I_n)\psi_1, \dots,\\ \psi_j = (A - \mu_j I_n)\psi_{j-1}, \dots, \quad \psi_m = (A - \mu_m I_n)\psi_{m-1} = 0.\end{gathered}$$

It is immediate that

$$\langle \operatorname{Im} A\varphi, \psi_{j-1}\rangle = \frac{1}{2i}[(\lambda_1 - \overline{\mu_j})\langle\varphi, \psi_{j-1}\rangle + \langle\varphi_1, \psi_{j-1}\rangle - \langle\varphi, \psi_j\rangle].$$

Here $j = 1, 2, \dots, m$; by convention $\psi_0 = \psi$. Therefore,

$$|\lambda_1 - \overline{\mu_j}| \cdot |\langle\varphi, \psi_{j-1}\rangle| \le 2|\langle \operatorname{Im} A\varphi, \psi_{j-1}\rangle| + |\langle\varphi_1, \psi_{j-1}\rangle| + |\langle\varphi, \psi_j\rangle|.$$

In view of (4) and (5), we conclude that

$$|\lambda_1 - \overline{\mu_j}| \cdot |\langle\varphi, \psi_{j-1}\rangle| \le (2 + \sqrt{2})kh\|A\| \cdot \|\varphi\| \cdot \|\psi_{j-1}\| + |\langle\varphi, \psi_j\rangle|.$$

It is easy to check that $|\lambda_1 - \overline{\mu_j}|^2 - |\lambda_1 - \mu_j|^2 = 4\operatorname{Im}\lambda_1 \operatorname{Im}\mu_j$, and so

$$|\lambda_1 - \overline{\mu_j}| \ge |\lambda_1 - \mu_j| \ge d\|A\|.$$

Thus,

$$|\langle\varphi, \psi_{j-1}\rangle| \le \frac{(2+\sqrt{2})kh}{d}\|\varphi\| \cdot \|\psi_{j-1}\| + \frac{1}{\|A\|d}|\langle\varphi, \psi_1\rangle|.$$

Write out the last inequality for $j = 1, 2, \dots, m-1, m$ (recall that $\psi_0 = \psi$ and $\psi_m = 0$):

$$\begin{gathered}|\langle\varphi, \psi\rangle| \le \frac{(2+\sqrt{2})kh}{d}\|\varphi\| \cdot \|\psi\| + \frac{1}{\|B\|d}|\langle\varphi, \psi_1\rangle|,\\ |\langle\varphi, \psi_1\rangle| \le \frac{(2+\sqrt{2})kh}{d}\|\varphi\| \cdot \|\psi_1\| + \frac{1}{\|B\|d}|\langle\varphi, \psi_2\rangle|,\\ \dots\dots\dots\dots\\ |\langle\varphi, \psi_{m-2}\rangle| \le \frac{(2+\sqrt{2})kh}{d}\|\varphi\| \cdot \|\psi_{m-2}\| + \frac{1}{\|B\|d}|\langle\varphi, \psi_{m-1}\rangle|,\\ |\langle\varphi, \psi_{m-1}\rangle| \le \frac{(2+\sqrt{2})kh}{d}\|\varphi\| \cdot \|\psi_{m-1}\|.\end{gathered}$$

Excluding the expressions $|\langle\varphi, \psi_j\rangle|$ subsequently and involving the estimates $\|\psi_j\| \le \|B\|^j\|\psi\|$, we arrive at (6).

In conclusion we observe that a version of (6) for the case when all λ_j and all λ_j coincide is proven in [1, p. 419].

REFERENCES

1. *Gohberg I. Ts. and Krein M. G.* An Introduction to the Theory of Linear Nonselfadjoint Operators in Hilbert Space [in Russian]. Moscow: Nauka, 1965.

December 16, 2014

V. M. Gordienko
Sobolev Institute of Mathematics, Novosibirsk, Russia
gordienk@math.nsc.ru

Yakutian Mathematical Journal
July—September, 2014. Vol. 21, No. 3

UDC 517.95

ON THE STATIONARY GALERKIN METHOD FOR A NONCLASSICAL NONLINEAR FORWARD–BACKWARD EQUATION OF THE THIRD ORDER WITH RESPECT TO TIME

I. E. Egorov and E. S. Efimova

Abstract. In the cylinder $Q = \Omega \times (0, T)$ we examine a boundary value problem for a nonclassical nonlinear forward-backward equation of the third order with respect to time. The existence and uniqueness theorem is proven for this boundary value problem. The stationary Galerkin method is involved. It is proven that every weakly convergent subsequence u_μ of a weakly compact sequence of approximate solutions u_m converges weakly to a solution of the boundary value problem.

Keywords: stationary Galerkin method, approximate solution, a priori estimate

Introduction

Boundary value problems for forward-backward equations lie at the frontiers of the theory of nonclassical boundary value problems for equations of mathematical physics. At present the studies of nonclassical boundary value problems for nonlinear forward-backward equations are of a great interest. The reasons for this fact are their applications in hydrodynamics, elasticity, and so on. By now, the nonstationary Galerkin method was used mainly for solving boundary value problems for nonclassical equations while the stationary method was employed only for elliptic-parabolic second order equations.

The study of boundary value problems for forward-backward equations began with Gevrey's articles [1, 2]. For the first time, nonlinear forward-backward parabolic equations were considered by N. N. Yanenko in [3] for describing complicated motions of a viscous fluid. Many articles (see, for instance, [4–8]) are devoted to nonclassical forward-backward equations.

At present, there is no sufficiently complete theory of boundary value problems for nonclassical equations of mathematical physics. Some results of the theory of boundary value problems for a wide class of linear and nonlinear differential equations including parabolic, inverse parabolic, degenerate parabolic equations, forward-backward parabolic equations, and stationary equations were announced in [9].

The stationary Galerkin method is a universal method widely applied to solving boundary value problems for linear and nonlinear elliptic equations of the second

The authors were supported by the Ministry of Education and Science of the Russian Federation Within the Federal Target Program "Scientific and Pedagogical Personnel of Innovative Russia" (Grant 14.132.21.1352) and partially by the State Maintenance Program of Research Activities (Grant 4402).

and higher orders. Some survey of fundamental results in this direction is presented in the well-known monograph [10].

In 1940 G. I. Petrov [11] proposed a generalization of the stationary Galerkin method for an ordinary differential fourth order equation. An approximate solution to the problem is sought as the span of functions satisfying the boundary conditions.

Boundary value problems for forward-backward parabolic equations are studied, for instance, in [12–18].

1. Statement of a Boundary Value Problem and Some Spaces

In the cylinder $Q = \Omega \times (0, T)$ we consider the equation

$$Lu = Pu - \sum_{i=1}^{n} \frac{\partial}{\partial x_i}(|u_{x_i}|^{p-2} u_{x_i}) + c(x)u = f(x,t), \tag{1}$$

where $p > 2$,

$$Pu = \sum_{i=1}^{3} k_i(x,t) D_t^i u,$$

and the coefficients $k_i(x,t)$ and $c(x)$ are sufficiently smooth functions.

Put

$$S_0^{\pm} = \{(x,0) : x \in \Omega,\ -k_3(x,0) \gtrless 0\}, \quad S_T^{\pm} = \{(x,T) : x \in \Omega,\ -k_3(x,T) \gtrless 0\}.$$

Boundary value problem. *Find a solution to* (1) *in* Q *satisfying*

$$u|_{S_T} = 0, \tag{2}$$

$$u|_{t=0} = 0, \quad u|_{t=T} = 0, \quad D_t u|_{\overline{S}_0^+} = 0, \quad D_t u|_{\overline{S}_T^-} = 0. \tag{3}$$

In what follows, we assume that $k_3(x,0) > 0$, $k_3(x,T) < 0$, $x \in \Omega$. Generally, (1) is a forward-backward equation.

Consider the Banach space

$$W_p^1(\Omega) = \{v : v \in L_p(\Omega),\ v_{x_i} \in L_p(\Omega),\ i = \overline{1,n}\}$$

with the norm

$$\|v\|_{W_p^1(\Omega)} = \|v\|_{L_p(\Omega)} + \sum_{i=1}^{n} \|v_{x_i}\|_{L_p(\Omega)},$$

and let $\overset{\circ}{W}{}_p^1(\Omega)$ be the closure of $C_0^\infty(\Omega)$ in the norm of $W_p^1(\Omega)$; let $W_{p'}^{-1}(\Omega)$ stand for the dual space of $\overset{\circ}{W}{}_p^1(\Omega)$, $\frac{1}{p} + \frac{1}{p'} = 1$, and $L_p\big((0,T), \overset{\circ}{W}{}_p^1(\Omega)\big)$ is the space endowed with the norm

$$\|f\| = \bigg(\int_0^T \|f(t)\|^p_{\overset{\circ}{W}{}_p^1(\Omega)}\, dt \bigg)^{1/p}.$$

Given $\varphi \in \overset{\circ}{W}{}_p^1(\Omega)$, put

$$A(\varphi) = -\sum_{i=1}^{n} \frac{\partial}{\partial x_i}(|\varphi_{x_i}|^{p-2} \varphi_{x_i}) + c(x)\varphi = A_0(\varphi) + c(x)\varphi.$$

In this case $\varphi \mapsto A_0(\varphi)$ is bounded as an operator from $\overset{\circ}{W}{}_p^1(\Omega)$ to $W_{p'}^{-1}(\Omega)$ [7, 19].

Put $\beta(\tau) = |\tau|^{(p-2)/2}\tau$.

2. Preliminary Lemmas

Let Ω be a bounded domain with boundary S of class C^∞. Let a function $\psi(x)$ be such that

$$\psi \in C^\infty(\overline{\Omega}), \quad \psi(x) > 0 \text{ for } x \in \Omega, \quad \psi|_S = 0, \quad \frac{\partial \psi}{\partial n}\bigg|_S = 0.$$

In Ω, we examine the Dirichlet problem

$$-\psi\Delta u + \mu u = f, \quad u|_S = 0. \tag{4}$$

Lemma 1 [7]. *If $f \in \overset{\circ}{W}{}_2^k(\Omega)$ then for sufficiently large μ there exists a unique solution to* (4) *such that*

$$u \in \overset{\circ}{W}{}_2^k(\Omega), \quad \sqrt{\psi} D^\alpha u \in L_2(\Omega), \quad |\alpha| = k+1.$$

Let $g_k(t)$, $k = 1, 2, \dots$, be the eigenfunctions of the spectral problem

$$-\frac{d^2 g_k}{dt^2} = \mu_k g_k, \quad g_k(0) = 0, \quad g_k(T) = 0, \tag{5}$$

with positive eigenvalues μ_k such that

$$\int\limits_0^T g_k^2(t) = 1.$$

Let $\xi_j(x)$ be a basis for $\overset{\circ}{W}{}_p^1(\Omega)$ such that $\xi_j \in \overset{\circ}{W}{}_2^k(\Omega)$ for sufficiently large k. Assume that $v_{km}(x)$ are solutions to the problem

$$-\psi\Delta v_{km} + (\lambda + \mu_k)v_{km} = \xi_m, \quad v_{km}|_S = 0, \tag{6}$$

where λ is a number and $m = 1, 2, \dots$.

Lemma 1 implies

Lemma 2 [7]. *For sufficiently large $\lambda > 0$,* (6) *has the unique smooth solution in $\overline{\Omega}$.*

Put

$$B\varphi \equiv -\frac{\partial^2 \varphi}{\partial t^2} - \psi\Delta\varphi + \lambda\varphi,$$

where λ is that of Lemma 2.

Lemma 3. *There exists a basis $\{\varphi_j\}$ of sufficiently smooth functions in $\overline{Q}$ such that $\{B\varphi_j\}$ form a basis for $L_p\big((0,T), \overset{\circ}{W}{}_p^1(\Omega)\big)$.*

The proof of Lemma 3 (which is similar to that of [7]) results from (5), (6), Lemma 2, and the equality

$$B(v_{km} \otimes g_k) = \xi_m \otimes g_k,$$

where $\otimes$ is the direct product.

3. The Main Result

Theorem 1. *Assume that*

$$k_3(x,0) > 0, \quad k_3(x,T) < 0, \quad x \in \Omega,$$

$$f, \sqrt{\psi}\frac{\partial f}{\partial x_i} \in L_2(Q), \ i = \overline{1,n},$$

$$-k_2 + \frac{1}{2}k_{3t} \geq \delta > 0, \quad -k_2 + \frac{3}{2}k_{3t} \geq \delta > 0, \quad \sum_{i=1}^{n}((k_3\psi)_{x_i})^2 \leq \frac{1}{4}\delta^2\psi,$$

and the coefficient $c(x) > 0$ is sufficiently large.

Then there exists a unique function $u(x,t)$ satisfying (1)–(3) *and such that*

$$u \in L_p\big((0,T); \overset{\circ}{W}{}^1_p(\Omega)\big),$$

$$u_t, u_{tt} \in L_2(Q), \quad k_3 D_t^3 u \in L_{p'}\big((0,T); W_{p'}^{-1}(\Omega)\big),$$

$$u_{tt}(x,0) \in L_2(Q), \quad u_{tt}(x,T) \in L_2(Q),$$

$$\sqrt{\psi}\frac{\partial}{\partial x_j}(|u_{x_i}|^{(p-2)/2}u_{x_i}) \in L_2(Q), \quad i,j = \overline{1,n},$$

$$\sqrt{\psi}v_{tx_i} \in L_2(Q), \quad \frac{\partial}{\partial t}(|u_{x_i}|^{(p-2)/2}u_{x_i}) \in L_2(Q), \quad i = \overline{1,n}.$$

Proof. Approximate solutions to (1)–(3) are sought in the form

$$u_m(x,t) = \sum_{k=1}^{m} c_k^m \varphi_k(x,t)$$

from simultaneous nonlinear algebraic equations

$$\int_0^T (Pu_m + Au_m, B\varphi_j)_0\,dt = \int_0^T (f, B\varphi_j)_0\,dt, \quad j = \overline{1,m}, \tag{7}$$

where

$$(u,v)_0 = \int_\Omega uv\,dx, \quad u,v \in L_2(\Omega).$$

First, we prove the existence of a function $u_m \equiv v$ satisfying (7), relying upon the inequality

$$\int_0^T (Pv + Av, Bv)_0\,dt \geq c\bigg\{\int_0^T \bigg[v_{tt}^2 + v_t^2 + \sum_{i=1}^{n}|v_{x_i}|^p + \sum_{i=1}^{n}\bigg(\frac{\partial}{\partial t}(\beta(v_{x_i}))\bigg)^2$$
$$+\psi\sum_{i=1}^{n}v_{tx_i}^2 + \sum_{i,j=1}^{n}\psi\bigg(\frac{\partial}{\partial x_j}(\beta(v_{x_i}))\bigg)^2\bigg]\,dQ + \int_\Omega \big[v_{tt}^2(x,0) + v_{tt}^2(x,T)\big]\,dx\bigg\}, \tag{8}$$

where $c > 0$.

Indeed,

$$\bigg|\int_0^T (f, Bv)_0\,dt\bigg| \leq c_*\bigg[\int_Q \bigg(v_{tt}^2 + \sum_{i=1}^{n}v_{x_i}^2\bigg)\,dQ\bigg]^{1/2}$$

with a constant $c_* > 0$. Hence,

$$\int\limits_0^T (Pu_m + Au_m - f, Bu_m)_0 \, dt \geq 0,$$

whenever $\int_Q \left(u_{mtt}^2 + \sum_{i=1}^n |u_{mx_i}|^2\right) dQ$ is sufficiently large. In this case (7) is solvable by Lemma 4.3 in [7, Chapter 1].

We have

$$\int\limits_0^T (Pv + Av, Bv)_0 \, dt = \sum_{k=1}^6 I_k, \tag{9}$$

where

$$I_1 = \int\limits_0^T (Pv, -v_{tt})_0 \, dt, \quad I_2 = \int\limits_0^T (Pv, -\psi \Delta v)_0 \, dt, \quad I_3 = \lambda \int\limits_0^T (Pv, v)_0 \, dt,$$

$$I_4 = \int\limits_0^T (A(v), -v_{tt})_0 \, dt, \quad I_5 = \int\limits_0^T (A(v), -\psi \Delta v)_0 \, dt, \quad I_6 = \lambda \int\limits_0^T (Av, v)_0 \, dt.$$

Transform I_k by using the conditions of the theorem and the available inequalities in [7]. We have

$$I_1 = \int\limits_Q \left[-\left(k_2 + \frac{1}{2} k_{3t}\right) v_{tt}^2 - k_1 v_t v_{tt} \right] dQ - \frac{1}{2} \int\limits_\Omega k_3 v_{tt}^2 \, dx \Big|_{t=0}^{t=T}$$

$$\geq \frac{\delta}{2} \int\limits_Q v_{tt}^2 \, dQ + C_0 \int\limits_\Omega \left[v_{tt}^2(x,0) + v_{tt}^2(x,T)\right] dx - C_1(\delta) \int\limits_Q v_t^2 \, dQ, \quad C_0, C_1(\delta) > 0. \tag{10}$$

After transformations we infer

$$\begin{aligned} I_2 = \int\limits_Q \Bigg\{ \left(-k_2 + \frac{3}{2} k_{3t} \right) \psi \sum_{i=1}^n v_{tx_i}^2 - v_{tt} \sum_{i=1}^n (k_3 \psi)_{x_i} v_{tx_i} \\ - \frac{1}{2} (k_{1t} - k_{2tt} + k_{3ttt}) \psi \sum_{i=1}^n v_{x_i}^2 + \frac{1}{2} \Delta(k_2 \psi) v_t^2 \\ - \sum_{i=1}^n [(k_{3t}\psi)_{x_i} v_{tt} + (k_{2t}\psi)_{x_i} v_t - (k_1 \psi)_{x_i} v_t] v_{x_i} \Bigg\} dQ - \frac{1}{2} \int\limits_\Omega k_3 \psi \sum_{i=1}^n v_{tx_i}^2 \, dx \Big|_{t=0}^{t=T}. \end{aligned}$$

Hence,

$$I_2 \geq \int\limits_Q \left[\frac{\delta}{2} \psi \sum_{i=1}^n v_{tx_i}^2 - \frac{\delta}{4} v_{tt}^2 - C_2 \left(v_t^2 + \sum_{i=1}^n v_{x_i}^2 \right) \right] dQ, \quad C_2 > 0. \tag{11}$$

Next,

$$I_3 = \lambda \int\limits_Q \left[\left(-k_2 + \frac{3}{2} k_{3t} \right) v_t^2 + \frac{1}{2} (-k_{1t} + k_{2tt} - k_{3ttt}) v^2 \right] dQ - \frac{1}{2} \lambda \int\limits_\Omega k_3 v_t^2 \, dx \Big|_{t=0}^{t=T}$$

and thus

$$I_3 \geq \lambda \int\limits_Q \left[\delta v_t^2 - C_3 v^2\right] dQ, \quad C_3 > 0. \tag{12}$$

We have

$$I_4 = \frac{4(p-1)}{p^2} \int\limits_Q \sum_{i=1}^n \left(\frac{\partial}{\partial t}(\beta(v_{x_i}))\right)^2 dQ + \frac{1}{2} \int\limits_Q c(x) v_t^2 \, dQ. \tag{13}$$

After some transformations we conclude that

$$I_5 = \int\limits_0^T (A_0(v), -\psi\Delta v)_0 \, dt + \int\limits_Q \left[c\psi \sum_{i=1}^n v_{x_i}^2 - \frac{1}{2}\Delta(c\psi) v^2\right] dQ,$$

and so

$$I_5 \geq C_4 \int\limits_Q \psi \sum_{i,j=1}^n \left(\frac{\partial}{\partial x_j}\beta(v_{x_i})\right)^2 dQ - C_5 \int\limits_Q \left[\sum_{i=1}^n v_{x_i}^2 + v^2\right] dQ, \quad C_4, C_5 > 0. \tag{14}$$

We have

$$I_6 = \lambda \int\limits_Q \left[\sum_{i=1}^n |v_{x_i}|^p + c(x) v^2\right] dQ. \tag{15}$$

In view of (10)–(15) and (9), we obtain

$$\begin{aligned}\int\limits_0^T (Pv + Av, Bv)_0 \, dt \geq \int\limits_Q \Bigg\{ & \frac{\delta}{4} v_{tt}^2 + \lambda \sum_{i=1}^n |v_{x_i}|^p + \Bigg[\frac{\delta}{2}\psi \sum_{i=1}^n v_{tx_i}^2 \\ & + \frac{4(p-1)}{p^2} \sum_{i=1}^n \left(\frac{\partial}{\partial t}\beta(v_{x_i})\right)^2 + C_4 \psi \sum_{i,j=1}^n \left(\frac{\partial}{\partial x_j}\beta(v_{x_i})\right)^2\Bigg] - (C_2 + C_5) \sum_{i=1}^n v_{x_i}^2 \\ & + \left(\lambda\delta + \frac{1}{2}c(x) - C_1 - C_2\right) v_t^2 + [\lambda(c(x) - C_3) - C_5] v^2 \Bigg\} \, dQ.\end{aligned} \tag{16}$$

Basing on the inequalities

$$\int\limits_Q v^2 \, dQ \leq k_1 \int\limits_Q \sum_{i=1}^n v_{x_i}^2 \, dQ, \quad \int\limits_Q \sum_{i=1}^n v_{x_i}^2 \, dQ \leq k_2 \int\limits_Q \sum_{i=1}^n |v_{x_i}|^p \, dQ, \quad k_1, k_2 > 0,$$

and (16), we arrive at (8) for

$$c(x) \geq C_3, \quad \lambda\delta \geq C_1 + C_2, \quad \lambda > C_5 k_1 k_2 + (C_2 + C_5) k_2.$$

Demonstrate that (8) ensures the existence of a solution to (1)–(3). Indeed, (8) implies that the sequence u_m (respectively, u_{mt} and u_{mtt}) is bounded in $L_p\big((0,T); \overset{\circ}{W}{}_p^1(\Omega)\big)$ (respectively, in $L_2(Q)$), the sequences $u_{mtt}(x,0)$ and $u_{mtt}(x,T)$ are bounded in $L_2(\Omega)$, and

$$\frac{\partial}{\partial t}(\beta(u_{mx_i})), \ \sqrt{\psi}\frac{\partial}{\partial x_j}(\beta(u_{mx_i})), \ \sqrt{\psi} u_{mtx_i} \quad \text{are bounded in } L_2(Q). \tag{17}$$

Thus, the sequence $\left|\frac{\partial u_m}{\partial x_i}\right|^{p-2}\frac{\partial u_m}{\partial x_i}$ is bounded in $L^{p'}(Q)$. Let G be an arbitrary domain in Ω such that $\overline{G}\subset\Omega$. In view of (17), the sequence $\beta(u_{mx_i})$ is bounded in $W_2^1(G\times(0,T))$. By compactness of the embedding of $W_2^1(G\times(0,T))$ in $L_2(G\times(0,T))$, there exists a subsequence u_μ of u_m such that

$$u_\mu\to u \quad \text{weakly in } L_p\big((0,T);\overset{\circ}{W}{}_p^1(\Omega)\big),$$

$$u_{\mu t}\to u_t,\ u_{\mu tt}\to u_{tt} \quad \text{weakly in } L_2(Q),$$

$$u_{\mu tt}(x,0)\to\chi_0(x),\ u_{\mu tt}(x,T)\to\chi_1(x) \quad \text{weakly in } L_2(\Omega),$$

$$\beta(u_{\mu x_i}) \quad \text{converge almost everywhere in } Q,$$

$$\frac{\partial}{\partial t}(\beta(u_{\mu x_i}))\to\xi_i \quad \text{weakly in } L_2(Q),$$

$$\sqrt{\psi}\frac{\partial}{\partial x_j}(\beta(u_{\mu x_i}))\to\eta_{ij} \quad \text{weakly in } L_2(Q),$$

$$|u_{\mu x_i}|^{p-2}u_{\mu x_i}\to\eta_i \quad \text{weakly in } L_{p'}(Q).$$

Thereby, the monotonicity of $\beta(\tau)$ implies that the subsequence $u_{\mu x_i}$ converges almost everywhere in Q. Lemma 1.3 in [7, Chapter 1] yields

$$\xi_i=\frac{\partial}{\partial t}(\beta(u_{x_i})),\quad \eta_{ij}=\sqrt{\psi}\frac{\partial}{\partial x_j}(\beta(u_{x_i})),\quad \eta_i=|u_{x_i}|^{p-2}u_{x_i}.$$

In this case $A_0(u_\mu)\to A(u)$ in $L_{p'}((0,T);W_{p'}^{-1}(\Omega))$, and (7) ensures that

$$\int_0^T\left[(-u_{tt},(k_3B\varphi_j)_t)_0+\left(\sum_{i=1}^2 k_iD_t^iu+Au,B\varphi_j\right)_0\right]dt=\int_0^T(f,B\varphi_j)_0\,dt,\quad j=\overline{1,n}. \tag{18}$$

Since $\{B\varphi_j\}$ is a basis for $L_p\big((0,T);\overset{\circ}{W}{}_p^1(\Omega)\big)$; therefore, (18) implies that $u(x,t)$ meets (1) and the boundary conditions (2), (3) and, moreover, $k_3D_t^3u\in L_{p'}\big((0,T);W_{p'}^{-1}(\Omega)\big)$.

By the trace theorems [20, Chapter 1], the traces $u_{tt}(x,0)u_{tt}(x,T)$ exist such that $u_{tt}(x,0)=\chi_0(x)\in L_2(\Omega)$ and $u_{tt}(x,T)=\chi_1(x)\in L_2(\Omega)$.

Proceed with the proof of uniqueness in Theorem 1. Let u_1 and u_2 be two solutions to the boundary value problem (1)–(3) from the class of Theorem 1. We have

$$Pu+A_0(u_1)-A_0(u_2)+c(x)u=0,$$

where $u=u_1-u_2$. The inequality $(A_0(u_1)-A_0(u_2),u_1-u_2)_0\geq 0$ yields

$$\int_Q\left[\delta u_t^2+(c(x)-C_3)u^2\right]dQ\leq 0.$$

Hence, $u=0$ whenever $c(x)\geq C_3$. Theorem 1 is proven.

In the proof of Theorem 1 we apply the stationary Galerkin method. We could not however establish an estimate of the error $u-u_m$ in $L_p\big((0,T);\overset{\circ}{W}{}_p^1(\Omega)\big)$ due to the strong nonlinearity of A_0u.

REFERENCES

1. *Gevrey M.* Sur les équations aux dérivées partielles du type parabolique // J. Math. Pures Appl. Ser. 6. 1913. V. 9. P. 305–471.

2. *Gevrey M.* Sur les équations aux dérivées partielles du type parabolique (suite) // J. Math. Pures Appl. Ser. 6. 1914. V. 10. P. 105–148.

3. *Lar′kin N. A., Novikov V. A., and Yanenko N. N.* Nonlinear Equations of Variable Type [in Russian]. Novosibirsk: Nauka, 1983.

4. *Tersenov S. A.* Forward-Backward Parabolic Equations [in Russian]. Novosibirsk: Nauka, 1985.

5. *Egorov I. E. and Fëdorov V. E.* Higher-Order Nonclassical Equations of Mathematical Physics [in Russian]. Novosibirsk: Vychisl. Tsentr Sibirsk. Otdel. Ros. Akad. Nauk, 1995.

6. *Beals R. and Protopopescu V.* Half-range completeness for the Fokker–Planck equation // J. Stat. Phys. 1983. V. 32, N 3. P. 565–584.

7. *Lions J.-L.* Some Methods for Solving Nonlinear Boundary Value Problems [Russian translation]. Moscow: Mir, 1973.

8. *Egorov I. E., Pyatkov S. G., and Popov S. V.* Nonclassical Operator-Differential Equations [in Russian]. Novosibirsk: Nauka, 2000.

9. *Glazatov S. N.* On solvability of nonclassical boundary value problems for differential equations of variable type // Nonclassical Equations of Mathematical Physics [in Russian], IV Siberian Congress on Applied and Industrial Mathematics (INPRIM-2000) (Novosibirsk, July 26–July 1 2000). Novosibirsk: Inst. Mat. (Novosibirsk), 2000. P. pp. 18–24.

10. *Ladyzhenskaya O. A. and Ural′tseva N. N.* Linear and Quasilinear Elliptic Equations. New York and London: Academic Press, 1968.

11. *Petrov G. I.,* “Application of the Galerkin method to the stability problem of a flow of a viscous fluid,” Prikl. Mat. Mekh., Vol. 4, No. 3 (1940).

12. *Egorov I. E. and Efimova E. S.* The stationary Galerkin method for a forward-backward parabolic equation // Mat. Zametki YaGU. 2011. V. 18, N 2. P. 41–46.

13. *Oleĭnik O. A. and Radkevich E. V.* Second-Order Equations with Nonnegative Characteristic Form [in Russian]. Moscow: Moscow Univ., 2010.

14. *Fichera G.* To the unique theory of boundary value problems for elliptic parabolic equations // Matematika. 1963. V. 7, N 6. P. 99–121.

15. *Egorov I. E. and Stepanova P. I.* On the Galerkin method for elliptic parabolic equations // Mat. Zametki YaGU. 2008. V. 15, N 2. P. 19–26.

16. *Egorov I. E.* Application of the Galerkin method to the third boundary value problem for an elliptic parabolic equation // Mat. Zametki YaGU. 2009. V. 16, N 1. P. 22–27.

17. *Nakhushev A. M.* Problems with a Shift for Partial Differential Equations [in Russian]. Moscow: Nauka, 2006.

18. *Popov S. V.* On solvability of the boundary value problem for a third order forward-backward equation // Differential Equations and Some of Their Applications [in Russian]. Yakutsk: YaF SO RAN SSSR, 1989. P. pp. 39–47.

19. *Gajewski H., Gröger K., and Zacharias K.* Nonlinear Operator Equations and Operator Differential Equations [Russian translation]. Moscow: Mir, 1978.

20. *Triebel H.* Interpolation Theory; Function Spaces; Differential Operators. Berlin: VEB Deutcher Verl. Wiss., 1977.

November 19, 2014

I. E. Egorov
North-Eastern Federal University
Institute of Mathematics, Yakutsk, Russia
`IvanEgorov51@mail.ru`

E. S. Efimova
North-Eastern Federal University
Institute of Mathematics, Yakutsk, Russia
`OslamE@mail.ru`

Yakutian Mathematical Journal
July—September, 2014. Vol. 21, No. 3

UDC 517.925.54:517.962.27/.8

EQUATIONS OF IDENTIFICATION METHODS FOR LINEAR DIFFERENTIAL EQUATIONS

A. O. Egorshin

Abstract. We consider a variational approach to identification of stationary linear dynamical models and compare it to other available approaches to estimating the coefficients of linear models of dynamical objects using the results of observations: orthogonal regression and algebraic identification methods. We express the estimates that are provided by these methods as functions of the length of the observation interval.

Keywords: dynamical piecewise-linear approximation, variational identification, algebraic identification, orthogonal regression, dynamical model, real-time identification

1. Introduction

Identification is a term for estimating the parameters of differential or difference equations of some dynamical process, which could be designed, studied, controlled, predicted using experimental or synthetic (initial) data. Sometimes this term is applied to estimating the characteristics of an integral description of a dynamical object, for instance, its momentum function or Green's function.

This article addresses variational *parametric* problems of identification and mathematical modeling, namely, estimating the coefficients of differential and difference equations of a certain class. We obtain nonlinear difference equations for some estimates for these equations as functions of the length of the observation interval of the object of identification. The class of models we use consists of ordinary linear differential or difference equations with constant coefficients on the finite interval under study (autonomous equations).

The identification problems, i.e., constructing mathematical models and estimating their characteristics with the use of initial data related to the object we study, are classified as *inverse* modeling problems in contrast to *direct* problems of *simulation*, which means reproducing the behavior of the processes modelled using specified descriptions of it, often approximate. We refer to inverse modeling problems as *mathematical* modeling. While stating the problems of mathematical modeling, the preliminary choice of the class of models is one of the methods for regularizing ill-posed inverse problems of reconstructing the operators of actual physical objects.

Identification refers sometimes to some special class of problems: the mathematical problems of estimating the characteristics of dynamical objects with *stochastic* initial data [1, p. 242]. In particular, these are the problems of mathematical modeling in which the models of *measurement* errors as random variables or processes are also given. In these problems the structure (the form of equations) of the object

The author was supported by the Russian Fund for Basic Research (Grant 13–01–00329) and Siberian Division of the Russian Academy of Sciences (Interdisciplinary Project No. 80).

is assumed to be known and *coincide* with the structure of the model. In these estimation problems, a rigorous study of the statistical properties of estimates is possible at least in principle. We use the term “identification” in a wider sense: as the mathematical modeling of dynamical objects or processes with unknown and possibly more complicated description than the assumed model. This is obviously typical in most applications.

Therefore, the problems of constructing mathematical models with *undetermined* errors in initial data are of great practical value. These are errors for which mathematical models are absent. As a rule, for these errors there are no theoretical methods for studying the results (properties of estimates) in the corresponding estimation and optimization problems. An important kind of undetermined errors are those related to the property that the object is, as a rule, more complicated than the model. This causes inevitable *structural* errors of the representation of the dynamical process under study by even the best model of the chosen class in the absence of measurement errors.

For both random measurement errors and inevitable undetermined errors, in particular structural errors in the initial data, it is natural to pose the corresponding optimization problems of mathematical modeling as problems of *approximating* the object by a model in a certain class. This is how we pose them in this article.

Most often, in particular for analytical reasons, we use the mean-square criteria for the quality of approximation of an object by a simpler model [2, p. 10, p. 16, p. 201; 3, 4]. Furthermore, the problems of mathematical modeling and identification become problems of projection onto (seeking the nearest element of) certain admissible sets of the model in the corresponding Euclidean or Hilbert space of initial data [5].

2. Notation

Suppose that on the interval $I_t = [0,t]$ of observation with the uniform h-mesh I_h of $L+1$ points $\tau_i = ih$, for $i = \overline{0,L}$, (i.e., $L = t/h$) there are specified readings $y_i = y(\tau_i)$, possibly with uncontrollable errors of measurement and/or structure (see above), of some solution $y(\tau)$ for $\tau \in I_t$ to a differential equation or a certain dynamical process under study. The readings constitute a vector $\mathbf{y} = \mathbf{y}_L = \{y_i\}_0^L \in E = E_L = E^{L+1}$ or a finite sequence $\mathbf{y} \in l^2[0,L] = l^2$. In the square norm on E and l^2; i.e., when

$$\|\mathbf{y}\|^2 = \sum_0^L |y_i|^2,$$

these are equivalent objects.

Denote the inner product of $\mathbf{x}$ and $\mathbf{y}$ in E and l^2 by $\langle \mathbf{y}, \mathbf{x}\rangle = \langle \mathbf{y}, \mathbf{x}\rangle_R = \mathbf{x}^* R\mathbf{y}$, where R is a positive definite selfadjoint matrix; the latter we express as $R > 0$. The involution $\mathbf{x}^*$ is the composition of the two involutions: the matrix transpose and complex conjugation. The symbol $\mathbf{x}'$ indicates an element of the dual space E'. Then, $x' = x^* R = \langle \cdot, \mathbf{x}\rangle$. The operators $K : E \to E$ satisfy $K' = R^{-1}K^*R$. Assume for simplicity that $R = I_L = I$ is the identity matrix. Then, $\langle \mathbf{y}, \mathbf{x}\rangle = \mathbf{y}'\mathbf{x} = \mathbf{x}^*\mathbf{y}$ and $\mathbf{x}' = \langle \cdot, \mathbf{x}\rangle = \mathbf{x}^*$; moreover, $K' = K^*$. Put $\|\mathbf{y}\|^2 = \|\mathbf{y}\|^2_{=}\langle \mathbf{y}, \mathbf{y}\rangle = \mathbf{y}'\mathbf{y} = \mathbf{y}^*\mathbf{y}$.

Define functionals on E via the inner product. Denote the standard basis vectors for E by e_i and for E', by $e_i' = \langle \cdot, e_i\rangle = e^*$, for $i = \overline{0,L}$. Then the components y_i of a vector $\mathbf{y} = \{y_i\}_0^L \in E$ are defined as $y_i = \langle \mathbf{y}, e_i\rangle = e_i'\mathbf{y} = e_i^*\mathbf{y}$.

Agree on the notation for vectors and matrices as sets. The constructs of the form $\{x_i\}_k^m$ stand for column vectors, $|x_j|_l^n$ or $|x_l, \dots, x_n|$ for row vectors, while

$\{x_{ij}\}_{kl}^{mn}$ for matrices (i the row index) of a set of components of the form x. If the elements x_i in $\{x_i\}_k^m$ are row $(n-l+1)$-vectors, while x_j in $|x_j|_l^n$ or $|x_l,\dots,x_n|$ are column $(m-k+1)$-vectors, then both constructs amount to the matrix $\{x_{ij}\}_{kl}^{mn}$.

Denote by $\mathbb{E}_{\overline{k,l}} = |e_k,\dots,e_l| = |e_i|_k^l$ the matrix of size $(L+1)\times(l-k+1)$ consisting of the corresponding basis vectors in E. This is also the operator $\mathbb{E}_{\overline{k,l}} : E^{l-k+1} \to E$. The collection of the corresponding basis vectors in E' is the matrix $\{e_i'\}_k^l$ of size $(l-k+1)\times(L+1)$. This is also the operator $\mathbb{E}'_{\overline{k,l}} : E \to E^{l-k+1}$. Assume that e_i are the standard unit basis vectors with 1 in slot i and 0 elsewhere: $e_i = \{\delta_{i,j}\}_0^L$ (here $\delta_{i,j}$ is the Kronecker symbol). Then we can write $\mathbb{E}_{\overline{kl}} = E_{\overline{kl}}$.

Since we have agreed that $R = I$, it follows that $\mathbb{E}_{\overline{0,L}} = \mathbb{E} = E = I_L = I$ is the identity matrix of size $L+1$, while

$$\mathbb{E}'_{\overline{k,l}} = E^*_{\overline{k,l}} = |0_{\overline{l-k+1,k}}, I_{l-k+1}, 0_{\overline{l-k+1,L-l}}| : E \to E^{l-k}$$

is the so-called excision matrix. Use the notation $0_{k,l}$ for the zero matrix of size $k\times l$ and 0_k for the zero vector of length k. Introduce also the operator of downward translation as $I^1 e_k = e_{k+1}$. Its matrix is $I^1 = \{\delta_{i-1,j}\}_0^L$; furthermore, $I^* = I^{-1}$. We can also define the translation operator as $sy_k = y_{k+1}$. Clearly, $I^{-1} = s$.

3. A Variational Problem of Approximation

The problem of mathematical modeling of a process $\mathbf{y}$ is posed as a variational problem of its dynamical piecewise-linear approximation. This refers to the problem of best mean-square approximation to the function $\mathbf{y}$ on a finite interval by a function $\hat{\mathbf{y}}$ satisfying on this interval the *linear* difference equation

$$D\hat{\mathbf{y}} = D_\alpha\hat{\mathbf{y}} = \left\{\sum_{i=0}^{n} \hat{y}_{k+i}\alpha_i^*\right\}_0^N = 0 = \left\{\sum_{i=0}^{n} \langle \hat{\mathbf{y}}, I^i e_k \alpha_i\rangle\right\}_0^N, \quad N = L-n, \tag{1}$$

with constant coefficients.

Given n local conditions (initial conditions $[y]_0 = \{y_i\}_0^{n-1}$, terminal conditions $[y]_{N+1} = \{y_i\}_{N+1}^L$, or other coordinates in $\ker D$) which determine the projection $\hat{\mathbf{y}}_\alpha \in \ker D_\alpha \subset E$, as well as the coefficients $\alpha^* = |\alpha_i^*|_0^n = |\alpha_0^*,\dots,\alpha_n^*|$ of (1) of the process $\hat{\mathbf{y}}$, we minimize the distance

$$J = J(\hat{\mathbf{y}}) = \|\mathbf{y}-\hat{\mathbf{y}}\|^2 = \|\mathbf{y}-\hat{\mathbf{y}}\|_E^2 = \|\mathbf{y}-\hat{\mathbf{y}}\|_{l^2}^2 = \sum_0^L |y_i - \hat{y}_i|^2. \tag{2}$$

The problem (1), (2) generalizes the classical n-polynomial approximation problem which also involves criterion (2), but the minimization condition uses

$$\Delta^n\hat{\mathbf{y}} = 0 = \sum_0^n \hat{y}_{k+i}(-1)^{n-i}C_n^i$$

instead of (1). This equation with given coefficients is a particular case of (1) with $\alpha_\Delta^* = |(-1)^{n-i}C_n^i|_0^n$.

The vector $\alpha \in \omega$, where ω is an admissible set for the values of α, is estimated in (1), (2) as a direction in E^{n+1}. If $\alpha_n \neq 0$ then we may assume that $\alpha_n = 1$. Then, $\omega \subset \mathbb{G} = \{\vartheta \in E^{n+1} : \vartheta_n = 1\}$ and the coefficients $\bar{\alpha}^* = |\alpha_i^*|_0^{n-1} = |\alpha_0^*,\dots,\alpha_{n-1}^*|$ are the estimated parameters in problem (1), (2). It is more robust to normalize the coefficient vector to the unit length, $\|\alpha\| = 1$, as it requires no a priori assumptions of this sort. Then, $\omega \subset \mathbb{S} = \{\vartheta \in E^{n+1} : \|\vartheta\| = 1\}$.

Minimizing the functional (2) for a *specified* coefficient vector α of (1) is called the *smoothing problem.* Minimizing the functional (2) also with respect to α is called the *identification problem.*

We call (1), (2) the *dynamical piecewise-linear approximation problem.* Observe that if the original realization $\mathbf{y}$ is the *precise* readings of some solution to the differential equation

$$\sum_{i=0}^{n} \hat{y}^{(i)} a_i^* = 0$$

on the interval I_t then $\widehat{J} = 0$ and $\hat{\mathbf{y}} = \mathbf{y}$, and so problem (1), (2) amounts to the variational method (in contrast to the analytical method based on the Hamilton–Cayley theorem) for uniform (on I_t) discretization of this differential equation [5, 6].

The space $E \ni \mathbf{y}$ is called the *space of initial data*; the subspace $\Psi = \Psi_\alpha = \ker D_\alpha \subset E$ of dimension n, the kernel of the operator D, is called the *subspace of the model.* Refer to the vector $\mathbf{y}$ as the *original realization*, to the vector $\hat{\mathbf{y}}$ as the *smoothed realization*, and to the vector $\hat{\mathbf{y}}_\alpha$ as the *optimal smoothed realization.* It yields the minimum of (2) for the specified value of α in (1).

In the problem (1), (2), introduce a parameter $M \leq L$ that is related to the length $M + 1$ of the initial segment $\mathbf{y}_M = \{y_i\}_0^M$ of the original realization $\mathbf{y}$, where $M = \overline{L_0, L}$. Here $L_0 + 1 \geq 2n$ is the length of some initial segment for which the solution to (1), (2) can exist and be unique. Henceforth we refer to the corresponding subproblems (1), (2) with parameter M as *M-problems*, denote their solutions by $\hat{\mathbf{y}}_{\hat{\alpha}}(M)$ and $\hat{\alpha}(M)$ and call them (partial) *M-solutions.*

One of the goals of this article is to obtain approximate difference equations for the vector function $\hat{\alpha}(M)$, where $M = \overline{L_0, L}$ (Section 8). See Remark 2 concerning the minimal value $2n$ of $L_0 + 1$.

Lemma 1. *In order to obtain partial M-solutions, for $M \leq L$, in* (1), (2) *it is necessary and sufficient to put $L = M$ and $M = \overline{L_0, L}$ in* (1).

Proof. If $L = M$ in (1) then restrictions (1) on the readings $\hat{y}_i$ for $i = \overline{M+2, L}$ are absent. Consequently, $\hat{y}_i = y_i$ for $i = \overline{M+2, L}$. Thus, $\|\cdot\|_E = \|\cdot\|_{E^{M+1}}$. □

By this property, while stating and solving the M-problems (1), (2), we need not pass from $E = E_L$ to E_M with $M < L$, but can solve the problems in E by putting $L = M$ in (1). In this case we denote by $K = M - n$ the number of conditions (1), where $M = \overline{L_0, L}$, and $K = \overline{N_0, N}$ with $N_0 = L_0 - n \geq n - 1$ and $L_0 \geq 2n - 1$.

In this article we deal with real equations: the coefficients α_i for $i = \overline{0, n}$ and the readings y_l and $\hat{y}_l$ for $l = \overline{0, L}$ in (1), (2) are assumed to be real scalars.

4. An Orthogonal Projection Problem

To express the variational problem (1), (2) as a projection problem in the Euclidean space E, express the condition (1) of minimization of (2) in matrix form. We can do this in two fundamentally (as becomes clear below) different ways. They are essential for solving (1), (2).

Use the special translation matrices that are formed by the coefficients and readings:

$$A = A_N = \left|\begin{matrix} \alpha_0 & 0 & 0 \\ \vdots & \ddots & 0 \\ \alpha_n & \dots & \alpha_0 \\ 0 & \ddots & \vdots \\ 0 & 0 & \alpha_n \end{matrix}\right|, \quad V = V_N = \left|\begin{matrix} y_0, & y_1, & \dots, & y_{n-1}, & y_n \\ y_1, & y_2, & \dots, & y_n, & y_{n+1} \\ y_2, & y_3, & \dots, & y_{n+1}, & y_{n+2} \\ y_3, & y_4, & \dots, & y_{n+2}, & y_{n+3} \\ \vdots & \vdots & \dots & \vdots & \vdots \\ y_N, & y_{N+1}, & \dots, & y_{L-1}, & y_L \end{matrix}\right| \tag{3}$$

This is a band Toeplitz matrix $A = A_N = A(\alpha)$ ($\widehat{A} = A(\hat{\alpha})$) of size $(L+1) \times (N+1)$. Henceforth, call $A = A(\alpha) = A_\alpha = A_N$ the *matrix of the sliding vector* α (or MSV α):

$$A = A_N = A_\alpha = A_N(\alpha) = |\eta_0, \eta_1, \dots, \eta_N| = |\eta_i|_0^N, \tag{4}$$

where $\eta_k = I^k \eta_0 \in E^{L+1}$ and $\eta_0 = |\alpha^*, 0_N^*|^*$. Since the MSV α is formed by an $(n+1)$-vector, the number $N+1$ of its columns is less by n than the number of rows. If the generating vector of the MSV, for instance λ, is of size $N+1$ then the MSV λ is a matrix of size $(L+1) \times (n+1)$, i.e., the number $n+1$ of its columns is N less than the number of its rows. Both $A(\alpha)$ and $\Lambda(\lambda)$, as well as the matrices A_K, occur below. The second matrix is the MSV of the vector of Lagrange multipliers. In Theorem 2 it determines a realization of the “errors” $\Delta\mathbf{y}_\alpha = \mathbf{y} - \hat{\mathbf{y}}_\alpha$. In the M-problems with $M < L$ the relation between the numbers of rows and columns is different. The columns of the MSV constitute special bases for certain subspaces of $E = E^{L+1}$. By Lemma 2, in the M-problems it is unnecessary to pass to the space of realizations of the lower dimension $M+1$.

The second special matrix in (3) is the Hankel matrix $V = V_N = V(\mathbf{y})$ of size $(N+1) \times (n+1)$. The columns of V^* are the vectors

$$v_k = \{y_i\}_k^{k+n} = [y]_{k(n+1)}, \quad k = \overline{0, N},$$

of readings of the original realization $\mathbf{y}$. The columns $\hat{v}_k$ for $k = \overline{0, N}$ of the matrix $\widehat{V}^* = V^*(\hat{\mathbf{y}})$ are the $(n+1)$-samples $[\hat{y}]_{k(n+1)} = \{\hat{y}_i\}_k^{k+n}$ of readings in the smoothed realization $\hat{\mathbf{y}}$. They constitute the model (1).

Introduce the *discrepancy vector* of (1). This is the $(N+1)$-vector of values of the operator $D = D_\alpha$ of (1) on the original realization $\mathbf{y}$:

$$\begin{aligned} m = m_\alpha = D_\alpha \mathbf{y} &= \{\alpha^*[y]_{k(n+1)}\}_0^N = \{\alpha^* v_k\}_0^N \\ &= \{\langle v_k, \alpha\rangle_{E^{n+1}}\}_0^N = \{\langle \mathbf{y}, \eta_k\rangle_E\}_0^N \in E^{N+1}. \end{aligned} \tag{5}$$

Lemma 2. *The image* $m = D_\alpha \mathbf{y}$ *of the real operator* D_α *of* (1) *on the original real realization* $\mathbf{y}$ *satisfies the identities* [3, 5]

$$m = A^* \mathbf{y} = V\alpha \quad \textit{or} \quad m = \langle \mathbf{y}, A\rangle = \langle \alpha, V^*\rangle_{E^{n+1}}. \tag{6}$$

Here and henceforth the vector and matrix arguments in inner products stand for the vectors and matrices of inner products of vectors and column vectors of matrix arguments.

The following theorem and its corollaries are now obvious.

Theorem 1. *Problem* (1), (2) *is a parametric projection problem in* E*:*

$$\textit{minimize} \quad \|\mathbf{y} - \hat{\mathbf{y}}\|^2 \quad \textit{provided that } A^* \hat{\mathbf{y}} = 0 = \widehat{V}\alpha, \tag{7}$$

where the two conditions are equivalent in E.

Corollary 1. *The variational projection problem* (1), (2), *or the constrained minimization of* (2), *is the problem of seeking the nearest element in the set*

$$\Omega = \{\Psi \subset E : \Psi = \Psi_\alpha = \ker D_\alpha,\ \alpha \in \omega\}$$

of admissible subspaces Ψ (the admissible values of the coefficient vector α determined by the set ω). *They are the orthogonal complements $S_\perp \in E$ to the closed linear spans $S = S(A) = S_\alpha$, where $A = A_\alpha$ is the MSV α of the coefficients of the difference equation* (1).

Corollary 2. *If $M - n = K \leq N$ then in the M-problem $A = A_K = |\eta_i|_0^K$ is a matrix of size $(L+1) \times (K+1)$ and V is a matrix of size $(K+1) \times (n+1)$, while* (1) *or* (7) *impose $K + 1 \leq N + 1$ conditions on the realization $\hat{\mathbf{y}}$ of length $L + 1$.*

Two types of condition (1) in problem (1), (2) specified in (7) enable us, firstly, to easily differentiate these conditions with respect to the two types of variables of this problem: the smoothing $\hat{\mathbf{y}}$ and the identification α. Secondly, they enable us to compare (1), (2) to the available simpler problem of estimation and identification in E^{n+1}. We do this in the next section.

We can use the "complete" $(n+1)$-samples $\hat{v}_k = [\hat{y}]_{k(n+1)}$, for $k = \overline{0, N}$ (the columns of $\widehat{V}^*$), as well as "short" n-samples $\hat{\bar{v}}_k$, the n-vectors of states $[\hat{y}]_{k(n)}$, to express (1) as

$$D_\alpha \hat{\mathbf{y}} = \{\alpha^* [\hat{y}]_{k(n+1)}\}_{k=0}^N = \{\alpha^* \hat{v}_k\}_{k=0}^N = 0 \tag{8}$$

or

$$\hat{y}_{k+n} = -\sum_0^{n-1} \hat{y}_{k+i}\alpha_i = -\bar{\alpha}^* \, [\hat{y}]_{k(n)} = -\bar{\alpha}^* \hat{\bar{v}}_k, \quad k = \overline{0, N}, \tag{9}$$

where

$$[\hat{y}]_{k(n+1)} = \{\hat{y}_i\}_k^{k+n} = \hat{v}_k, \quad [\hat{y}]_{k(n)} = \{\hat{y}_i\}_k^{k+n-1} = \hat{\bar{v}}_k. \tag{10}$$

The short n-samples $[\hat{y}]_{k(n)} = \overline{v}_k$, where $k = \overline{0, N+1}$, are the states of the model (1). In view of the recursive form (9) of (1), they enable us to calculate the realization $\hat{\bar{\mathbf{y}}} = \hat{\bar{\mathbf{y}}}_N = \{\hat{y}_i\}_n^L$ successively, starting with the initial conditions $[\hat{y}]_0 = [\hat{y}]_{0(n)}$, provided that $\alpha_n = 1$, i.e., $\alpha \in \mathbb{G} \supset \omega$. The boundary state $[\hat{y}]_{N+1(n)}$ corresponds to $k = N + 1$. Its prediction for the reading $\hat{y}_{L+1}$ using (9) lies outside the observation interval $I_t = [0, Lh]$.

Therefore, in the minimization conditions of (7) we can express the Hankel matrix V of (3) of samples of size $(N+1) \times (n+1)$ as

$$\begin{aligned} V = V_N = \{v_k^*\}_{k=0}^N = \{[y]_{k(n+1)}^*\}_{k=0}^N \\ \longrightarrow V = \{\overline{v}_k^*, y_{k+n}\}_{k=0}^N = \{[y]_{k(n)}^*, y_{k+n}\}_{k=0}^N = |\overline{V}, \overline{\mathbf{y}}|, \end{aligned} \tag{11}$$

where

$$\overline{V} = \overline{V}_N = \{[y]_{k(n)}^*\}_{k=0}^N \in ((N+1) \times n), \quad \overline{\mathbf{y}} = \overline{\mathbf{y}}_N = \{y_i\}_n^L \longrightarrow \hat{\bar{\mathbf{y}}} = -\widehat{\overline{V}}\bar{\alpha}. \tag{12}$$

The matrix $\widehat{\overline{V}}$ of short samples $\hat{\bar{v}}_k^*$ is the matrix $[\hat{y}]_{k(n)}^*$ of states of the model (1) in the observation interval I_t on the mesh I_h (except for the last for $k = N + 1$). These states determine in (9) the components of the realization $\hat{\bar{\mathbf{y}}}$, the last column of the matrix $\widehat{V}$: $\hat{\bar{\mathbf{y}}} = -\widehat{\overline{V}}\bar{\alpha}$ of (12).

5. Alternative Approaches

Let us indicate two alternative approaches to estimating the coefficients of (1). The first, called algebraic identification (AI), involves the minimization of a certain discrepancy functional $m = V(\mathbf{y})\alpha$ of (8). The algebraic approach includes numerous and diverse groups of relatively simple methods of identification of models without feedback (unclosed identification) [3, 6]. The algebraic approach is the simplest and best-known approach to the identification of objects of the form (1) since half a century ago, when the capabilities of computers were minimal. The second approach involves the minimization of the discord of samples (Levin's method [1, p. 294]). There are similar approaches to estimation (when the matrix of the system is regarded as perturbed by errors) called the *orthogonal regression method* (OR) [2, p. 287] and the *total least squares method* (MLS).

Denote by $\|V\|^2 = \mathrm{Sp}(V^*V)$ the squared norm of a matrix V. We can compare the three approaches (VI, OR, and AI) to estimating the coefficients of (1) intuitively and concisely using the simple table:

approach	criterion	conditions
VI	$J_{vi} = \|\mathbf{y} - \hat{\mathbf{y}}\|^2$	$A^*\hat{\mathbf{y}} = 0 = \widehat{V}\alpha$
OR	$J_{or} = \|V - \widehat{V}\|^2$	$\widehat{V}\alpha = 0$
AI	$J_{ai} = \|V(\hat{\mathbf{y}})\alpha\|^2$	$\mathbf{y} - \hat{\mathbf{y}} = 0$

(13)

The definitions (13) immediately imply the following claims.

Lemma 3. **(a)** *In problem* VI *we project one realization* $\mathbf{y} \in E = E^{L+1}$ *of length* $L+1$ *onto a dimension* n *subspace of* E. *In problem* OR *we project* $N+1$ *samples* v_k *of length* $n+1$ *onto the dimension* n *subspace* $\alpha_\perp$ *of* E^{n+1}.

(b) *In problems* OR *and* AI *we regard the rows of the matrix* V, *which are complete* $(n+1)$*-samples of readings contained in* $\mathbf{y}$, *as independent vectors in* E^{n+1}.

(c) *Problem* OR *is the sum of* $N+1$ *independent* $(n+1)$*-problems* VI *for the rows of* V *with one estimated vector* α.

Proof. The first claim is obvious from the first two rows of (13). The second claim follows since we can reorder the rows of V in problems OR and AI in (13) arbitrarily without changing the results of solving these problems. In problem VI the Hankel structure of not only the original matrix V of samples, but also the smoothed matrix $\widehat{V}$ is uniquely determined by the property that the readings of the smoothed realization $\hat{\mathbf{y}}$ are uniquely ordered by the optimization criterion (the distance in E) and the Toeplitz matrix A^* in the leftmost identity in (13) VI (Lemma 2). The third claim follows since in problem VI one $(L+1)$-realization in $E = E^{L+1}$ projects to the subspace $S_\perp(A) \subset E$ of dimension n, while in problem OR a "cloud" [2, p. 20] of $N+1$ vectors in E^{n+1} (the columns of V^*) independently project to the subspace $S_\perp(\alpha) = \alpha_\perp \subset E^{n+1}$ (of dimension n as well). □

Corollary 1. *In Problem* VI, *to determine the transitional process* (*smoothed realization*) $\hat{\mathbf{y}}$ *of length* $L+1$, *it is necessary and sufficient to optimize* n *variables: the* n*-vector of its initial conditions* (*or arbitrary other local conditions if* $\alpha_0 \neq 0$). *In problem* OP *of projecting* $N+1$ *samples* E^{n+1} *onto* $\alpha_\perp$, *the number of optimized initial conditions equals* $(N+1)n$.

Proof. The number of optimized variables for smoothing VI and OR is easy to calculate using (9) and the form of V. □

Definition 1. A model $\mathbb{M}(\alpha, \hat{\mathbf{y}})$ is called *closed* (with feedback) whenever its representation of the realization $\hat{\mathbf{y}}$ in the space E of initial data $\mathbf{y}$ is the solution to an initial value problem.

Corollary 2. (a) *In problem* AI (13) *the original realization of* $\mathbf{y}$ *is regarded as an "output" of the model. Smoothing is absent. The optimized parameters are* $\bar{\alpha}$.

(b) *The output errors as discrepancy* $m = V\alpha$ *of* (5) *are transferred to the input of the model.*

(c) *the model* $\mathbb{M}(\alpha, \hat{\mathbf{y}})$ of AN (14) *lacks feedback.*

Proof. As (13) implies, the AI methods use as a model for reading the realization $\overline{\mathbf{y}}_N$ (see (11) and (12)) the algebraic relation called the perturbed sliding average:

$$y_{k+n} = -\sum_{0}^{n-1} y_{k+i}\alpha_i + m_{k+1}, \quad k = \overline{0, N}. \tag{14}$$

Minimize with respect to $\bar{\alpha}$ the "input" of (14); i.e., $\|m\|^2_{E^{N+1}}$. The errors in the original realization of $\mathbf{y}$ are absent. Thus, no smoothing problem appears here, and so the calculation of $\hat{\mathbf{y}}$ proceeds.

(b) The relation (14) looks similar to the model (1) in the form (9), but has different physical and informational meanings. The errors in the original realization $\mathbf{y}$ in this model carry over to its "input" m, a suitable functional of which is to be minimized. This follows from (13) and (14).

(c) From the statement of problem AI in (13) together with (14), it is clear that the original realization $\mathbf{y}$, used as the output sequence of (14) is not a solution to the initial value problem. The inverse relation, used in (9) to solve the initial value problem for (8), in (14) is "destroyed" by an arbitrary sequence, the discrepancy m. It is used as the input of the model (14). □

To transfer errors from output to input is incorrect for a physically realizable operator. This leads to the low stability of the AI methods with respect to errors in the original data. The relative simplicity of the use and analysis of AI methods is their characteristic feature. For instance, using (11) and (12), we obtain from (14) the inconsistent system $V\alpha \approx 0$ and the corresponding MLS estimates that minimize the discrepancy $m = V(\mathbf{y})\alpha$ of (6) and (13):

$$V\alpha = m \approx 0 \longrightarrow \overline{\mathbf{y}} \approx -\overline{V}\bar{\alpha} \longrightarrow \bar{\alpha} = -(\overline{V}^*\overline{V})^{-1}\overline{V}^*\overline{\mathbf{y}}. \tag{15}$$

Remark 1. These estimates are known to be biased due to errors in $\overline{V}$ [1, p. 283]. Thus, processing the original realization $\mathbf{y}$ for decrease in the errors in it usually precedes the methods of AI. The best processing method is to apply a model most adequate for the process under study. In our case this is model (1). Precisely joint smoothings $\mathbf{y}$ and estimation α are performed in problem VI. This is also done in problem OR, but for the row $\{v_i^*\}_0^N$ of V.

Remark 2. Definitions (11) and (12) together with (15) imply that $L_0+1 \geq 2n$ in Lemma 1. Here $2n$ is the minimal number of readings necessary for calculating certain estimates of the coefficient vector $\bar{\alpha}$ using (15). These readings must be such that the square matrix $\overline{V}_{n-1}$ of size $n \times n$ is nondegenerate (see (12)). For this minimal number of readings of this type, three estimates in (13) coincide. We calculate them using the formula $\bar{\alpha} = -\overline{V}_{n-1}^{-1}[y]_{n(n)}$.

The estimates $\hat{\alpha}(M)$ of all partial M-problems (1), (2) VI, OR, and AI coincide not only for $L = 2n$ (and a nondegenerate matrix $\overline{V}_{n-1}$), but in one more case.

Lemma 4. *Assume that some realization* $\mathbf{y}$ *is an exact solution to the difference equation* (1) *with* $L = M$ *and some coefficients* $\alpha = |\alpha_0, \dots, \alpha_{n-1}, 1|$*. In addition, assume that for this solution the initial matrix of samples* $\overline{V}_{n-1}$ *is nondegenerate. Then the solutions to identification problems* VI, OR, AI (13) *for* $M = \overline{L_0, L}$ *coincide and can be obtained using* (15).

Proof. These assumptions yield $V\alpha = \widehat{V}\alpha = 0$ because $\hat{\mathbf{y}} = \mathbf{y}$. It is clear from the table in (13) that the conditions for minimization of both VI and OR are fulfilled, all functionals J_{vi}, J_{or}, and J_{ai} are minimal and vanish on the solution α.

In the hypotheses of the lemma, for all $M = \overline{L_0, L}$ the matrices of samples V_K for $K = M - n$ are of rank n. Thus, the solution (15) is unique for all problems in (13). □

6. An Analytical Solution of Problem VI

Two n-vectors of independent variables, for instance $[\hat{y}]_0$ and α, of problem (1), (2) determine two stages of optimization. The first stage, smoothing, requires no search, as we calculate the smoothed realization $\hat{\mathbf{y}}_\alpha$ for a specified coefficient vector α using orthogonal projection formulas. The second stage, identification, involves search. In the framework of analytical solution, in this section we obtain an expression for the identification functional, whose unconstrained global minimum yields the solution to the identification problem in the approximation problem (1), (2).

The formulas of orthogonal projection to a subspace when a basis is chosen for it or its orthogonal complement are well-known. We can easily deduce them from the statement of problem (1), (2) as (7). Denote by $C = C_N = \langle A, A\rangle$ the Gram matrix of the system of vectors $A = A_N$.

Theorem 2. **(a)** *The method of Lagrange multipliers* (LM) *yields orthogonal projection formulas in problem* (1), (2)*; the vector of* LM *equals* $\lambda = C^{-1}\langle \mathbf{y}, A\rangle$.

(b) *Denote by* Π *the projection onto* $\ker D_\alpha = \Psi_\alpha = S_\perp = S(A_\perp)$*, and by* $P = P(A) = I - \Pi$ *the projection onto* $S = S(A) = E \ominus \Psi_\alpha$*, where* $N = L - n$*. Then we can express the smoothed realization* $\hat{\mathbf{y}}_\alpha = \hat{\mathbf{y}}_\alpha(L)$ *and the perpendicular* $\Delta\mathbf{y}_\alpha = \mathbf{y} - \hat{\mathbf{y}}_\alpha$ *as*

$$\hat{\mathbf{y}}_\alpha = \Pi\mathbf{y}, \quad \Pi = \Pi_\alpha = \Pi_N = I - P, \quad P = P_\alpha = P_N, \quad \Delta\mathbf{y}_\alpha = \mathbf{y} - \hat{\mathbf{y}}_\alpha = A\lambda = P\mathbf{y}, \tag{16}$$

where

$$P = P(A) = A\langle A, A\rangle^{-1}\langle \cdot, A\rangle = AC^{-1}A^*, \quad C = \langle A, A\rangle = A^*A. \tag{17}$$

(c) *the partial* M*-solutions* $\hat{\mathbf{y}}_\alpha(M)$ *and* $\Delta\mathbf{y}_\alpha(M)$ *to the* M*-problem* (1), (2) *in* (16) *and* (17) *satisfy* $A = A_K$, $\Pi = \Pi_K$, *and* $P = P_K$, *where* $K = M - n = \overline{L_0 - n, N}$.

The next theorem provides important formulas for the squared length of the perpendicular $\Delta\mathbf{y}_\alpha = \mathbf{y} - \hat{\mathbf{y}}_\alpha$.

Theorem 3. *The value* $\widehat{J}$ *of* J *of* (2) *at the projections* $\hat{\mathbf{y}}_\alpha$ *of* (16) *of the original realization* $\mathbf{y}$ *of length* $L + 1$ *onto the subspace*

$$\Psi_\alpha = \ker D_\alpha = S_\perp \subset E^{L+1},$$

the kernel of the difference operator D_α *of the model* (1)*, is determined by*

$$\widehat{J} = \widehat{J}_\alpha = \rho^2(\alpha) = \rho_N^2 = \|\Delta\mathbf{y}_\alpha\|^2 = \langle \mathbf{y}, \Delta\mathbf{y}_\alpha\rangle = \langle \mathbf{y}, P(A)\mathbf{y}\rangle = m^*C^{-1}m. \tag{18}$$

Corollary 1. *The length* $\rho = \|\Delta\mathbf{y}_\alpha\| = \rho(\alpha)$ *of the perpendicular* $\Delta\mathbf{y}_\alpha = \mathbf{y} - \hat{\mathbf{y}}_\alpha$ *is independent of the smoothing parameters of problem* (1), (2) *and depends only on the coefficient vector* α.

Refer to the function $\rho^2(\alpha) = \widehat{J}(\alpha)$ of α in (18) as the *identification functional.*

Corollary 2. (a) *The optimization problem of the model* (14) *with respect to its coefficient vector* α *and the equivalent problem* AI *in* (13) *become equivalent to problem* VI (1), (2) *in* (13) *when the norms in the space* E^{N+1} *of discrepancies* $V\alpha = m \in E^{N+1}$ *are defined as* $\| * \|_{C^{-1}}$, *where* $C = C(\alpha)$ *is the Toeplitz Gram matrix of the system of vectors* $A(\alpha)$ *of* (3) *in the projection formulas* (17) *and* (18) *of Theorem* 2.

(b) *Then the criteria for minimization of the discrepancy* $m \in E^{N+1}$ *of optimization problem* AI *in* (13) *of the coefficients* α *of* (14) *are the identification functionals in* E^{N+1}: $\|V\alpha\|^2_{C^{-1}} = \|m\|^2_{C^{-1}}$.

Proof. Theorem 3 reduces the variational identification problem (13) to the problem of unconstrained minimization of the identification functional (18). The form of this functional implies the corollary. □

Theorem 4. *The optimization problem of the unknown coefficients of the difference equation in the variational approximation problem* (1), (2) *reduces to the unconstrained minimization problem for the identification functional*

$$\rho^2(\alpha) = J(\hat{\mathbf{y}}_\alpha) = \widehat{J} = \langle A, \mathbf{y}\rangle \langle A, A\rangle^{-1} \langle \mathbf{y}, A\rangle = \mathbf{y}^* A(A^*A)^{-1}A^*\mathbf{y} \tag{19}$$

in $E = E^{L+1}$.

The equalities of Theorem 4 and the identities of Lemma 2 lead to the following result.

Theorem 5. (a) *We can also express the identification functional, defined in Corollary* 2 *of Theorem* 3 *as a functional on* E^{N+1}, *and in Theorem* 4 *as a functional on* $E = E^{L+1}$, *as a functional on the sphere in* E^{n+1}.

(b) *We can express it as a quadratic form with identifying matrix* $Q(\alpha)$ *which is nonconstant with respect to* α *on the sphere* $\mathbb{S}(c) = \{\alpha \in E^{n+1} : \|\alpha\| = c\}$:

$$\widehat{J} = \widehat{J}_{vi} = \widehat{J}_\alpha = \rho^2_\alpha = \alpha^* Q \alpha = \widehat{J}_\alpha(L), \tag{20}$$

where

$$Q = Q_{vi} = Q(\alpha) = V^*(A^*A)^{-1}V = V^*C^{-1}V.$$

Refer to a vector $\hat{\alpha} = \hat{\alpha}(M) = \arg\min \widehat{J}_\alpha(M) = \widehat{J}_{vi}$ on which the identification functional attains its global minimum as a *solution to the partial variational identification* M*-problem* (1), (2) for the realization of length $M + 1$, where $M = \overline{L_0, L}$.

Remark 3. In problem OR of (13) we have the following analogy with (20) [3, 5, 7]:

$$\widehat{J}_{or} = \alpha^* Q_{or} \alpha, \quad Q_{or} = V^*V/\|\alpha\|^2.$$

7. Equations of Smoothing and Filtering

As we realize the solutions obtained, the following questions arise: to find the minimum of the identification functional $\widehat{J}_\alpha(M)$ and, if the length of the realization L (the number $N = L - n$) is large, to invert the Gram matrix C. Suitable recursive algorithms (with respect to M) rely on the equations of the two-sided (*counter*) Gram–Schmidt orthogonalization [5, 7–9]. Below we give the equations for successively solving the M-problems (1), (2) for $M = \overline{L_0, L}$.

Introduce the projections $P_{\overline{k,l}}$ onto the subspaces $S_{\overline{k,l}} = S(A_{\overline{k,l}})$, the closed linear spans of the listed columns of the matrix A, namely, $A_{\overline{k,l}} = |\eta_k, \dots, \eta_l|$. Put $P_{\overline{0,k}} = P_k$, $A_{\overline{0,k}} = A_k$, and $S_{\overline{0,k}} = S_k$. Use similar indices at the projections $\Pi_{\overline{k,l}} = I - P_{\overline{k,l}}$ and $\Pi_{\overline{0,k}} = \Pi_k$ onto the orthogonal complement in $\Psi_{\overline{k,l}}$ to the subspaces $S_{\overline{k,l}}$. Clearly, $\Pi_{-1} = I$ and $P_{-1} = 0$.

Define the *orthogonalizing vectors* f_k and $\tilde{f}_k$, for $k = \overline{0, N}$, of the opposite (*forward* and *backward*) processes in the Gram–Schmidt orthogonalization of the system of vectors A_k. In M-problems the vectors f_k, where $k = \overline{0, K}$ and $K = M - n$, determine the chain of projections Π_K, for $K = \overline{0, N}$, of M-problems (1), (2). Here $K + n = M = \overline{L_0, L}$.

By the definition of the processes of successive orthogonalization [7],

$$f_k = \Pi_{k-1}\eta_k \longrightarrow \Pi_k = I - \sum_0^k f_i \langle f_i, f_i \rangle^{-1} \langle \cdot, f_i \rangle, \quad k = \overline{0, N}, \ \tilde{f}_k = \Pi_{\overline{1,k}}\eta_0. \qquad (21)$$

Theorem 6. (a) *Put* $a_k = \|f_k\|^{-2}$, *and* $\tilde{a}_k = \|\tilde{f}_k\|^{-2}$, *while* $\tilde{f}_0 = f_0 = \eta_0$.

To calculate the orthogonalizing vectors f_{k+1} *and* $\tilde{f}_{k+1}$ *for* $k = \overline{-1, N-1}$, *we can use the following nonlinear equations of the two-sided (counter) orthogonalization process:*

$$f_{k+1} = I^1 f_k - \tilde{f}_k \theta^*_{k+1}, \quad \tilde{f}_{k+1} = \tilde{f}_k - I^1 f_k \theta_{k+1}, \qquad (22)$$

where

$$\theta_{k+1} = a_k \mu_{k+1}, \quad \mu_{k+1} = \langle \tilde{f}_k, I^1 f_k \rangle,$$
$$a_{k+1} = \tilde{a}_{k+1} = (I - \theta_{k+1}\theta^*_{k+1})^{-1} a_k, \quad a_0 = \|\eta_0\|^{-1} = \|\alpha\|^{-2}.$$

(b) *The numbers* $1 - |\theta_{k+1}|^2 > 0$ *to be inverted are nonzero for all* $k = \overline{0, N-1}$ *provided that the coefficient vector* α *of the difference equation* (1) *has at least one nonzero component, i.e.,* $\operatorname{rank} A_N = N + 1$.

Corollary 1. *The projection* $\hat{\mathbf{y}}_{k+1} = \Pi_{k+1}\mathbf{y}$ *is a solution to the* M*-problem of smoothing* (1), (2) *for* $M = k + n + 1 = K + n$ *and* $K = M - n = k + 1$, *where* $k = \overline{-1, N-1}$.

Corollary 2. *The following recurrences describe the partial* M*-smoothed* ($M = k + n + 1$) *realization* $\hat{\mathbf{y}}_{k+1} = \Pi_{k+1}\mathbf{y} = \hat{\mathbf{y}}(M)$:

$$\hat{\mathbf{y}}_{k+1} = \hat{\mathbf{y}}_k - f_{k+1} a_{k+1} \pi_{k+1}, \quad \textit{where} \quad \pi_{k+1} = \langle \hat{\mathbf{y}}_k, \eta_{k+1} \rangle = y_{k+1+n} - \hat{y}_{k+1+n/k}, \quad (23)$$

i.e., the renewal process: the error of the prediction $\hat{y}_{k+1+n/k}$ *of* (9) *for the value of the reading* y_{k+1+n}.

Proof. The first corollary follows from Lemma 1 and Corollary 2 to Theorem 1. The second corollary follows from the structure of the basis A of (3), its constituent vectors η_k of (4), as well as (21) and (22). □

Lemma 5. *Assume that* $M = k + n + 1 = \overline{n, L}$ *and* $k = \overline{-1, N-1}$. *To calculate in* (23) *the predictions* $\hat{y}_{k+1+n/k}$ *for* $k = \overline{0, N-1}$, *it suffices to have a solution to the filtration* $(M-1)$*-problem, i.e., to have an* $(M-1)$*-estimate of the previous state of the model:* $[\hat{y}]_{k+1(n)/k} = \{\hat{y}_{i/k}\}_{i=k+1}^{k+n}$ (10).

Corollary 1. *To continue recursion in* (23)*, it suffices to calculate only the n-vector of the last readings $\{\hat{y}_{i/k+1}\}_{k+2}^{k+n+1}$ of the M-smoothed realization $\hat{\mathbf{y}}_{k+1}$.*

Corollary 2. *To predict and continue recursion, it suffices to calculate in* (22) *only the last n components $[f]_k = \{f_{ik}\}_{k+1}^{k+n}$ and $[\tilde{f}]_k = \{\tilde{f}_{ik}\}_{k+1}^{k+n}$ of direct and inverse vectors f_k and $\tilde{f}_k$, for $k = \overline{0,N}$, of the two-sided orthogonalization process.*

Theorem 7. **(a)** *The calculation in* (22) *of the last n components $[f]_k = \{f_{ik}\}_{k+1}^{k+n}$ and $[\tilde{f}]_k = \{\tilde{f}_{ik}\}_{k+1}^{k+n}$ of the orthogonalizing vectors f_k and $\tilde{f}_k$ is necessary and sufficient for solving M-problems* (1), (2) *of filtration and prediction for all $M = K + n = \overline{n, L}$ and $K = k = \overline{0,N}$.*

(b) *Suppose that $\alpha_0 \neq 0$. Then the calculation of the last n components of f_k and $\tilde{f}_k$ in* (22) *is also necessary and sufficient for solving the smoothing problems, the complete one and all partial ones, namely, for calculating the smoothed realizations $\hat{\mathbf{y}}_K = \hat{\mathbf{y}}(M)$ of all $M = K + n$-problems* (1), (2) *for $M = \overline{n, L}$. If $M = L$ then we obtain a solution to the complete smoothing problem* (1), (2).

Proof. Claim (a) is obvious in view of Lemma 5 and its corollaries. To justify claim (b), observe that if $\alpha_0 \neq 0$ then we can solve (9) for the lowest reading $\hat{y}_k$. Therefore, it becomes possible to solve this difference equation backwards in time:

$$\hat{y}_{k/k} = -\sum_{1}^{n} \hat{y}_{k+i/k}\alpha_i/\alpha_0,$$

for $k = \overline{N, 0}$. Consequently, we can calculate all $(K+n)$-smoothed realizations $\hat{\mathbf{y}}_k = \hat{\mathbf{y}}_K = \hat{\mathbf{y}}(M)$, for $M = \overline{n, L}$, with the boundary conditions $\{\hat{y}_{i/k}\}_{k+1}^{k+n}$. □

Claim (a) of Theorem 7 also holds for the equations of real-time variational identification thanks to Theorems 5 and 6. We obtain these equations below.

8. Equations for the Identifying Matrix

Lemma 6. **(a)** *Suppose that the matrix $|\mathbf{w}_j|_0^n$ of size $(L+1) \times (n+1)$ is the collection of $n+1$ realizations $\mathbf{w}_j$, for $j = \overline{0,n}$, such that their discrepancies $A^*|\mathbf{w}_j|_0^n$ amount to the matrix of samples V, i.e., $\langle |\mathbf{w}_j|_0^n, A\rangle = V$.*

(b) *Denote by W the matrix $|\mathbf{w}_j|_0^n$ with the minimal squared norm $\|W\|^2 = \mathrm{Sp}(W^*W) = \mathrm{Sp}\langle W, W\rangle$.*

Then the identifying matrix Q of (20) *of problem* VI (1), (2) *is $Q = Q_{vi} = \langle W, W\rangle$, which is the Gram matrix of the minimal matrix of realizations W.*

Proof. The claim follows because the minimal solution to the system $A^*W = V$ is $W = W_\alpha = A^{(-1)*}V = AC^{-1}V$ [10, p. 37]. □

Lemma 7. *Take the last column $c_k = \langle \eta_k, A_k\rangle = |\overline{c}_k^{\,*}, c_{kk}|^*$ of the selfadjoint $(k+1)$-matrix $C_k = \langle A_k, A_k\rangle$. Here $\overline{c}_k = \langle \eta_k, A_{k-1}\rangle$ and $c_{kk} = \langle \eta_k, \eta_k\rangle = \|\alpha\|^2$. Take the last column $F_k b_k = |\overline{F}_k^{\,*}, 1|^* b_k$ of the inverse matrix C_k^{-1}, where b_k is the diagonal lower right entry, $C_{k-1/0}^{-1}$ is the $(k+1)$-matrix C_{k-1}^{-1} bordered by zeroes in the last row and column, while $\Delta_k = C_k^{-1} - C_{k-1/0}^{-1}$. Then for $k = \overline{0,N}$ we have*

$$F_k = \Delta_k c_k, \tag{24.1}$$

$$F_k^* c_k = b_k^{-1}, \tag{24.2}$$

$$A_k F_k = f_k, \tag{24.3}$$

$$b_k = a_k = \|f_k\|^{-2}. \tag{24.4}$$

Proof. The Frobenius formula for inverting bordered matrices [10, p. 60] implies (24.1) and (24.2). Express the Frobenius formula for C_k as

$$C_k^{-1} = C_{k-1/0}^{-1} + F_k(F_k^* c_k)^{-1} F_k^* \longrightarrow \Delta_k = \Delta_k c_k (c^* \Delta_k c_k)^{-1} c_k^* \Delta_k.$$

Here $F = |\overline{F}^*, 1|^* = \Delta_k c_k$ and $\overline{F}_k = -C_{k-1/0}\overline{c}_k$. The chain of equalities

$$\begin{aligned} A_k F_k = A_k \Delta_k c_k = A_k\big(C_k^{-1} - C_{k-1/0}^{-1}\big)\langle \eta_k, A_k\rangle = P_k \eta_k - P_{k-1}\eta_k \\ = \eta_k - P_{k-1}\eta_k = \Pi_{k-1}\eta_k = f_k \longrightarrow a_k^{-1} = \langle A_k F_k, A_k F_k\rangle \\ = F_k^* C_k F_k = b_k^{-1} |0_k^*, 1| F_k = b_k^{-1} \quad \square \end{aligned} \tag{25}$$

yields (24.3) and (24.4).

Denote by $W_k = A_k C_k^{-1} V_k$, for $k = \overline{0, N}$, the minimal solutions to the partial M-systems of equations $A_k^* W_k = V_k$ for $M = k + n$. Here $C_k = \langle A_k, A_k\rangle$, while $A_k = |\eta_i|_0^k$ and $\eta_i \in E$ for $i = \overline{0, N}$.

Theorem 8. *The minimal solutions W_{k+1} to the equations $A_{k+1}^* W_{k+1} = V_{k+1}$ as functions of the parameter $k = \overline{-1, N-1}$ can be described by the difference equations:*

$$\begin{aligned} W_{k+1} = W_k + f_{k+1} a_{k+1} (v_{k+1}^* - \hat{v}_{k+1/k}^*), \\ \hat{v}_{k+1/k}^{\,*} = -\overline{F}_{k+1}^{\,*} V_k = \langle W_k, \eta_{k+1}\rangle, \quad W_{-1} = 0. \end{aligned} \tag{26}$$

Here $\hat{v}_{k+1/k}^*$ is the prediction of the system $A_{k+1}^* W_k = V_{k+1/k}$ for the row v_{k+1}^* of the matrix V_{k+1}, $v_{k+1}^* - \hat{v}_{k+1/k}^*$ is the error of this prediction, $f_{k+1} = \Pi_k \eta_{k+1}$ is the final vector of the direct Gram–Schmidt orthogonalization for the subsystem A_{k+1}, and $a_{k+1} = \|f_{k+1}\|^{-2}$.

Proof. For $k = \overline{-1, N-1}$ and $W_{-1} = 0$ The Frobenius formula and Lemma 6 yield

$$W_{k+1} = A_{k+1}\big(C_{k/0}^{-1} + F_{k+1} a_{k+1} F_{k+1}^*\big) V_{k+1} = W_k + f_{k+1} a_{k+1} c_{k+1}^* \Delta_{k+1} V_{k+1}.$$

Since $\Delta_{k+1} = C_{k+1}^{-1} - C_{k/0}^{-1}$, the claim follows:

$$c_{k+1}^* C_{k/0}^{-1} V_{k+1} = \langle A_{k+1}, \eta_{k+1}\rangle C_{k/0}^{-1} V_{k+1} = \langle W_k, \eta_{k+1}\rangle,$$

$$\begin{aligned} c_{k+1}^* C_{k+1}^{-1} V_{k+1} = \big|0_{k+1}^T, 1\big| V_{k+1} = v_{k+1}^* \\ \longrightarrow F_{k+1}^* V_{k+1} = c_{k+1}^* \Delta_{k+1} V_{k+1} = v_{k+1}^* - \hat{v}_{k+1/k}^{\,*}, \end{aligned}$$

where

$$\hat{v}_{k+1/k}^* = -\overline{F}_{k+1}^{\,*} V_k = \langle W_{k+1}, \eta_{k+1}\rangle. \quad \square$$

Corollary 1. *The identifying Gram matrix $Q = (W, W)$ satisfies the following recurrences for $k = \overline{-1, N-1}$ and the zero initial conditions $Q_{-1} = 0$:*

$$Q_{k+1} = Q_k + q_{k+1} a_{k+1} q_{k+1}^*, \quad \text{where} \quad q_{k+1} = v_{k+1} - \hat{v}_{k+1/k}. \tag{27}$$

Proof. Two terms in (26) are orthogonal to each other; hence,

$$Q_{k+1} = \langle W_k + f_{k+1} a_{k+1} q_{k+1}^*, W_k + f_{k+1} a_{k+1} q_{k+1}^*\rangle = \langle W_k, W_k\rangle + q_{k+1} a_k q_{k+1}^*. \quad \square$$

Corollary 2. *If the matrix A is the MSV* (3), (4) *then we can calculate the vectors $\hat{v}^*_{k+1/k}$ of predictions $\langle W_k, \eta_{k+1}\rangle$ in* (26) *and* (27) *in the filtration regime described in Theorem* 7, *Lemma* 5, *and its Corollary* 2. *This means that the vector of predictions $\hat{v}^*_{k+1/k}$ is determined by only the last n rows of the matrix of realizations W_k. Consequently, in order to calculate them, it is necessary and sufficient to know only the last n components of the orthogonalizing vectors f_k and $\tilde{f}_k$.*

These properties of the identification problem (announced following the proof of Theorem 7) complement the statements of Theorem 7, Lemma 5, and its Corollary 2. The analogous properties in these statements apply to the problems of filtration and prediction in accordance with the general smoothing equations (23).

Proof. Recall that the filtration regime means that in (22) of Theorem 6 we calculate only the last n components of the orthogonalizing vectors f_k and $\tilde{f}_k$ mentioned in Theorem 7. This suffices for both solving the filtration problem and calculating predictions in accordance with the general smoothing equations (23). These components also suffice for calculating the predictions $\hat{v}^*_{k+1/k} = \langle W_k, \eta_{k+1}\rangle$ indicated in Corollary 2.

Indeed, the last formula, from the second equation in (26), and the structure of the vector η_{k+1} defined in (3) and (4) show that to calculate the vector of predictions $\hat{v}^*_{k+1/k}$, it suffices to know only the last n rows of the matrix of realizations W_k. As the first equation in (26) implies, to calculate these rows, it suffices to know only the last n components of the orthogonalizing vectors f_k and $\tilde{f}_k$. These components, which underlie the filtration regime, are defined in Theorem 7. According to this theorem, as well as Lemma 5 and its Corollary 2, the knowledge of these components is necessary and sufficient for solving the filtration problem in accordance with the general smoothing equation (23). According to the claim being proved, these components are also necessary and sufficient for solving the identification problem. □

Equations (22), (26), and (27) constitute a system of nonlinear difference equations for the identifying matrix $Q_k = Q(M)$ with $M = k + n = \overline{L_0, L}$ (see Lemma 1) of partial M-problems (1), (2).

9. Identification Equations in Real Time

This is the term for equations for the estimates $\hat{\alpha}(M)$ as functions of the exponents of the length of realization, the numbers M or $K = M - n$. Real-time identification is used, in particular, in certain problems of automatic control with identificator (adaptive and self-adjusting control systems).

In problems AI of (13) it is possible to obtain exact equations for the estimates (15). They rely on the matrix Riccati equations for inverting matrices with additive factorized increments [11, p. 20]. We can express the system of equations for calculating the matrix function Q^{-1}_{k+1}, where $k = \overline{L_0 - n, N - 1}$ (and $L_0 \geq 2n$), as

$$Q^{-1}_{k+1} = (Q_k + q_{k+1}a_{k+1}q^*_{k+1})^{-1} = Q^{-1}_k - Q^{-1}_k q_{k+1}(a^{-1} + q^*_{k+1}Q^{-1}_k q_{k+1})^{-1} q^*_{k+1} Q^{-1}_k \tag{28}$$

provided that the matrix Q_{L_0} of (27) of the L_0-problem (1), (2) is invertible (see Remark 2). Using these general equations in the formula (15) for inverting the matrices

$$V^*_{k+1}V_{k+1} = V^*_k V_k + \overline{v}_{k+1}\overline{v}^*_{k+1},$$

we obtain the equations of the recursive least squares method [11, p. 279].

Exact equations for the estimates in problems VI and OR are beyond reach. Below we obtain equations for one form of approximate estimates for the coefficients α in problems VI and OR.

The problems VI and OR have common features not only in their statements (13) (Remark 3). For the M-problems (1), (2) introduce the $K+1$-vectors λ_{vi} and λ_{or} of Lagrange multipliers, as well as the functionals

$$J_{viM} = J_{vi} + \lambda_{vi}^* A^* \hat{\mathbf{y}} = J_{vi} + \lambda_{vi}^* A^* \widehat{V}\alpha, \quad J_{orL} = J_{or} + \lambda_{or}^* \widehat{V}\alpha,$$

while for problem VI, moreover, the size $(M+1)\times(n+1)$ matrix Λ_{vi} of the form (4) of the sliding LM vector λ_{vi}. Observe that the identificator functionals of problems VI (19) and OR (Theorem 9) are invariant with respect to the length of coefficient vector α; therefore, their gradients are orthogonal to this vector.

Using [3, 5, 12], we can obtain the following formulas for the identification functionals $\widehat{J}_{viM}$ and $\widehat{J}_{orM}$, as well as their derivatives with respect to α.

Theorem 9. **(a)** *The* LM *and identification functionals of the M-problems* VI *and* OR *are of the form*

$$\lambda_{vi} = C^{-1}V\alpha, \quad \lambda_{or} = (1/\|\alpha\|^2)V\alpha,$$

$$\widehat{J}_{vi} = \alpha^*\Lambda^*\Lambda\alpha = \alpha^* Q_{vi}\alpha, \quad \widehat{J}_{or} = \alpha^*\alpha\lambda_{or}^*\lambda_{or} = \alpha^* Q_{or}\alpha.$$

(b) *The gradients of $\widehat{J}'_{or}$ and $\widehat{J}'_{vi}$, which are the vectors of derivatives of the functionals $\widehat{J}_{or}$ and $\widehat{J}_{vi}$ with respect to the coefficients α can be expressed in the comparable forms:*

$$\begin{gathered} \widehat{J}'_{vi} = (Q_{vi} - \Lambda^*\Lambda)\alpha, \quad Q_{vi} = V^*C^{-1}V, \quad Q_{or} = V^*V/\|\alpha\|^2, \\ \widehat{J}'_{or} = (Q_{or} - I_{n+1}\cdot\lambda_{or}^*\lambda_{or})\alpha = (Q_{or} - I_{n+1}\cdot\widehat{J}_{or}/\|\alpha\|^2)\alpha. \end{gathered} \tag{29}$$

To minimize $\widehat{J}_{vi}$ and $\widehat{J}_{or}$, use the iterative gradient procedures (IP): $\alpha_{[j+1]} = \alpha_{[j]} - T_{[j]}^{-1}J'_{[j]}$. Here T is a positive definite matrix. Inserting into T the matrices $Q_{vi} = V^*C^{-1}V$ or $Q_{or} = V^*V/\|\alpha\|^2$, we obtain these IP for problems VI and OR, which converge to the minimum of the functionals $\widehat{J}_{vi}$ and $\widehat{J}_{or}$ under certain conditions:

$$(a)\ \alpha_{[j+1]vi} = Q_{[j]vi}^{-1}\Lambda_{[j]}^*\Lambda_{[j]}\alpha_{[j]vi}, \quad (b)\ \alpha_{[j+1]or} = Q_{[j]or}^{-1}\lambda_{[j]or}^*\lambda_{[j]or}\alpha_{[j]or}. \tag{30}$$

For OR in (30(b)) we obtain IP applied to minimize the functionals of this kind: seeking the eigenvector corresponding to the minimal eigenvalue, with the factor $\lambda^*\lambda$ restricting the growth of the length of α in these iterations.

Experiments showed that the iterations of the form (30(a)) and (30(b)) have a high rate and a large domain of convergence. For the single-valued convergence of iterations in problem OR (30(b)) the fulfillment of the two conditions is necessary and sufficient: the simplicity of the minimal eigenvalue and the nonzero projection of the initial vector $\alpha_{[0]}$ onto the corresponding eigenvector.

Theoretical analysis of the functional (20) and the iteration (30(a)) of problem VI is rather complicated. There are no methods for this analysis, but only encouraging experimental results. They show that the iteration (30a) of problem VI have even greater rate and larger domain of convergence (for the same relative errors in the original realization $\mathbf{y}$, which determines the sharpness of the extremum and the rate of convergence) than the iteration in problem OR. We can explain this by the smaller number of independent variables in the approximation problem VI in comparison with problem OR of (13) (Corollary 1 of Lemma 3).

Suppose that the level of errors is low. Then it is possible to obtain an approximation $\tilde{\alpha}(M)$ to the exact solution $\hat{\alpha}(M)$ with practically sufficient accuracy in one iteration of the form $(30(a))$:

$$\tilde{\alpha}(M) = \vec{\alpha}_{[1]}(k+n), \quad \vec{\alpha} = \alpha/\|\alpha\|,$$

where $\alpha_{[1]}(k+n) = Q_k^{-1}(\vec{\alpha}_{[0]}) \cdot \vec{\alpha}_{[0]}$, where $k = \overline{K_0, N}$ with $K_0 > n$. Thanks to the expansion (27) of Q, we calculate Q_k^{-1} using (28).

Suppose that errors in the original realization $\mathbf{y}$ are absent or their level is very low, i.e., the smallest eigenvalue of $\widehat{J}_{\min} = \min_\alpha \widehat{J}$ is zero or very small. Then the rate of convergence IP (30) is very high, but the matrices Q are degenerate or ill-conditioned. In these cases, we have to invert the matrices $Q + I\varepsilon$, where $\varepsilon > 0$. This involves translating the eigenvalues of Q by ε, or otherwise, adding the regularizing term $\varepsilon\|\alpha\|^2$ to the approximation functional J of (2) and identification functional $\widehat{J}$ of (20).

REFERENCES

1. *Eickhoff F.* Principles of Identification of Control Systems [Russian translation]. Moscow: Mir, 1975.
2. *Linnik Yu. V.* The Least-Squares Method and Foundations of the Theory of Observation Processing [in Russian]. Moscow: Fizmatgiz, 1961.
3. *Egorshin A. O.* Computational closed methods of identification of linear objects // Optimal and Self-Adjusting Systems [in Russian]. Novosibirsk: Izdat. IAiE SO AN SSSR, 1971. P. pp. 40–53.
4. *Aoki M. and Yue P. C.* On priory error estimates of some identification methods // IEEE Trans. Automat. Control. 1970. V. 15, N 5. P. 541–548.
5. *Egorshin A. O.* Optimization of parameters of stationary models in a unitary space // Automation and Remote Control. 2004. V. 65, N 12. P. 1885–1903.
6. *Egorshin A. O.* Identification and discretization of linear differential equations with constant coefficients // Vestnik NGU Ser. Mat. Mekh. Informat. 2014. V. 14, N 3. P. 29–42.
7. *Egorshin A. O.* On a method of estimation of modeling coefficients for sequences // Sibirsk. Zh. Indust. Mat. 2000. V. 3, N 2. P. 78–96.
8. *Egorshin A. O.* On a variational problem of smoothing // Vestnik Udmurt. Univ. Mat. Mekh. Komp′yut. Nauki. 2011. N 4. P. 9–22.
9. *Egorshin A. O.* On a variational problem of dynamical piecewise-linear approximation // Vestnik Udmurt. Univ. Mat. Mekh. Komp′yut. Nauki. 2012. N 4. P. 30–45.
10. *Gantmakher F. R.* The Theory of Matrices [in Russian]. Moscow: Nauka, 1966.
11. *El′yasberg P. E.* Determination of Motion from Measurements of Results [in Russian]. Moscow: URSS, 2011.
12. *Egorshin A. O.* On watching the extremum parameters in the identification variational problem // Vestnik NGU Ser. Mat. Mekh. Informat. 2011. V. 11, N 3. P. 95–114.

December 15, 2014

A. O. Egorshin
Sobolev Institute of Mathematics, Novosibirsk, Russia
egorshin@math.nsc.ru

Yakutian Mathematical Journal
July—September, 2014. Vol. 21, No. 3

UDC 519.21

A PROOF OF THE GENERALIZED ITÔ—WENTZELL FORMULA VIA THE DELTA–FUNCTION AND THE DENSITY OF NORMAL DISTRIBUTION

E. V. Karachanskaya

Abstract. We prove the generalized Itô–Wentzell formula on using a stochastic approximation and the density function of the normal distribution.

Keywords: Itô–Wentzell formula, generalized Itô equation, Poisson measure, δ-function, density of normal distribution, mean square convergence

Introduction

The rules for constructing stochastic differentials, e.g., the chain rule, are very important in the theory of stochastic random processes. These are Itô's formula [1, 2] for the differential of a nonrandom function of a random process and the Itô–Wentzell formula [3] enabling us to construct the differential of a function which is itself a solution to a stochastic equation. Many articles address the derivation of these formulas for various classes of processes by extending Itô's formula and the Itô–Wentzell formula to a larger class of functions (for instance, see [4–12] for Itô's formula and [13–19] for the Itô–Wentzell formula).

The next level is to obtain a new formula for the generalized Itô equation [2] which involves Wiener and Poisson components. Doobko in 2002 presented [20] a generalization of stochastic differentials of random functions satisfying the generalized Itô equation based on expressions for the kernels of integral invariants (only the ideas of a possible proof were sketched in [20]). The result is called the *generalized Itô–Wentzell formula.* In 2007 a generalized Itô–Wentzell formula was constructed [21] for one-dimensional processes in the absence of Wiener component, basing on the classical theory of stochastic differential equations by Itô; and in 2013, the "Itô–Wentzell formula with jumps" for one-dimensional random processes with both Wiener and Poisson components appeared [22] along with a reference to the proof in [21].

In contrast to [20], the generalized Itô–Wentzell formula for the Poisson measure is suggested in [23]. In this case the requirement on the character of the Poisson distribution is only a general restriction, as the knowledge of its explicit form is unnecessary. The proof [24] of the generalized Itô–Wentzell formula uses the method of stochastic integral invariants and equations for their kernels.

In this article we present a proof that is based on traditional stochastic analysis and the use of approximations to random functions related to stochastic differential equations by averaging their values at each point. For this approximation we apply a sequence of density functions of normal distributions with the variance tending to zero. Since the exponential is an infinitely differentiable function (while the second derivative suffices) and the corresponding sequence converges quickly, the prelimit expression enables us to apply the generalized Itô formula without additional remarks.

The generalized Itô–Wentzell formula relying on the kernels of integral invariants [23, 24] requires stricter conditions on the coefficients of all equations under consideration: the existence of second derivatives [25]. The reason is that the kernels of invariants for differential equations exist under certain restrictions on the coefficients.

1. Notation and Preliminaries

Consider a filtered probability space $\big(\Omega, \mathscr{F}, \{\mathscr{F}_t\}_0^T, \mathbf{P}\big)$ and an m-dimensional Wiener process $\mathbf{w}(t) = (w_1(t), \dots, w_m(t))^T$ such that the one-dimensional Wiener processes $w_k(t)$ for $k = 1, \dots, m$ defined on the space are $\mathscr{F}_t$-measurable and mutually independent.

Take a vector γ with values in $\mathbb{R}^{n'}$. Denote by $\nu(\Delta t, \Delta\gamma)$ the standard Poisson measure on $[0, T] \times \mathbb{R}^{n'}$ modeling independent random variables on disjoint intervals and sets. The one-dimensional Wiener processes $w_k(t)$ for $k = 1, \dots, m$ and the Poisson measure $\nu([0; T], \mathscr{A})$ defined on the specified space are $\mathscr{F}_t$-measurable and independent of one another. The random processes and functions appearing below are $\mathscr{F}_t$-measurable and agree with the processes mentioned.

Our notation is similar to that of [26]:

$$H_2[0,T] = \Big\{\alpha(t),\ \alpha : [0,T] \to \mathbb{R}^m \mid \int\limits_0^T |\alpha(t)|^2\, dt < \infty \ \text{ almost surely}\Big\},$$

$$H_s(\Pi) = \Big\{\varphi(t,y) = \varphi(t,y;\omega),\ \varphi : [0,T] \times \mathbb{R} \times \Omega \to \mathbb{R}^{m'} \\ \mid \int\limits_0^T \int\limits_{\mathbb{R}} |\varphi(t,y)|^s \Pi(dy)\, dt < \infty \ \text{ almost surely}\Big\}.$$

Consider a random process $\mathbf{x}(t)$ with values in $\mathbb{R}^n$ defined by the equation [26, pp. 271–272]

$$d\mathbf{x}(t) = A(t)dt + B(t)d\mathbf{w}(t) + \int g(t,\gamma)\nu(dt, d\gamma), \tag{1}$$

where $A(t) = \{a_1(t), \dots, a_n(t)\}^*$, $g(t,\gamma) = \{g_1(t,\gamma), \dots, g_n(t,\gamma)\}^* \in \mathbb{R}^n$ and $\gamma \in \mathbb{R}^{n'}$, while $\mathbf{w}(t)$ is an m-dimensional Wiener process; moreover,

$$\begin{gathered} |g(t,\gamma)| \in H_{1,2}(\Pi), \quad \sqrt{|a_j(t)|}, |b_{j,k}(t)| \in H_2[0,T], \\ B(t) = (b_{j,k}(t)), \quad j = 1, \dots, n,\ k = 1, \dots, m. \end{gathered} \tag{2}$$

In general the coefficients $A(t)$, $B(t)$, and $g(t,\gamma)$ are random functions depending also on $\mathbf{x}(t)$. Since the restrictions on these coefficients relate explicitly only to the variables t and γ, we use precisely this notation for the coefficients of (1) instead of writing $A(t, \mathbf{x}(t))$, $B(t, \mathbf{x}(t))$, and $g(t, \mathbf{x}(t); \gamma)$.

The differential of a random function $F(t, \mathbf{x}(t))$, where $\mathbf{x}(t)$ is a solution to (1), can be expressed by the *generalized Itô formula* [2] (which we use as in [26, pp. 271–

272]):

$$d_tF(t,\mathbf{x}(t)) = \Big[\frac{\partial F(t,\mathbf{x})}{\partial t}\Big|_{\mathbf{x}=\mathbf{x}(t)} + \sum_{i=1}^{n} a_i(t,\mathbf{x}(t))\frac{\partial F(t,\mathbf{x})}{\partial x_i}\Big|_{\mathbf{x}=\mathbf{x}(t)}$$

$$+\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{m} b_{i\,k}(t,\mathbf{x}(t))b_{j\,k}(t,\mathbf{x}(t))\frac{\partial^2 F(t,\mathbf{x})}{\partial x_i\partial x_j}\Big|_{\mathbf{x}=\mathbf{x}(t)}\Big]\,dt$$

$$+\sum_{i=1}^{n}\sum_{k=1}^{m} b_{i\,k}(t,\mathbf{x}(t))\frac{\partial F(t,\mathbf{x})}{\partial x_i}\Big|_{\mathbf{x}=\mathbf{x}(t)} dw_k(t)$$

$$+\int [F(t,\mathbf{x}(t)+g(t,\mathbf{x}(t);\gamma)) - F(t,\mathbf{x}(t))]\nu(dt,d\gamma). \tag{3}$$

The δ-function is a convenient tool when we need to extract the value of a function at a certain point while disregarding its continuity or smoothness. Consider the properties of the δ-function:

$$\int\limits_{-\infty}^{+\infty} \delta(x)\,dx = 1, \tag{4}$$

$$f(x) = \int\limits_{-\infty}^{+\infty} f(y)\delta(y-x)\,dy. \tag{5}$$

The first of them enables us to use the distribution density as the $\delta(x)$-function. Choose the distribution density function satisfying the conditions:

(1) it is symmetric about the point x;

(2) it attains maximum at x;

(3) it is concentrated near x (this is necessary to extract the value $f(x)$);

(4) it is at least twice differentiable at x (Itô's formula requires this condition).

The density function of the normal distribution $\mathscr{N}(x,\varepsilon^2)$ as $\varepsilon\to 0$ meets these requirements. The second property of the δ-function enables us to introduce for each section $f_\varepsilon(t_\alpha,x)$, where $t=t_\alpha$ with $t_\alpha\in[0,T]$, the function

$$f_\varepsilon(t_\alpha,x) = \frac{1}{\varepsilon\sqrt{2\pi}}\int\limits_{-\infty}^{\infty} f(t_\alpha,y)\exp\Big\{-\frac{(y-x)^2}{2\varepsilon^2}\Big\}\,dy. \tag{6}$$

Thus, for every point x we approximate the random function $f(t,x)$ by a sequence of nonrandom functions $f_\varepsilon(t,x)$ as $\varepsilon\to 0$.

2. The Formula and Its Proof

Theorem 1 (generalized Itô–Wentzell formula). *Consider the real function $F(t,\mathbf{x})$ of $(t,\mathbf{x})\in[0;T]\times\mathbb{R}^n$ with generalized stochastic differential of the form*

$$d_tF(t,\mathbf{x}) = Q(t,\mathbf{x})\,dt + \sum_{k=1}^{m} D_k(t,\mathbf{x})dw_k(t) + \int G(t,\mathbf{x};\gamma)\nu(dt,d\gamma) \tag{7}$$

whose coefficients satisfy the conditions:

(a) *in general $Q(t,\mathbf{x})$, $D_k(t,\mathbf{x})$, $G(t,\mathbf{x};\gamma)\in\mathbb{R}$ are random functions measurable with respect to the same nondecreasing flow $\{\mathscr{F}_t\}_0^T$ of σ-algebras as the processes $w(t)$ and $\nu(t,\mathscr{A})$ for every $\mathscr{A}\in\mathfrak{B}$ in a fixed Borel σ-algebra* [26, p. 266]*;*

(b) *the sections of random functions* $Q(t_\alpha, \mathbf{x})$, $D_k(t_\alpha, \mathbf{x})$, $G(t_\alpha, \mathbf{x}; \gamma) \in \mathbb{R}$ *for all* $t = t_\alpha$ *with* $\alpha \in [0, T]$ *have normal distribution* $\mathscr{N}(x, \varepsilon^2)$ *as* $\varepsilon \to 0$*;*

(c) *with probability 1, the functions* $Q(t, \mathbf{x})$, $D_k(t, \mathbf{x})$, *and* $G(t, \mathbf{x}; u)$, *together with their first and second partial derivatives with respect to the components of* $\mathbf{x}$, *are continuous, bounded, and satisfy Hölder's condition.*

If a random process $\mathbf{x}(t)$ *obeys* (1) *and the restrictions* (2) *then the stochastic differential exists and*

$$
\begin{aligned}
d_t F(t, \mathbf{x}(t)) &= Q(t, \mathbf{x}(t))\, dt + \sum_{k=1}^{m} D_k(t, \mathbf{x}(t)) dw_k \\
&+ \Bigg[\sum_{i=1}^{n} a_i(t) \frac{\partial F(t, \mathbf{x})}{\partial x_i} \bigg|_{\mathbf{x}=\mathbf{x}(t)} + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{m} b_{i,k}(t) b_{j,k}(t) \frac{\partial^2 F(t, \mathbf{x})}{\partial x_i \partial x_j} \bigg|_{\mathbf{x}=\mathbf{x}(t)} \\
&+ \sum_{i=1}^{n} b_{i,k}(t) \frac{\partial D_k(t, \mathbf{x})}{\partial x_i} \bigg|_{\mathbf{x}=\mathbf{x}(t)} \Bigg] dt + \sum_{i=1}^{n} \sum_{k=1}^{m} b_{i,k}(t) \frac{\partial F(t, \mathbf{x})}{\partial x_i} \bigg|_{\mathbf{x}=\mathbf{x}(t)} dw_k \\
&+ \int [(F(t, \mathbf{x}(t) + g(t, \gamma)) - F(t, \mathbf{x}(t))] \nu(dt, d\gamma) \\
&+ \int G(t, \mathbf{x}(t) + g(t, \gamma); \gamma) \nu(dt, d\gamma). \qquad (8)
\end{aligned}
$$

Since the available properties of the δ-function relate to deterministic functions, while the functions we face are random in general, the proof rests on the following propositions.

Proposition 1. *If a function* $f(t, x)$ *is bounded with probability 1 for all* $t \in [0, T]$ *and satisfies Hölder's condition in* x *then*

$$
\begin{aligned}
f(t, x) &= \lim_{\varepsilon \downarrow 0} f_\varepsilon(t, x) = \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon \sqrt{2\pi}} \int_{-\infty}^{\infty} f(t, y) \exp\left\{ -\frac{(y - x)^2}{2\varepsilon^2} \right\} dy \\
&= \int_{-\infty}^{\infty} f(t; y) \delta(y - x)\, dy. \qquad (9)
\end{aligned}
$$

Proof. Assume that $f(t, x)$ satisfies Hölder's condition

$$|f(t, y_1) - f(t, y_2)| \le L(t) |y_1 - y_2|^\varsigma, \quad 0 < \varsigma \le 1.$$

In the proofs we treat t as a fixed parameter, and we so simplify expressions by omitting t. Therefore, instead of $f(t, x)$ we write $f(x)$ and instead of $L(t)$, simply L.

Change the variables: $(y - x)\varepsilon^{-1} = z$, and so $y = \varepsilon z + x$. Adding and subtracting the same expression and using the properties of the probability integral, we obtain

$$
\begin{aligned}
\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(\varepsilon z + x) \exp\{-z^2/2\}\, dz &= f(x) \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\{-z^2/2\}\, dz \\
&+ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (f(\varepsilon z + x) - f(x)) \exp\{-z^2/2\}\, dz \\
&= f(x) + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (f(\varepsilon z + x) - f(x)) \exp\{-z^2/2\}\, dz. \qquad (10)
\end{aligned}
$$

Estimate the last integral:

$$\left|\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{\infty}(f(\varepsilon z+x)-f(x))\exp\{-z^2/2\}\,dz\right|\le\frac{L}{\sqrt{2\pi}}\int\limits_{-\infty}^{\infty}|(\varepsilon z+x)-x|^{\varsigma}\exp\{-z^2/2\}\,dz$$

$$\le\varepsilon L\frac{2}{\sqrt{2\pi}}\Bigg[\int\limits_{0}^{1}z^{\varsigma}\exp\{-z^2/2\}\,dz-\int\limits_{1}^{\infty}\exp\{-z^2/2\}d(-z^2/2)$$

$$\le\varepsilon L\frac{2}{\sqrt{2\pi}}[z^{\varsigma}|_{z=1}+1]\le\varepsilon L\frac{4}{\sqrt{2\pi}}. \qquad (11)$$

Hence, as $\varepsilon\to0$ with $\varepsilon>0$, we have (9) for arbitrary functions $f(t,x)$ which for all $t\in[0,T]$ are bounded with probability 1 and satisfy Hölder's condition in x.

Consequently, the result of the convolution of the δ-function with a random function is also a random function, i.e., we obtain a result similar to the deterministic case.

Since Itô's formula is inapplicable to the δ-function, we need the following result. For convenience, in the subsequent manipulations, put

$$\delta_\varepsilon(y-x)=\frac{1}{\varepsilon\sqrt{2\pi}}\exp\left\{-\frac{(y-x)^2}{2\varepsilon^2}\right\}.$$

Proposition 2. *If $f(t,x)$ and its first and second derivatives with respect to x are bounded with probability 1 for all $t\in[0,T]$ and satisfy Hölder's condition in x then*

$$\frac{\partial}{\partial x}f(t,x)=-\lim_{\varepsilon\downarrow0}\int\limits_{-\infty}^{\infty}f(t,y)\frac{\partial}{\partial y}\delta_\varepsilon(y-x)\,dy$$

$$=\lim_{\varepsilon\downarrow0}\frac{1}{\varepsilon\sqrt{2\pi}}\int\limits_{-\infty}^{\infty}\exp\left\{-\frac{(y-x)^2}{2\varepsilon^2}\right\}\frac{\partial}{\partial y}f(t,y)\,dy, \qquad (12)$$

$$\frac{\partial^2}{\partial x^2}f(t,x)=\lim_{\varepsilon\downarrow0}\int\limits_{-\infty}^{\infty}f(t,y)\frac{\partial^2}{\partial y^2}\delta_\varepsilon(y-x)\,dy$$

$$=\lim_{\varepsilon\downarrow0}\frac{1}{\varepsilon\sqrt{2\pi}}\int\limits_{-\infty}^{\infty}\exp\left\{-\frac{(y-x)^2}{2\varepsilon^2}\right\}\frac{\partial^2}{\partial y^2}f(t,y)\,dy. \qquad (13)$$

Proof. Verify that in this case we also have estimates similar to the above based on the corresponding limit expressions. Differentiate (9) and integrate by parts:

$$\frac{\partial}{\partial x}f(x)=\int\limits_{-\infty}^{\infty}f(y)\frac{\partial}{\partial x}\delta(y-x)\,dy=\lim_{\varepsilon\downarrow0}\frac{1}{\varepsilon\sqrt{2\pi}}\int\limits_{-\infty}^{\infty}f(y)\frac{\partial}{\partial x}\exp\left\{-\frac{(y-x)^2}{2\varepsilon^2}\right\}dy$$

$$=-\lim_{\varepsilon\downarrow0}\frac{1}{\varepsilon\sqrt{2\pi}}\int\limits_{-\infty}^{\infty}f(y)\frac{\partial}{\partial y}\exp\left\{-\frac{(y-x)^2}{2\varepsilon^2}\right\}dy$$

$$= -\lim_{\varepsilon\downarrow 0} f(y)\frac{1}{\varepsilon\sqrt{2\pi}}\exp\left\{-\frac{(y-x)^2}{2\varepsilon^2}\right\}\Bigg|_{-\infty}^{+\infty}$$

$$+\lim_{\varepsilon\downarrow 0}\frac{1}{\varepsilon\sqrt{2\pi}}\int\limits_{-\infty}^{\infty}\exp\left\{-\frac{(y-x)^2}{2\varepsilon^2}\right\}\frac{\partial}{\partial y}f(y)\,dy, \quad \varepsilon > 0.$$

Provided that $f_y'(y)$ satisfies Hölder's condition, for the asymptotic expression of the derivative of $\delta(x)$ the proof coincides with the previous argument. Similarly we establish the relations that involve the second derivatives of $\delta(x)$ whenever $f_y''(y)$ satisfies Hölder's condition. Generalizing the restrictions used, we see that we must require $f_y'(y)$ to exist and be continuous, and the second derivatives to be Hölder.

Proceed to justify (8) by using the properties of $\delta(x)$ on the class of all continuous functions bounded with probability 1 which are established in Propositions 1 and 2.

Proof of the theorem. Consider a solution $F(t,\mathbf{x}(t))$ to (8). Introduce the function $F_\varepsilon(t,\mathbf{x}(t))$ as

$$F_\varepsilon(t,\mathbf{x}(t)) = \int\limits_{\mathbb{R}^n}\prod_{l=1}^{n}\delta_\varepsilon(y_l - x_l(t))F(t,\mathbf{y})\,d\Gamma(\mathbf{y}), \quad d\Gamma(\mathbf{y}) = \prod_{l=1}^{n}dy_l. \tag{14}$$

Let us show that the assumption $F(0,\mathbf{x}(0)) = F_\varepsilon(0,\mathbf{x}(0))$ yields the mean square convergence:

$$F(t,\mathbf{x}(t)) = \underset{\varepsilon\downarrow 0}{\text{l.i.m.}}F_\varepsilon(t,\mathbf{x}(t)). \tag{15}$$

By definition, this is equivalent to

$$\lim_{\varepsilon\downarrow 0}\mathbf{M}[|F_\varepsilon(t,\mathbf{x}(t)) - F(t,\mathbf{x}(t))|^2] = 0, \tag{16}$$

which we are going to verify.

Using the generalized Itô formula (3), differentiate the expression

$$\prod_{l=1}^{n}\delta_\varepsilon(y_l - x_l(t))F(t,\mathbf{y})$$

in (14), putting $h(t,\xi) = h(t,\mathbf{x};F)$ and

$$h(t,\xi(t)) = h(t,\mathbf{x}(t);F(t,\mathbf{y})) = \prod_{l=1}^{n}\delta_\varepsilon(y_l - x_l(t))F(t,\mathbf{y}),$$

where $\mathbf{x}(t)$ obeys (1), while $F(t,\mathbf{x})$ obeys (7).

We arrive at the following result (we omit the sign summation and assume summation over repeated indices):

$$d_t\left[\prod_{l=1}^{n}\delta_\varepsilon(y_l - x_l(t))F(t,\mathbf{y})\right] = \left[a_i(t)F(t,\mathbf{y})\frac{\partial}{\partial x_i}\prod_{l=1}^{n}\delta_\varepsilon(y_l - x_l)\Bigg|_{\xi=\xi(t)}\right.$$

$$+Q(t,\mathbf{y})\prod_{l=1}^{n}\delta_\varepsilon(y_l - x_l(t)) + \frac{1}{2}b_{i\,k}(t)b_{j\,k}(t)F(t,\mathbf{y})\frac{\partial^2}{\partial x_i\partial x_j}\prod_{l=1}^{n}\delta_\varepsilon(y_l - x_l)\Bigg|_{\xi=\xi(t)}$$

$$\left.+b_{i\,k}(t)D_k(t,\mathbf{y})\frac{\partial}{\partial x_i}\prod_{l=1}^{n}\delta_\varepsilon(y_l - x_l)\Bigg|_{\xi=\xi(t)}\right]dt$$

$$
\begin{aligned}
+\bigg[F(t,\mathbf{y})b_{i\,k}(t)\frac{\partial}{\partial x_i}\prod_{l=1}^{n}\delta_\varepsilon(y_l-x_l)\bigg|_{\xi=\xi(t)} &+D_k(t,\mathbf{y})\prod_{l=1}^{n}\delta_\varepsilon(y_l-x_l(t))\bigg]dw_k(t)\\
+\int\bigg[\prod_{l=1}^{n}\delta_\varepsilon(y_l-x_l(t)-g_l(t,\gamma))&(F(t,\mathbf{y})+G(t,\mathbf{y};\gamma))\\
&-\prod_{l=1}^{n}\delta_\varepsilon(y_l-x_l(t))F(t,\mathbf{y})\bigg]\nu(dt,d\gamma).
\end{aligned}
\tag{17}
$$

Using Lemmas 1 and 2, on the right-hand side we pass from the partial derivatives with respect to the components of $\mathbf{x}$ to the partial derivative with respect to the components of $\mathbf{y}$, and express the resulting stochastic ordinary differential equation in integral form, integrating over $\mathbb{R}^n$ and taking (14) into account:

$$
\begin{aligned}
F_\varepsilon(t,\mathbf{x}(t)) &= \int\limits_{\mathbb{R}^n}\prod_{l=1}^{n}\delta_\varepsilon(y_l-x_l(t))F(t,\mathbf{y})\,d\Gamma(\mathbf{y}) = \int\limits_{\mathbb{R}^n}\prod_{l=1}^{n}\delta_\varepsilon\big(y_l^0-x_l(0)\big)F(0,\mathbf{y}_0)\,d\Gamma(\mathbf{y})\\
&-\int\limits_{\mathbb{R}^n}\int\limits_0^t a_i(\tau)F(\tau,\mathbf{y})\frac{\partial}{\partial y_i}\prod_{l=1}^{n}\delta_\varepsilon(y_l-x_l(\tau))\,d\tau d\Gamma(\mathbf{y})\\
&+\int\limits_{\mathbb{R}^n}\int\limits_0^t Q(\tau,\mathbf{y})\prod_{l=1}^{n}\delta_\varepsilon(y_l-x_l(\tau))\,d\tau d\Gamma(\mathbf{y})\\
&+\frac{1}{2}\int\limits_{\mathbb{R}^n}\int\limits_0^t b_{i\,k}(\tau)b_{j\,k}(\tau)F(\tau,\mathbf{y})\frac{\partial^2}{\partial y_i\partial y_j}\prod_{l=1}^{n}\delta_\varepsilon(y_l-x_l(\tau))\,d\tau d\Gamma(\mathbf{y})\\
&-\int\limits_{\mathbb{R}^n}\int\limits_0^t b_{i\,k}(\tau)D_k(\tau,\mathbf{y})\frac{\partial}{\partial y_i}\prod_{l=1}^{n}\delta_\varepsilon(y_l-x_l(\tau))\,d\tau d\Gamma(\mathbf{y})\\
&-\int\limits_{\mathbb{R}^n}\int\limits_0^t F(\tau;\mathbf{y})b_{i\,k}(\tau)\frac{\partial}{\partial y_i}\prod_{l=1}^{n}\delta_\varepsilon(y_l-x_l(\tau))\,dw_k(\tau)\,d\Gamma(\mathbf{y})\\
&+\int\limits_{\mathbb{R}^n}\int\limits_0^t D_k(\tau;\mathbf{y})\prod_{l=1}^{n}\delta_\varepsilon(y_l-x_l(\tau))\,dw_k(\tau)\,d\Gamma(\mathbf{y})\\
&+\int\limits_{\mathbb{R}^n}\int\limits_0^t\int\bigg[\prod_{l=1}^{n}\delta_\varepsilon(y_l-x_l(\tau)-g_l(\tau;\gamma))(F(\tau,\mathbf{y})+G(\tau,\mathbf{y};\gamma))\\
&\qquad-\prod_{l=1}^{n}\delta_\varepsilon(y_l-x_l(\tau))F(\tau,\mathbf{y})\bigg]\nu(d\tau;d\gamma)\,d\Gamma(\mathbf{y}).
\end{aligned}
\tag{18}
$$

Put (8) into integral form to find that

$$
\begin{aligned}
F(t,\mathbf{x}(t)) = F(0;\mathbf{x}(0)) &+ \int\limits_0^t Q(\tau;\mathbf{x}(\tau))\,d\tau + \int\limits_0^t D_k(\tau;\mathbf{x}(\tau))\,dw_k(\tau)\\
&+\int\limits_0^t b_{i,k}(\tau)\frac{\partial F(\tau;\mathbf{x})}{\partial x_i}\bigg|_{\mathbf{x}=\mathbf{x}(\tau)}dw_k(\tau)
\end{aligned}
$$

$$+\int\limits_0^t\biggl[a_i(\tau)\frac{\partial F(\tau;\mathbf{x})}{\partial x_i}\bigg|_{\mathbf{x}=\mathbf{x}(\tau)}+\frac{1}{2}b_{i,k}(\tau)b_{j,k}(\tau)\frac{\partial^2 F(\tau;\mathbf{x})}{\partial x_i\partial x_j}\bigg|_{\mathbf{x}=\mathbf{x}(\tau)}$$

$$+b_{i,k}(\tau)\frac{\partial D_k(\tau;\mathbf{x})}{\partial x_i}\bigg|_{\mathbf{x}=\mathbf{x}(\tau)}\biggr]\,d\tau$$

$$+\int\limits_0^t\int[F(\tau;\mathbf{x}(\tau)+g(\tau;\gamma))-F(\tau;\mathbf{x}(\tau))]\nu(d\tau;d\gamma)$$

$$+\int\limits_0^t\int G(\tau;\mathbf{x}(\tau)+g(\tau;\gamma);\gamma)\nu(d\tau;d\gamma). \tag{19}$$

Consider the difference between (18) and (19), taking into account the prelimit properties of the δ-function and the possibility of changing the order of integration:

$$F_\varepsilon(t,\mathbf{x}(t))-F(t,\mathbf{x}(t))=F_\varepsilon(0,\mathbf{x}(0))-F(0;\mathbf{x}(0))$$

$$+\int\limits_0^t a_i(\tau)\biggl[-\int\limits_{\mathbb{R}^n}F(\tau,\mathbf{y})\frac{\partial}{\partial y_i}\prod_{l=1}^n\delta_\varepsilon(y_l-x_l(\tau))\,d\Gamma(\mathbf{y})-\frac{\partial F(\tau;\mathbf{x})}{\partial x_i}\bigg|_{\mathbf{x}=\mathbf{x}(\tau)}\biggr]\,d\tau$$

$$+\int\limits_0^t\biggl[\int\limits_{\mathbb{R}^n}Q(\tau,\mathbf{y})\prod_{l=1}^n\delta_\varepsilon(y_l-x_l(\tau))\,d\Gamma(\mathbf{y})-Q(\tau;\mathbf{x}(\tau))\biggr]\,d\tau$$

$$+\frac{1}{2}\int\limits_0^t b_{i\,k}(\tau)b_{j\,k}(\tau)\biggl[\int\limits_{\mathbb{R}^n}F(\tau,\mathbf{y})\frac{\partial^2}{\partial y_i\partial y_j}\prod_{l=1}^n\delta_\varepsilon(y_l-x_l(\tau))\,d\Gamma(\mathbf{y})-\frac{\partial^2 F(\tau;\mathbf{x})}{\partial x_i\partial x_j}\bigg|_{\mathbf{x}=\mathbf{x}(\tau)}\biggr]\,d\tau$$

$$+\int\limits_0^t b_{i\,k}(\tau)\biggl[-\int\limits_{\mathbb{R}^n}D_k(\tau,\mathbf{y})\frac{\partial}{\partial y_i}\prod_{l=1}^n\delta_\varepsilon(y_l-x_l(\tau))\,d\Gamma(\mathbf{y})-\frac{\partial D_k(\tau;\mathbf{x})}{\partial x_i}\bigg|_{\mathbf{x}=\mathbf{x}(\tau)}\biggr]\,d\tau$$

$$-\int\limits_0^t b_{i\,k}(\tau)\biggl[-\int\limits_{\mathbb{R}^n}F(\tau;\mathbf{y})\frac{\partial}{\partial y_i}\prod_{l=1}^n\delta_\varepsilon(y_l-x_l(\tau))\,d\Gamma(\mathbf{y})-\frac{\partial F(\tau;\mathbf{x})}{\partial x_i}\bigg|_{\mathbf{x}=\mathbf{x}(\tau)}\biggr]dw_k(\tau)$$

$$+\int\limits_0^t\biggl[\int\limits_{\mathbb{R}^n}D_k(\tau;\mathbf{y})\prod_{l=1}^n\delta_\varepsilon(y_l-x_l(\tau))\,d\Gamma(\mathbf{y})-D_k(\tau;\mathbf{x}(\tau))\biggr]\,dw_k(\tau)$$

$$+\int\limits_0^t\int\biggl(\int\limits_{\mathbb{R}^n}\biggl[\prod_{l=1}^n\delta_\varepsilon(y_l-x_l(\tau)-g_i(\tau;\gamma))(F(\tau,\mathbf{y})+G(\tau,\mathbf{y};\gamma))$$

$$-\prod_{l=1}^n\delta_\varepsilon(y_l-x_l(\tau))F(\tau,\mathbf{y})\biggr]\,d\Gamma(\mathbf{y})$$

$$-F(\tau;\mathbf{x}(\tau)+g(\tau;\gamma))+F(\tau;\mathbf{x}(\tau))-G(\tau;\mathbf{x}(\tau)+g(\tau;\gamma);\gamma)\biggr)\nu(d\tau;d\gamma). \tag{20}$$

Since the δ-function is not Itô differentiable, use Propositions 1 and 2, namely, (9), (12), and (13):

$$|F_\varepsilon(t,\mathbf{x}(t))-F(t,\mathbf{x}(t))|\le|F_\varepsilon(0,\mathbf{x}(0))-F(0;\mathbf{x}(0))|$$

$$
\begin{aligned}
&+\Bigg|\int\limits_0^t a_i(\tau)\Bigg[\frac{1}{(\varepsilon\sqrt{2\pi})^n}\int\limits_{\mathbb{R}^n}\prod_{l=1}^n \exp\Bigg\{-\frac{(y_l-x_l(\tau))^2}{2\varepsilon^2}\Bigg\}\\
&\qquad\times\frac{\partial}{\partial y_l}F(\tau,\mathbf{y})\,d\Gamma(\mathbf{y})-\frac{\partial F(\tau;\mathbf{x})}{\partial x_i}\Bigg|_{\mathbf{x}=\mathbf{x}(\tau)}\Bigg]\,d\tau\Bigg|\\
&+\Bigg|\int\limits_0^t\Bigg[\frac{1}{(\varepsilon\sqrt{2\pi})^n}\int\limits_{\mathbb{R}^n}\prod_{l=1}^n \exp\Bigg\{-\frac{(y_l-x_l(\tau))^2}{2\varepsilon^2}\Bigg\}Q(\tau,\mathbf{y})\,d\Gamma(\mathbf{y})-Q(\tau;\mathbf{x}(\tau))\Bigg]\,d\tau\Bigg|\\
&+\Bigg|\frac{1}{2}\int\limits_0^t\Bigg[\frac{1}{(\varepsilon\sqrt{2\pi})^n}\int\limits_{\mathbb{R}^n}\prod_{l=1}^n \exp\Bigg\{-\frac{(y_l-x_l(\tau))^2}{2\varepsilon^2}\Bigg\}\frac{\partial^2}{\partial y_l\partial y_j}F(\tau,\mathbf{y})\,d\Gamma(\mathbf{y})\\
&\qquad-\frac{\partial^2 F(\tau;\mathbf{x})}{\partial x_i\partial x_j}\Bigg|_{\mathbf{x}=\mathbf{x}(\tau)}\Bigg]b_{i\,k}(\tau)b_{j\,k}(\tau)\,d\tau\Bigg|\\
&+\Bigg|\int\limits_0^t b_{i\,k}(\tau)\Bigg[\frac{1}{(\varepsilon\sqrt{2\pi})^n}\int\limits_{\mathbb{R}^n}\prod_{l=1}^n \exp\Bigg\{-\frac{(y_l-x_l(\tau))^2}{2\varepsilon^2}\Bigg\}\frac{\partial}{\partial y_l}D_k(\tau,\mathbf{y})\,d\Gamma(\mathbf{y})\\
&\qquad-\frac{\partial D_k(\tau;\mathbf{x})}{\partial x_i}\Bigg|_{\mathbf{x}=\mathbf{x}(\tau)}\Bigg]\,d\tau\Bigg|\\
&+\Bigg|\int\limits_0^t b_{i\,k}(\tau)\Bigg[\frac{1}{(\varepsilon\sqrt{2\pi})^n}\int\limits_{\mathbb{R}^n}\prod_{l=1}^n \exp\Bigg\{-\frac{(y_l-x_l(\tau))^2}{2\varepsilon^2}\Bigg\}\frac{\partial}{\partial y_l}F(\tau;\mathbf{y})\,d\Gamma(\mathbf{y})\\
&\qquad-\frac{\partial F(\tau;\mathbf{x})}{\partial x_i}\Bigg|_{\mathbf{x}=\mathbf{x}(\tau)}\Bigg]dw_k(\tau)\Bigg|\\
&+\Bigg|\int\limits_0^t\Bigg[\frac{1}{(\varepsilon\sqrt{2\pi})^n}\int\limits_{\mathbb{R}^n}\prod_{l=1}^n \exp\Bigg\{-\frac{(y_l-x_l(\tau))^2}{2\varepsilon^2}\Bigg\}D_k(\tau;\mathbf{y})\,d\Gamma(\mathbf{y})-D_k(\tau;\mathbf{x}(\tau))\Bigg]dw_k(\tau)\Bigg|\\
&+\Bigg|\int\limits_0^t\int\Bigg[\frac{1}{(\varepsilon\sqrt{2\pi})^n}\int\limits_{\mathbb{R}^n}\prod_{l=1}^n \exp\Bigg\{-\frac{(y_l-x_l(\tau)-g_l(\tau;\gamma))^2}{2\varepsilon^2}\Bigg\}\\
&\qquad\times F(\tau,\mathbf{y})d\Gamma(\mathbf{y})-F(\tau;\mathbf{x}(\tau)+g(\tau;\gamma))\Bigg]\nu(d\tau;d\gamma)\Bigg|\\
&+\Bigg|\int\limits_0^t\int\Bigg[\frac{1}{(\varepsilon\sqrt{2\pi})^n}\int\limits_{\mathbb{R}^n}\prod_{l=1}^n \exp\Bigg\{-\frac{(y_l-x_l(\tau))^2}{2\varepsilon^2}\Bigg\}F(\tau,\mathbf{y})\,d\Gamma(\mathbf{y})-F(\tau;\mathbf{x}(\tau))\Bigg]\nu(d\tau;d\gamma)\Bigg|\\
&+\Bigg|\int\limits_0^t\int\Bigg[\frac{1}{(\varepsilon\sqrt{2\pi})^n}\int\limits_{\mathbb{R}^n}\prod_{l=1}^n \exp\Bigg\{-\frac{(y_l-x_l(\tau)-g_l(\tau;\gamma))^2}{2\varepsilon^2}\Bigg\}\\
&\qquad\times G(\tau,\mathbf{y};\gamma)\big)d\Gamma(\mathbf{y})-G(\tau;\mathbf{x}(\tau)+g(\tau;\gamma);\gamma)\Bigg]\nu(d\tau;d\gamma)\Bigg|.
\end{aligned}
\tag{21}
$$

Apply integration by parts to the integrals with exponentials, then rearrange by analogy with (10) and find the upper bounds similar to (11):

$$\Big|\int\limits_0^t\Big[\frac{1}{(\varepsilon\sqrt{2\pi})^n}\int\limits_{\mathbb{R}^n}\prod_{l=1}^n\exp\Big\{-\frac{(y_l-x_l(\tau))^2}{2\varepsilon^2}\Big\}Q(\tau,\mathbf{y})\,d\Gamma(\mathbf{y})-Q(\tau;\mathbf{x}(\tau))\Big]\,d\tau\Big|$$

$$\le\int\limits_0^t\Big|\frac{1}{(\sqrt{2\pi})^n}\int\limits_{\mathbb{R}^n}\prod_{l=1}^n\exp\Big\{-\frac{z_l^2(\tau)}{2}\Big\}[Q(\tau,\varepsilon\mathbf{z}(\tau)+\mathbf{x}(\tau))-Q(\tau;\mathbf{x}(\tau))]\,d\Gamma(\mathbf{z})\Big|\,d\tau$$

$$\le\int\limits_0^t\frac{\varepsilon^n2^{n+1}}{(\sqrt{2\pi})^n}L^{(1)}(\tau)\,d\tau=\frac{\varepsilon^n2^{n+1}}{(\sqrt{2\pi})^n}\int\limits_0^tL^{(1)}(\tau)\,d\tau. \tag{22}$$

Similarly,

$$\Big|\int\limits_0^t\Big[\frac{1}{(\varepsilon\sqrt{2\pi})^n}\int\limits_{\mathbb{R}^n}\prod_{l=1}^n\exp\Big\{-\frac{(y_l-x_l(\tau))^2}{2\varepsilon^2}\Big\}D_k(\tau,\mathbf{y})\,d\Gamma(\mathbf{y})-D_k(\tau;\mathbf{x}(\tau))\Big]dw_k(\tau)\Big|$$

$$\le\Big|\int\limits_0^t\frac{\varepsilon^n2^{n+1}}{(\sqrt{2\pi})^n}L^{(2)}(\tau)\,dw_k(\tau)\Big|=\frac{\varepsilon^n2^{n+1}}{(\sqrt{2\pi})^n}\Big|\int\limits_0^tL^{(2)}(\tau)dw_k(\tau)\Big|. \tag{23}$$

Furthermore,

$$\Big|\int\limits_0^t\int\Big[\frac{1}{(\varepsilon\sqrt{2\pi})^n}\int\limits_{\mathbb{R}^n}\prod_{l=1}^n\exp\Big\{-\frac{(y_l-x_l(\tau)-g_l(\tau;\gamma))^2}{2\varepsilon^2}\Big\}F(\tau,\mathbf{y})\,d\Gamma(\mathbf{y})$$

$$-F(\tau;\mathbf{x}(\tau)+g(\tau;\gamma))\Big]\nu(d\tau;d\gamma)\Big|\le\frac{\varepsilon^n2^{n+1}}{(\sqrt{2\pi})^n}\int\limits_0^t\int L^{(3)}(\tau;\gamma)\nu(d\tau;d\gamma), \tag{24}$$

$$\Big|\int\limits_0^t\int\Big[\frac{1}{(\varepsilon\sqrt{2\pi})^n}\int\limits_{\mathbb{R}^n}\prod_{l=1}^n\exp\Big\{-\frac{(y_l-x_l(\tau)-g_l(\tau;\gamma))^2}{2\varepsilon^2}\Big\}G(\tau,\mathbf{y};\gamma))\,d\Gamma(\mathbf{y})$$

$$-G(\tau;\mathbf{x}(\tau)+g(\tau;\gamma);\gamma)\Big]\nu(d\tau;d\gamma)\Big|\le\frac{\varepsilon^n2^{n+1}}{(\sqrt{2\pi})^n}\int\limits_0^t\int L^{(4)}(\tau;\gamma)\nu(d\tau;d\gamma). \tag{25}$$

Inserting $z=\frac{y-x}{\varepsilon}$, we obtain

$$\frac{\partial f(\varepsilon z+x)}{\partial z}=\varepsilon\frac{\partial f(\varepsilon z+x)}{\partial x},\qquad\frac{\partial^2f(\varepsilon z+x)}{\partial z^2}=\varepsilon^2\frac{\partial^2f(\varepsilon z+x)}{\partial x^2}.$$

Rearrange the expressions with the first derivatives by analogy with (10), and find the upper bounds similar to (11):

$$\Big|\int\limits_0^ta_i(\tau)\Big[\frac{1}{(\varepsilon\sqrt{2\pi})^n}\int\limits_{\mathbb{R}^n}\prod_{l=1}^n\exp\Big\{-\frac{(y_l-x_l(\tau))^2}{2\varepsilon^2}\Big\}\frac{\partial}{\partial y_l}F(\tau,\mathbf{y})\,d\Gamma(\mathbf{y})$$

$$-\frac{\partial F(\tau;\mathbf{x})}{\partial x_i}\Big|_{\mathbf{x}=\mathbf{x}(\tau)}\Big]d\tau\Big|\le\frac{\varepsilon^n2^{n+1}}{(\sqrt{2\pi})^n}\int\limits_0^t|a_i(\tau)|L^{(5)}(\tau)\,d\tau, \tag{26}$$

$$\Big|\int\limits_0^t b_{i\,k}(\tau)\Big[\frac{1}{(\varepsilon\sqrt{2\pi})^n}\int\limits_{\mathbb{R}^n}\prod_{l=1}^n \exp\Big\{-\frac{(y_l-x_l(\tau))^2}{2\varepsilon^2}\Big\}\frac{\partial}{\partial y_l}D_k(\tau,\mathbf{y})\,d\Gamma(\mathbf{y})$$

$$-\frac{\partial D_k(\tau;\mathbf{x})}{\partial x_i}\Big|_{\mathbf{x}=\mathbf{x}(\tau)}\Big]\,d\tau\Big| \le \frac{\varepsilon^n 2^{n+1}}{(\sqrt{2\pi})^n}\int\limits_0^t |b_{i\,k}(\tau)|L^{(6)}(\tau)\,d\tau, \tag{27}$$

$$\Big|\int\limits_0^t b_{i\,k}(\tau)\Big[\frac{1}{(\varepsilon\sqrt{2\pi})^n}\int\limits_{\mathbb{R}^n}\prod_{l=1}^n \exp\Big\{-\frac{(y_l-x_l(\tau))^2}{2\varepsilon^2}\Big\}\frac{\partial}{\partial y_l}F(\tau;\mathbf{y})\,d\Gamma(\mathbf{y})$$

$$-\frac{\partial F(\tau;\mathbf{x})}{\partial x_i}\Big|_{\mathbf{x}=\mathbf{x}(\tau)}\Big]dw_k(\tau)\Big| \le \frac{\varepsilon^n 2^{n+1}}{(\sqrt{2\pi})^n}\Big|\int\limits_0^t |b_{i\,k}(\tau)|L^{(7)}(\tau)\,dw_k(\tau)\Big|. \tag{28}$$

The expression involving the second derivatives satisfies

$$\Big|\frac{1}{2}\int\limits_0^t\Big[\frac{1}{(\varepsilon\sqrt{2\pi})^n}\int\limits_{\mathbb{R}^n}\prod_{l=1}^n \exp\Big\{-\frac{(y_l-x_l(\tau))^2}{2\varepsilon^2}\Big\}\frac{\partial^2}{\partial y_l\partial y_j}F(\tau,\mathbf{y})\,d\Gamma(\mathbf{y})$$

$$-\frac{\partial^2 F(\tau;\mathbf{x})}{\partial x_i\partial x_j}\Big|_{\mathbf{x}=\mathbf{x}(\tau)}\Big]b_{i\,k}(\tau)b_{j\,k}(\tau)\,d\tau\Big| \le \frac{\varepsilon^n 2^{n+1}}{(\sqrt{2\pi})^n}\int\limits_0^t |b_{i\,k}(\tau)b_{j\,k}(\tau)|L^{(8)}(\tau)\,d\tau. \tag{29}$$

Since $L^{(s)}(t)$ with $s = 1, 3, 4, 5, 6, 8$ are bounded nonrandom functions, while $a_i(t)$ and $b_{j\,k}(t)$ satisfy (2), as $\varepsilon \downarrow 0$ the expressions (22), (24)–(27), and (29) are at most 0.

The integral over the Wiener processes $\int_0^t f(\tau)dw(\tau)$ is defined for the functions $f(\tau) \in H_2[0;t]$ such that the condition

$$\int\limits_0^t f^2(\tau)\,d\tau < \infty$$

holds with probability 1 [26, p. 12], i.e., the restrictions (2) are imposed; consequently, as $\varepsilon \downarrow 0$ the expressions in (23) and (28) are at most 0 as well.

Therefore, if the initial conditions coincide then

$$\lim_{\varepsilon\downarrow 0}|F_\varepsilon(t,\mathbf{x}(t)) - F(t,\mathbf{x}(t))| = 0.$$

This means that (16) and (15) hold:

$$F(t,\mathbf{x}(t)) = \underset{\varepsilon\downarrow 0}{\mathrm{l.\,i.\,m.}}F_\varepsilon(t,\mathbf{x}(t)).$$

Thus, (8) is valid.

The proof of the theorem is complete.

Conclusion

By analogy with the classical Itô and Itô–Wentzell formulas, the generalized Itô–Wentzell formula is promising for various applications. In particular, it helped to obtain equations for the first and stochastic first integrals of the stochastic Itô system [24], equations for the density of stochastic dynamical invariants, Kolmogorov equations for the density of transition probabilities of random processes described by the generalized stochastic Itô differential equation [27], as well as the construction of program controls with probability 1 for stochastic systems [28].

The author is grateful to Professor Doobko for the idea of the proof.

REFERENCES

1. *Itô K.* Stochastic differential equations in a differentiable manifold // Nagoya Math. J. 1950. V. 1. P. 35–473.

2. *Kunita H. and Watanabe S.* On square integrable martingales // Nagoya Math. J. 1967. N 30. P. 209–245.

3. *Venttsel′ A. D.* On equations of the theory of conditional Markov processes // Theory Probab. Appl. 1965. V. 10, N 2. P. 357–360.

4. *Kabanov Yu. M.* A generalized Itô formula for an extended stochastic integral with respect to Poisson random measure // Uspekhi Mat. Nauk. 1974. V. 29, N 4. P. 167–168.

5. *Liptser R. Sh. and Shiryaev A. N.* Martingale Theory [in Russian]. Moscow: Nauka, 1986.

6. *Krylov N. V.* On a proof of Itô's formula // Proc. Steklov Inst. Math. 1994. V. 202. P. 139–142.

7. *Norin N. V.* Itô formula for an extended stochastic integral with nonanticipating kernel // Theory Probab. Appl. 1994. V. 39, N 4. P. 573–592.

8. *Perel′man G. V.* Towards the validity of Itô's formula for discontinuous functions // Theory Probab. Appl. 2011. V. 56, N 3. P. 443–456.

9. *Bismut J.-M.* A generalized formula of Itô and some other properties of stochastic flows // Z. Wahrsch. Verw. Geb. 1981. V. 55. P. 331–350.

10. *Es-Sebaiy K. and Tudor C.* Levy processes and Itô–Skorohod integrals // Theory Stoch. Process. 2008. V. 14, N 2. P. 10–18.

11. *Krylov N. V.* A relatively short proof of Itô's formula for SPDEs and its applications // SPDE Anal. Comp. 2013. N 1. P. 152–174.

12. *Purtukhia O. and Jaoshvili V.* Itô type formula for Poisson anticipating integral // Rep. Enlarged Session Seminar I. Vekua Inst. Appl. Math. 2001. V. 25. P. 103–108.

13. *Doobko V. A.* Questions of the Theory and Application of Stochastic Differential Equations [in Russian]. Vladivostok: DVNTs AN SSSR, 1989.

14. *Rozovskiĭ B. L.* On Itô–Ventsel′ formula // Vestnik NGU Ser. Mat. Mekh. Informat. 1973. N 1. P. 26–32.

15. *Flandoli F. and Russo F.* Generalized integration and stochastic ODEs // Ann. Prob. 2002. V. 30, N 1. P. 270–292.

16. *Krylov N. V.* On the Itô–Ventzel's formula for distribution-valued processes and related topics // Probab. Theory Relat. Fields. 2011. V. 120, N 1–2. P. 295–319.

17. *Ocone D., Pardoux E.* A generalized Itô–Ventzel's formula // Ann. Inst. Henri Poincaré. 1989. V. 25, N 1. P. 39–71.

18. *Purtukhia O.* Itô–Ventsel's formula for antisipative processes // New Trends Probab. Stat. 1991. P. 503–527.

19. *Toronjadze T. and Lazrieva N.* Asymptotic properties of the maximum likelihood estimator, Itô–Ventzel's formula for semimartingales and its application to the recursive estimation in a general scheme of statistical models // Proc. 1st World Congress Bernoulli Soc. (Tashkent, 1986). Utrecht: VNU Sci. Press, 1987. V. 2. P. pp. 63–66.

20. *Doobko V. A.* Open evolving systems // The First International Conference "Open Evolving Systems" (26–27 October 2002). Kiyv: VNZ VMURoL, 2002. P. pp. 14–31.

21. *Øksendal B. and Zhang T.* The Itô–Ventzel's formula and forward stochastic differential equation driven by Poisson random measures // Osaka J. Math. 2007. V. 44. P. 207–230.

22. *Øksendal B., Sulem A., and Zhang T.,* "A stochastic HJB equation for optimal control of forward-backward SDEs," 2013. http:arxiv.org/abs/1312.1472v1.

23. *Karachanskaya E. V.* On one generalization of the Itô–Wentzell formula // Obozrenie Prikl. i Promyshl. Mat. 2011. V. 18, No. 2. P. 494–496.

24. *Karachanskaya E. V.* The generalized Itô–Wentzell formula for noncentered Poisson measure, stochastic first integral and first integral // Mat. Tr. 2014. V. 17, N 1. P. 99–122.

25. *Doobko V. A. and Karachanskaya E. V.,* On Two Approaches for Obtaining of the Generalized Itô–Wentzell Formula [in Russian] [Preprint No. 174], Pacific National University, Khabarovsk (2012).

26. *Gikhman I. I. and Skorokhod A. V.* Stochastic Differential Equations. Berlin: Springer, 1972.

27. *Doobko V. A. and Karachanskaya E. V.* Stochastic first integrals, kernel of integral invariants and Kolmogorov equations // Dal′nevostochn. Mat. Sb. 2014. V. 14, N 2. P. 1–17.

28. *Karachanskaya E. V.* Construction of program control with probability one for a dynamical system with Poisson perturbations // Bulletin of PNU. 2011. N 2. P. 51–60.

November 27, 2014

E. V. Karachanskaya
Pacific National University, Khabarovsk, Russia
EKarachanskaya@mail.khstu.ru

Yakutian Mathematical Journal
July—September, 2014. Vol. 21, No. 3

UDC 512.53

SEMIGROUPS WITH FINITELY APPROXIMATED FINITE ACTS

I. B. Kozhukhov and A. R. Khaliullina

Abstract. We study the semigroups over which all right acts are residually finite. We characterize the groups with this property. We prove that if all acts over a semigroup are approximated by finite ones whose orders are bounded then the semigroup is uniformly locally finite, i.e., there exists a function $f(n)$ such that the order of each n-generated subsemigroup is less than $f(n)$.

Keywords: act over a semigroup, residually finite act, uniformly locally finite algebra

It is well known that the category of acts over a semigroup carries a lot of information on its structure (see [1, Chapter 6] for instance). The semigroups S studied in [2–4] satisfy the conditions:

$(*)$ all right S-acts are residually finite;

$(**)$ all right S-acts are approximated by acts with n or fewer elements.

As [2] shows, all acts over a semigroup S of this kind are approximated by the acts consisting of at most two elements if and only if S is a semilattice (a commutative semigroup of idempotents). The periodicity of the semigroups satisfying $(**)$ is established in [3]. Commutative semigroups and nilsemigroups with $(*)$ and $(**)$ were studied in [3, 4]. This article continues these efforts. We prove that in every semigroup with $(*)$ each right congruence is always the intersection of right congruences of finite indices. For groups this condition turns out sufficient as well, i.e., a group satisfies $(*)$ if and only if its every subgroup is an intersection of finite index subgroups.

Following Pinus [5], we refer to a universal algebra as *uniformly locally finite* whenever there exists a function $h(t)$ such that the orders of all t-generated subalgebras are at most $h(t)$. Similarly, we call this semigroup *uniformly periodic* whenever it satisfies the identity $x^{p+q} = x^p$. The uniform periodicity of semigroups with $(**)$ is proved in [3]; namely, the identity $x^{2n!} = x^{n!}$ holds in these semigroups. Actually, the arguments of [3] enable us to establish the stronger identity $x^{2m} = x^m$, where $m = \operatorname{lcm}(1, 2, \dots, n)$ (the least common multiple). In addition, we can strengthen this result to the uniform local finiteness of semigroups with $(**)$, which follows from a general statement established here: every universal algebra of finite signature approximated by finite algebras of jointly bounded orders is uniformly locally finite.

As for the basic references, see [6] for semigroup theory, [1] for the theory of acts, and [7] for universal algebra. Let us recall some definitions and notation. Given an equivalence ρ on a set X, denote by X/ρ the set of cosets of ρ (the quotient). For $a \in X$, denote by $a\rho$ the coset of ρ containing a. Hence, $X/\rho = \{a\rho \mid a \in X\}$. Furthermore, $\Delta_X = \{(x, x) \mid x \in X\}$ is the identity relation on X. Given two equivalences ρ and σ on a set, $\rho \vee \sigma$ stands for the join of ρ and σ in the lattice of equivalences. It is known that if ρ and σ are congruences of a universal algebra then the join of ρ and σ in the lattice of congruences coincides with their join in the lattice of equivalences.

Say that an algebra A is *approximated* by algebras in a given class $\mathscr{K}$ of universal algebras whenever there exists a set $\{\rho_i \mid i \in I\}$ of congruences of A such that $A/\rho_i \in \mathscr{K}$ for all $i \in I$ and $\bigcap\{\rho_i \mid i \in I\} = \Delta_A$. This is known to be equivalent to A being a subdirect product of algebras in $\mathscr{K}$ (see Corollary 7.2 of [7]).

A *right act* over a semigroup S (or a *right S-act*) is a set X equipped with an action of S; i.e, there is a mapping $X \times S \to X$, $(x, s) \mapsto xs$, satisfying the condition $x(st) = (xs)t$ for all $x \in X$ and $s, t \in S$ (see [1]). We write X_S to emphasize that X is a right S-act. Since we never consider left acts, we write simply "acts" rather than "right acts." The semigroup S itself is a right S-act. Clearly, the congruences of the act S_S are precisely the right congruences of the semigroup S.

Assume that S is a semigroup with unity e. Then an S-act X is called *unitary* whenever $xe = x$ for all $x \in X$. If we can express an S-act X as a disjoint union of subacts X_i, i.e., $X = \bigcup\{X_i \mid i \in I\}$ and $X_i \cap X_j = \varnothing$ for $i \neq j$, then we call X the *coproduct* of the acts X_i and write $X = \coprod\{X_i \mid i \in I\}$.

Given a group G and its subgroup H, not necessarily normal, denote by G/H the set of right cosets Hg. The set G/H is a unitary right G-act with respect to the action $Hg \cdot g' = Hgg'$. It is not difficult to verify that each cyclic unitary G-act is isomorphic to one of the acts G/H.

Given a semigroup S, denote by $S^1 = S \cup \{1\}$ the semigroup that is obtained by adjoining a unit element to S, even if S already has such a unit.

A universal algebra is called *subdirectly indecomposable* whenever it cannot decompose as a nontrivial subdirect product of algebras. The interest in subdirectly indecomposable algebras is explained by Birkhoff's theorem stating that every algebra is a subdirect product of subdirectly indecomposable algebras (see Theorem 7.3 of [7]). In view of this theorem, we can assert that for every semigroup S condition $(*)$ is equivalent to condition

$(*')$ all subdirectly indecomposable S-acts are finite,

while condition $(**)$ is equivalent to condition

$(**')$ every subdirectly indecomposable S-act X satisfies $|X| \leq n$.

The following theorem shows that we can strengthen Theorem 2 of [3] by replacing $n!$ with $\operatorname{lcm}(1, 2, \ldots, n)$. Furthermore, the proof carries over almost verbatim.

Theorem 1. *If a semigroup S satisfies $(**)$ then the identity $x^{2m} = x^m$ holds in S, where $m = \operatorname{lcm}(1, 2, \ldots, n)$.*

Proof. Assume the contrary. Then there exists $a \in S$ with $a^{2m} \neq a^m$. Using Zorn's lemma, we easily verify that there exists a congruence ρ of the act $(S^1)_S$ which is maximal with respect to the condition $(a^{2m}, a^m) \notin \rho$. Standard arguments show that S^1/ρ is a subdirectly indecomposable S-act. Then $(**')$ implies that $|S^1/\rho| \leq n$.

Let us verify that the elements $1, a, a^2, \ldots, a^n$ lie in distinct cosets of ρ. Indeed, if $(a^i, a^j) \in \rho$ for some $i < j \leq n$ then $(a^i, a^{i+k}) \in \rho$ for some $k \leq n$, and so $(a^i, a^{i+kl}) \in \rho$ for all $l \in \mathbb{N}$. Since $m = \operatorname{lcm}(1, 2, \ldots, n)$, it follows that $m = kt$ for some $t \in \mathbb{N}$. Hence, $(a^i, a^{i+m}) \in \rho$. Multiplying by a^{m-i} on the right, we obtain $(a^m, a^{2m}) \in \rho$ which contradicts the choice of a. □

It is well known that each finite index subgroup contains a finite index normal subgroup. Semigroup theory provides a similar statement.

Lemma 2 [2, Lemma 7]. *For every finite index right congruence ρ of a semigroup S, there is a finite index two-sided congruence $\rho' \subseteq \rho$. Furthermore, if $|S/\rho| = n$ then we can choose ρ' so that $|S/\rho'| \leq n^{n+1}$.*

Lemma 3. *If a semigroup S satisfies $(*)$ then every right congruence is an intersection of finite index right congruences.*

Proof. Take a right congruence ρ of a semigroup S. By $(*)$, the act S/ρ is residually finite. Hence, the equality relation $\Delta_{S/\rho}$ is an intersection of finite index congruences of S/ρ, which means that ρ is an intersection of finite index right congruences of S.

The following statement is pointed out in [2]. Let us justify it to make our exposition fuller.

Lemma 4. *If a semigroup S satisfies $(*)$ then all its homomorphic images are residually finite.*

Proof. Take a congruence ρ of S and consider the quotient semigroup $T = S/\rho$ as a right S-act. By $(*)$, the act T is residually finite. Therefore, there exist congruences ρ_i of this act with $|T/\rho_i| < \infty$ and $\bigcap_i \rho_i = \Delta_T$. This implies that S has finite index right congruences σ_i with $\bigcap_i \sigma_i = \rho$. By Lemma 2, there exist finite index two-sided congruences $\sigma_i' \subseteq \sigma_i$. We have

$$\rho \subseteq \bigcap_i (\rho \vee \sigma_i') \subseteq \bigcap_i (\rho \vee \sigma_i) = \bigcap_i \sigma_i = \rho.$$

Hence,

$$\rho = \bigcap_i (\rho \vee \sigma_i').$$

Since the congruences $\rho \vee \sigma_i'$ are of finite index, the semigroup S/ρ is residually finite. □

Below we need a description of subdirectly indecomposable acts over a group. Let us establish firstly a lemma on subdirectly indecomposable acts over a monoid.

Lemma 5. *Given a monoid S with unity e, every subdirectly indecomposable nonunitary S-act X can be expressed as $X = Y \cup \{a\}$, where $Y = Xe$ is a unitary subdirectly indecomposable S-act and $aS \subseteq Y$.*

Proof. Take an act X satisfying the hypotheses. It is obvious that $Y = Xe = XS$ is a unitary subact. To verify that Y is subdirectly indecomposable, assume the contrary. Then there exists a family $\{\rho_i \mid i \in I\}$ of congruences of Y such that $\rho_i \neq \Delta_X$ and $\bigcap\{\rho_i \mid i \in I\} = \Delta_X$. Put $\rho_i' = \rho_i \cup \Delta_X$. Then ρ_i' is a congruence of X and

$$\bigcap\{\rho_i' \mid i \in I\} = \Delta_X.$$

But this contradicts the indecomposability of X as a subdirect product. It remains to show that $|X \backslash Y| = 1$.

If $a, b \in X \backslash Y$ with $a \neq b$ then $ae \neq a$ and $be \neq b$. It is obvious that the relations $\rho_a = \{(a, ae), (ae, a)\} \cup \Delta_X$ and $\rho_b = \{(b, be), (be, b)\} \cup \Delta_X$ are congruences of X and $\rho_a \cap \rho_b = \Delta_X$. This contradicts the assumption that X is subdirectly indecomposable. □

Given an act X over a group G with unity e, we have $X = Xe \cup (X \backslash Xe)$, where Xe is the maximal unitary subact, while $A = X \backslash Xe$ is the nonunitary part.

The unitary subact Xe is the union of disjoint orbits, $Xe = \coprod_{i\in I} x_iG$; furthermore, $x_iG \cong G/H_i$, where $H_i = \{g \mid x_ig = x_i\}$. Thus,

$$X \cong \coprod_{i\in I}(G/H_i) \coprod A,$$

where $AG \subseteq Xe$.

Theorem 6. *A right act X over a group G is subdirectly indecomposable if and only if $|X| = 1$ or X is isomorphic to one of the acts:*
(1) $\{x, a\}$, where $xG = aG = \{x\}$;
(2) G/H and, furthermore, G includes the smallest subgroup $H' \supset H$;
(3) $(G/H)\coprod\{z\}$, where H is as in (2) and $zG = \{z\}$.

Proof. We can verify directly that the acts of the form (1)–(3) are subdirectly indecomposable.

Take a subdirectly indecomposable G-act X with $|X| \geq 2$. Assume firstly that X is unitary. Then

$$X = \bigcup\{x_iG \mid i \in I\},$$

where $x_iG \cong G/H_i$ are the orbits. To show that $|I| \leq 2$, suppose that $|I| \geq 3$ and choose distinct elements $i_1, i_2, i_3 \in I$. The sets

$$P = x_{i_1}G, \quad Q = x_{i_2}G, \quad R = \bigcup\{x_iG \mid i \neq i_1, i_2\}$$

are nonempty. Put

$$\rho_1 = ((P \cup Q) \times (P \cup Q)) \cup \Delta_X,$$
$$\rho_2 = ((R \cup Q) \times (R \cup Q)) \cup \Delta_X,$$
$$\rho_3 = ((P \cup R) \times (P \cup R)) \cup \Delta_X.$$

Then ρ_1, ρ_2, and ρ_3 are nontrivial congruences of X and $\rho_1 \cap \rho_2 \cap \rho_3 = \Delta_X$, which is impossible for a subdirectly indecomposable act.

Therefore, we see that $X = xG$ or $X = x_1G \cup x_2G$. If $X = xG$ then $X \cong G/H$ and, since every nontrivial subdirectly indecomposable act must contain the smallest nontrivial congruence, the smallest subgroup $H' \supset H$ exists. This means that X is of the form (2). Assume now that $X = x_1G \cup x_2G$. If $|x_1G|, |x_2G| \geq 2$ then $\rho_1 = (x_1G \times x_1G) \cup \Delta_X$ and $\rho_2 = (x_2G \times x_2G) \cup \Delta_X$ is a nontrivial congruence with $\rho_1 \cap \rho_2 = \Delta_X$; this is contradiction. Therefore, we can assume that $|x_2G| = 1$. Consequently,

$$X = x_1G \cup \{z\} \cong (G/H)\coprod\{z\}.$$

It is not difficult to see that we obtain a act of the form (3).

If X is nonunitary then $X = Xe \cup \{a\}$ by Lemma 5. To verify that $|Xe| = 1$, suppose that $|Xe| > 1$. Then

$$\rho_1 = \{(a, ae), (ae, a)\} \cup \Delta_X, \quad \rho_2 = (Xe \times Xe) \cup \Delta_X$$

are nontrivial congruences of X and $\rho_1 \cap \rho_2 = \Delta_X$, and we also arrive at a contradiction. If $|Xe| = 1$ then X is of the form (1). □

All homomorphic images of every infinite cyclic semigroup S are residually finite (in fact, the nontrivial homomorphic images are finite). But some example shows [2] that an infinite subdirectly indecomposable act over S exists. Consequently, the converse to Lemma 4 is false. The example shows also that the converse to Lemma 3 is false. But in the case that S is a group, Lemma 3 admits a converse.

Theorem 7. *A group G satisfies condition $(*)$ if and only if each of its subgroups is an intersection of finite index subgroups.*

Proof. *Necessity* follows from Lemma 3.

Sufficiency. Since $(*)$ and $(*')$ are equivalent conditions, it suffices to prove $(*')$.

Take a subdirectly indecomposable G-act X, which is of the form (1), (2), or (3) by Theorem 5. Acts of the form (1) are finite. Let us show that every act of the form (2) is finite as well. Hence, $X \cong G/H$ and there the smallest subgroup $H' \supset H$ exists. By assumption, H is an intersection of finite index subgroups, i.e., $H = \bigcap_\alpha H_\alpha$ and $[G : H_\alpha] < \infty$ for all α. In addition, either $H_\alpha \supseteq H'$ or $H_\alpha = H$ for each α. Consequently, $H_\alpha = H$ for some α, and so H is of finite index. Therefore, X is a finite act. For X of the form (3) we establish in the same fashion that H is a finite index subgroup, and so $X \cong (G/H) \cup \{z\}$ is a finite act. □

Proceed to condition $(**)$. To show that the semigroups satisfying it are uniformly locally finite, we need one general statement. The next theorem follows from Theorem 0.1 of [8]. We include a proof because in our case we can sharpen the upper bound on the orders of t-generator subalgebras.

Theorem 8. *Consider a universal algebra A of finite signature. If A is approximated by some algebras A_i, $i \in I$, of finite jointly bounded orders then A is uniformly locally finite. Furthermore, if $|A_i| \leq n$ for all $i \in I$ then each subalgebra generated by t elements has at most $\exp(\psi(n) \cdot n^t \cdot \log n)$ elements, where $\psi(n)$ is the number of nonisomorphic algebras of this signature with orders at most n.*

Proof. Enumerate the elements of the signature of A as $\Sigma = \{f_1, \dots, f_k\}$. By assumption, A is a subdirect product of some algebras A_i of signature Σ with $|A_i| \leq n$. Consequently, A is isomorphic to a subalgebra of the direct product $\bar{A} = \prod_{i \in I} A_i$. Therefore, it suffices to prove that $\bar{A}$ is a uniformly locally finite algebra.

Since Σ is a finite signature, there is only a finite number $\psi(n)$ of pairwise nonisomorphic algebras with n or fewer elements. Hence,

$$\bar{A} \cong B_1 \times B_2 \times \cdots \times B_{\psi(n)}, \tag{$\circ$}$$

where $B_\alpha = \prod_{I_\alpha}(C_\alpha)$ and C_α is an algebra of cardinality at most n, while $\alpha = 1, 2, \dots, \psi(n)$. Since the direct product of finitely many uniformly locally finite algebras is a uniformly locally finite algebra, it suffices to show that all B_α are uniformly locally finite algebras. To simplify notation, we omit the indices.

Thus, we have to show that $B = \prod_{j \in J} C$ is a uniformly locally finite algebra provided that C is a finite algebra of finite signature.

Choose $b^{(1)}, b^{(2)}, \dots, b^{(t)} \in B$ and consider the subalgebra $\langle b^{(1)}, b^{(2)}, \dots, b^{(t)} \rangle$ generated by these elements. We have $b^{(s)} = (b_{sj})_{j \in J}$ for $s = 1, 2, \dots, t$, where $b_{sj} \in C$. For each tuple $\xi = (c_1, \dots, c_t) \in C^t$, put

$$J_\xi = \{j \in J \mid b_{1j} = c_1, \dots, b_{tj} = c_t\}.$$

Some of these sets may be empty. It is obvious that $J_\xi \cap J_\eta = \varnothing$ for $\xi \neq \eta$. Consequently, the nonempty J_ξ form a partition of J, i.e., $J = \bigcup\{J_\xi \mid \xi \in C^t\}$. Denote the equivalence corresponding to this partition by θ.

Put $B' = \{(c_j)_{j \in J}$ for $(j, j') \in \theta\}$. It is obviously a finite subalgebra and its order equals $|C|^p$, where $p = |J/\theta|$. Moreover, $b^{(1)}, b^{(2)}, \dots, b^{(t)} \in B'$. Consequently, the subalgebra $\langle b^{(1)}, b^{(2)}, \dots, b^{(t)} \rangle$ is finite. Let us find an upper bound on its order. Since $p \leq |C^t| \leq n^t$, we see that

$$|\langle b^{(1)}, b^{(2)}, \dots, b^{(t)} \rangle| \leq |B'| = |C|^p \leq \exp(n^t \log n).$$

So, in each B_α the subalgebra generated by t elements contains at most $\exp(n^t \log n)$ elements. Using ($\circ$), we infer that every t-generated subalgebra of A has order at most $\exp(\psi(n) \cdot n^t \cdot \log n)$. $\square$

The theorem above enables us to establish the uniform local finiteness of some class of semigroups that contains the semigroups satisfying ($**$).

Theorem 9. *Every semigroup S such that the act S_S is approximated by S-acts whose orders are bounded above by the same positive integers is uniformly locally finite.*

Proof. By assumption, $\Delta_S = \bigcap\{\rho_i \mid i \in I\}$, where ρ_i is a right congruence such that S/ρ_i is a subdirectly indecomposable act and $|S/\rho_i| \leq n$. Lemma 2 implies that ρ_i contains a two-sided congruence ρ_i' of index at most n^{n+1}. Therefore, $\Delta_S = \bigcap\{\rho_i' \mid i \in I\}$, where $|S/\rho_i'| \leq n^{n+1}$. Since ρ_i' is a two-sided congruence, it follows that S/ρ_i' is a semigroup. Consequently, the semigroup S is approximated by semigroups with n^{n+1} or fewer elements. By Theorem 8, it is uniformly locally finite. The order of each t-generated subsemigroup is at most $\exp(\psi(n^{n+1}) \cdot n^{t(n+1)} \cdot (n+1) \cdot \log n)$, where $\psi(k)$ is the number of nonisomorphic semigroups with k or fewer elements. $\square$

Corollary 10. *Every semigroup with condition* ($**$) *is uniformly locally finite.*

Remark. If a semigroup is uniformly locally finite then it is uniformly periodic. But the bound on the order of a cyclic subsemigroup provided by Theorem 9 is weaker than that in Theorem 1. Indeed, taking $t = 1$ in the formula of Theorem 9, we obtain the inequality $|\langle a\rangle| \leq \exp(\psi(n^{n+1}) \cdot n^{n+1} \cdot (n+1) \cdot \log n)$ for every element a of this semigroup, while Theorem 1 yields $|\langle a\rangle| \leq 2 \cdot \mathrm{lcm}(1, 2, \dots, n) - 1$.

REFERENCES

1. *Kilp M., Knauer U., and Mikhalev A. V.* Monoids, Acts and Categories. New York and Berlin: W. de Gruyter, 2000.
2. *Kozhukhov I. B.* One characteristical property of semilattices // Comm. Algebra. 1997. V. 25, N 8. P. 2569–2577.
3. *Kozhukhov I. B.* Finiteness conditions for subdirectly irreducible S-acts and modules // Fundam. Prikl. Mat. 1998. V. 4, N 2. P. 763–767.
4. *Kozhukhov I. B.* Semigroups over which all acts are residually finite // Fundam. Prikl. Mat. 1998. V. 4, N 4. P. 1335–1344.
5. *Pinus A. G.* Inner homomorphisms and positive-conditional terms // Algebra and Logic. 2001. V. 40, N 2. P. 87–95.
6. *Clifford A. H. and Preston G. B.* The Algebraic Theory of Semigroups. Vol. 1. Providence: Amer. Math. Soc., 1972.
7. *Cohn P.* Universal Algebra. New York, Evanston, and London: Harper and Row, Publishers, 1965.
8. *Hobby D. and McKenzie R.* The Structure of Finite Algebras. Providence: Amer. Math. Soc., 1993.

November 29, 2014

I. B. Kozhukhov; A. R. Khaliullina
National Research University of Electronic Technology
Moscow, Russia
kozhuhov_i_b@mail.ru; haliullinaar@gmail.com

Yakutian Mathematical Journal
July—September, 2014. Vol. 21, No. 3

UDC 539.311

THE EQUILIBRIUM PROBLEM FOR A TIMOSHENKO PLATE IN CONTACT WITH AN OBLIQUE OBSTACLE

N. P. Lazarev

Abstract. Under study is the equilibrium of a plate in contact with a fixed rigid obstacle on a part of the outer edge. On the contact boundary we propose a condition that is described the no-penetration of the points of the plate and obstacle, on assuming that the normal to the surface of a possible contact of the plate with the obstacle is at a small angle to the midplane of the plate. We prove that the variational equilibrium problem of the plate with Signorini-type conditions has a unique solution. We find a differential statement equivalent to the original statement in case of sufficiently smooth solutions.

Keywords: plate, crack, no-penetration condition, variational problem

Introduction

Many articles study plates and shells (see [1–13] for instance). As a rule, the edges of the plate are vertical and defined by a cylindrical surface, whose normal at each point is parallel to the midplane of the plate. Variational methods are applicable in many model problems of the mechanics of deformable solids with nonlinear no-penetration conditions (in the form of a system of equalities and inequalities) defined on the curve (surface) corresponding to the zone of a possible contact or a crack in the solid [6–13].

Nonlinear equilibrium problems of a plate in the Kirchhoff–Liav model with the no-penetration conditions for an oblique crack were studied in [7, 8]. The no-penetration condition in [7] is given for a crack specified as a smooth surface $z = F(x_1, x_2)$, where $F(x_1, x_2)$ is a function on the (mid)plane (x_1, x_2). A no-penetration condition for an almost vertical crack is derived in [8]. Both articles rely on the additional assumption (as compared to the Kirchhoff–Liav model) that we can determine the displacements at all points of the crack faces by using the displacements of the points on the midplane of the plate. The equilibrium problem for the Timoshenko plate with a nonlinear no-penetration boundary condition modeling the mutual no-penetration of the opposite faces of an oblique crack is studied in [12]; furthermore, the no-penetration condition, by analogy with [8], is derived on assuming that the obliqueness of the crack surface is small.

In this article we propose a mathematical model with a nonlinear no-penetration condition for the contact equilibrium problem of a plate. We derive the boundary condition on assuming that the normal to the contact surface is at a small angle to the midplane. We state the variational equilibrium problem of a plate as the minimization problem for an energy functional on a set of admissible functions satisfying

The author was supported by the Russian Foundation for Basic Research (Grant 13–01–00017A).

the no-penetration condition and show that this problem has the unique solution. We describe the contact of the plate with a fixed rigid obstacle. When the solution to the minimization problem is sufficiently smooth, for the original variational statement we obtain an equivalent statement as a boundary problem.

1. Statement of the Problem

Consider a bounded simply-connected region $\Omega \subset \mathbb{R}^2$ with smooth boundary $\Gamma = \gamma \cup \Gamma_0$, where $\gamma \cap \Gamma_0 = \varnothing$ and $\operatorname{mes} \Gamma_0 > 0$ (Fig. 1). Assume that the curve γ does not include its endpoints. Denote by $\boldsymbol{\nu} = (\nu_1, \nu_2)$ the outer normal to Γ. Define the three-dimensional Cartesian coordinates $\{x_1, x_2, z\}$ so that the set $\Omega \times \{0\} \subset \mathbb{R}^3$ corresponds to the midplane of the plate.

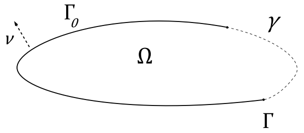


Fig. 1

As in [8, 12], define the surface

$$\Xi = \big\{(\bar{x}_1, \bar{x}_2, z) \mid |z| \le h,\ (\bar{x}_1, \bar{x}_2) = \bar{\mathbf{x}},\ \bar{\mathbf{x}} = \mathbf{x} - z\boldsymbol{\nu}(\mathbf{x}) \tan \alpha(\mathbf{x}),\ \mathbf{x} \in \gamma\big\}, \tag{1}$$

where $\alpha(\mathbf{x}) \in C(\gamma)$ with $|\alpha(\mathbf{x})| < \pi/2$ for $\mathbf{x} \in \gamma$. Assume that the normal $\boldsymbol{n}(\bar{\mathbf{x}}, z)$ to Ξ for fixed $\mathbf{x} \in \gamma$ remains constant, i.e.,

$$\boldsymbol{n}(\bar{\mathbf{x}}, z) = \boldsymbol{n}(\mathbf{x}, 0) = (\boldsymbol{\nu}(\mathbf{x}) \cos \alpha(\mathbf{x}), \sin \alpha(\mathbf{x})), \quad \bar{\mathbf{x}} = \mathbf{x} - z\boldsymbol{\nu}(\mathbf{x}) \tan \alpha(\mathbf{x}), \quad |z| \le h.$$

For example, the surfaces formed using the plane or a conical surface enjoy this property.

In the original undeformed state the plate touches a rigid obstacle along some surface Σ (Fig. 2) and is clamped along the rest of its outer edge.

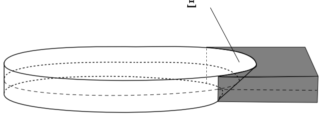


Fig. 2

Denote by $\boldsymbol{\chi} = \boldsymbol{\chi}(\mathbf{x}) = (\boldsymbol{W}, w)$, for $\mathbf{x} \in \Omega$, the displacement vector of the points of the middle surface, where $\boldsymbol{W} = (w_1, w_2)$ and w are horizontal (along the plane (x_1, x_2)) and vertical displacements respectively. Denote the rotation angles of the normal sections by $\boldsymbol{\psi} = \boldsymbol{\psi}(\mathbf{x}) = (\psi_1, \psi_2)$ for $\mathbf{x} \in \Omega$.

Recall some tensor relations of elasticity theory, valid for transversally isotropic elastic plates [14]. The tensors

$$\varepsilon_{ij}(\boldsymbol{\psi}) = \frac{1}{2}\left(\frac{\partial \psi_i}{\partial x_j} + \frac{\partial \psi_j}{\partial x_i}\right), \quad \varepsilon_{ij}(\boldsymbol{W}) = \frac{1}{2}\left(\frac{\partial w_i}{\partial x_j} + \frac{\partial w_j}{\partial x_i}\right), \quad i, j = 1, 2,$$

describe the deformation of the plate. The momentum and stress tensors are calculated as

$$m_{ij}(\boldsymbol{\psi}) = c_{ijkl}\varepsilon_{kl}(\boldsymbol{\psi}), \quad \sigma_{ij}(\boldsymbol{W}) = 3h^{-2}c_{ijkl}\varepsilon_{kl}(\boldsymbol{W}) \tag{2}$$

(with summation over the repeated indices); the nonzero constant components of the tensor c_{ijkl} are defined as

$$c_{iiii} = D, \quad c_{iijj} = D\varkappa, \quad c_{ijij} = c_{ijji} = D(1-\varkappa)/2, \quad i,j = 1,2,\ i \neq j,$$

where D is the cylindrical rigidity of the plate and $\varkappa$ is Poisson's coefficient. The vector of transverse forces $\boldsymbol{q} = (q_1, q_2)$ satisfies [14]

$$q_i(w, \boldsymbol{\psi}) = \Lambda(w_{,i} + \psi_i), \quad i = 1,2, \quad \left(v_{,i} = \frac{\partial v}{\partial x_i}\right) \tag{3}$$

with $\Lambda = 2\kappa' G$, where κ' is the shear coefficient, G is the shear modulus in the area elements orthogonal to the midplane of the plate; furthermore, Λ, κ', and G are constant.

Denote by $H^1(\Omega)$ the Sobolev space and by $H^{1,0}_{\Gamma_0}(\Omega)$ the subspace of $H^1(\Omega)$ of all functions vanishing on a part Γ_0 of the outer boundary. Put

$$H = H^{1,0}_{\Gamma_0}(\Omega)^5, \quad \|\cdot\| = \|\cdot\|_H.$$

For $\boldsymbol{\eta} = (\boldsymbol{W}, w, \boldsymbol{\psi}) \in H$ and $\bar{\boldsymbol{\eta}} = (\overline{\boldsymbol{W}}, \bar{w}, \bar{\boldsymbol{\psi}}) \in H$, define the bilinear form

$$B(\boldsymbol{\eta}, \bar{\boldsymbol{\eta}}) = \int\limits_{\Omega} \left(\sigma_{ij}(\boldsymbol{W})\varepsilon_{ij}(\overline{\boldsymbol{W}}) + m_{ij}(\boldsymbol{\psi})\,\varepsilon_{ij}(\bar{\boldsymbol{\psi}}) + \Lambda(w_{,i} + \psi_i)(\bar{w}_{,i} + \bar{\psi}_i)\right).$$

The potential energy of the deformed plate is

$$\Pi(\boldsymbol{\eta}) = \frac{1}{2}B(\boldsymbol{\eta}, \boldsymbol{\eta}) - \int\limits_{\Omega} \boldsymbol{F}\boldsymbol{\eta}, \quad \eta = (\boldsymbol{W}, w, \boldsymbol{\psi}) \in H, \tag{4}$$

where $\boldsymbol{F} = (f_1, f_2, f_3, \mu_1, \mu_2) \in L^2(\Omega)^5$ is the prescribed vector of exterior loads [14].

Impose the boundary conditions of rigid clamping on the part Γ_0 of the outer boundary Γ:

$$\boldsymbol{\eta} = \mathbf{0} \quad \text{on } \Gamma_0, \quad \text{where } \mathbf{0} = (0,0,0,0,0),\ \boldsymbol{\eta} = (\boldsymbol{W}, w, \boldsymbol{\psi}). \tag{5}$$

In the Timoshenko model we can express the displacements

$$\boldsymbol{\chi}(\mathbf{x}, z) = (\boldsymbol{W}(\mathbf{x}, z), w(\mathbf{x}, z))$$

of the points of the plate at distance $|z| \leq h$ from the middle surface in terms of the displacements $\boldsymbol{\chi}(\mathbf{x}, 0) = \boldsymbol{\chi}(\mathbf{x}) = (\boldsymbol{W}, w)$ in the middle surface and the rotation angles $\boldsymbol{\psi} = \boldsymbol{\psi}(\mathbf{x})$ of the normal sections. Furthermore [14],

$$\boldsymbol{W}(\mathbf{x}, z) = \boldsymbol{W}(\mathbf{x}) + z\boldsymbol{\psi}(\mathbf{x}), \quad w(\mathbf{x}, z) = w(\mathbf{x}), \quad |z| \leq h,\ \mathbf{x} \in \Omega.$$

We derive the no-penetration condition by analogy with [8, 12]. Assume that the angle $\alpha(\mathbf{x})$ is sufficiently small for all $\mathbf{x} \in \gamma$ and that we can express the displacements at the points $(\bar{\mathbf{x}}, z) \in \Xi$ (on the possible contact surface) using the traces of the functions $\boldsymbol{W}(\mathbf{x})$, $w(\mathbf{x})$, and $\boldsymbol{\psi}(\mathbf{x})$ along the curve γ the equalities

$$w^{\pm}(\bar{\mathbf{x}}, z) = w^{\pm}(\mathbf{x}), \quad \boldsymbol{W}^{\pm}(\bar{\mathbf{x}}, z) = \boldsymbol{W}^{\pm}(\mathbf{x}) + z\boldsymbol{\psi}^{\pm}(\mathbf{x}), \quad |z| \leq h,\ \mathbf{x} \in \gamma, \tag{6}$$

where $\bar{\mathbf{x}} = \mathbf{x} - z\boldsymbol{\nu}(\mathbf{x}) \tan \alpha(\mathbf{x})$.

The no-penetration condition amounts to requiring the projection of the displacement vector $\boldsymbol{\chi}(\bar{\mathbf{x}}, z)$ (for $x \in \Xi$) onto the normal $\boldsymbol{n}(\bar{\mathbf{x}}, z)$ to be nonpositive:

$$\boldsymbol{\chi}(\bar{\mathbf{x}}, z) \cdot (\boldsymbol{\nu}(\mathbf{x}) \cos\alpha(\mathbf{x}), \sin\alpha(\mathbf{x})) \leq 0, \quad (\bar{\mathbf{x}}, z) \in \Xi.$$

By (6), inserting the extremal values $z = h$ and $z = -h$, we find that

$$\boldsymbol{W}_{\boldsymbol{\nu}} \cos\alpha + h|\boldsymbol{\psi}_{\boldsymbol{\nu}}| \cos\alpha + w \sin\alpha \leq 0, \quad \mathbf{x} \in \gamma,$$

where $\boldsymbol{\psi}_{\boldsymbol{\nu}} = \psi_i \nu_i$ and $\boldsymbol{W}_{\boldsymbol{\nu}} = w_i \nu_i$. Dividing the last relation by $\cos\alpha$, we deduce the no-penetration condition for the oblique crack:

$$\boldsymbol{W}_{\boldsymbol{\nu}} + w \tan\alpha \leq -h|\boldsymbol{\psi}_{\boldsymbol{\nu}}|, \quad \mathbf{x} \in \gamma. \tag{7}$$

Observe that this inequality with $\alpha \equiv 0$ recovers the no-penetration condition for the contact problem with a vertical edge [13].

Let us state the equilibrium problem of a plate in contact with an oblique rigid obstacle as the minimization

$$\inf_{\boldsymbol{\eta} \in K} \Pi(\boldsymbol{\eta}) \tag{8}$$

of the energy functional, where

$$K = \{\boldsymbol{\eta} \in H \mid \boldsymbol{\eta} = (\boldsymbol{W}, w, \boldsymbol{\psi}) \text{ satisfies (7)}\}$$

is the set of admissible functions. The functional $\Pi(\boldsymbol{\eta})$ on the space H is coercive, convex, and weakly lower-semicontinuous [13]. In addition, it is $\Pi(\boldsymbol{\eta})$ differentiable. We can show that K is a convex and closed set; consequently, it is weakly closed in the reflexive space H. These properties of $\Pi(\boldsymbol{\eta})$ and K ensure the existence of a unique solution $\boldsymbol{\xi} = (\boldsymbol{U}, u, \boldsymbol{\phi})$ satisfying the variational inequality

$$B(\boldsymbol{\xi}, \boldsymbol{\eta} - \boldsymbol{\xi}) \geq \int_\Omega \mathbf{F}(\boldsymbol{\eta} - \boldsymbol{\xi}), \quad \boldsymbol{\xi}, \boldsymbol{\eta} \in K. \tag{9}$$

By the properties of the energy functional and the set of admissible functions, the variational inequality (9) and the minimization problem (8) are equivalent [13].

2. Formulation as a Boundary Value Problem

In this section we obtain a formally equivalent differential statement of problem (8). To this end, starting from the variational inequality (9) and choosing suitable test functions, we deduce a complete collection of boundary conditions along the curve γ. In order to extract from (9) a relation on γ, we use Green's formula. Assume that the solution $\boldsymbol{\xi}$ to (8) is sufficiently smooth.

Compare two inequalities obtained by inserting the test functions $\boldsymbol{\eta} = \boldsymbol{\xi} + \tilde{\boldsymbol{\eta}}$ and $\boldsymbol{\eta} = \boldsymbol{\xi} - \tilde{\boldsymbol{\eta}}$ into (9), where $\tilde{\boldsymbol{\eta}} = (\widetilde{\boldsymbol{W}}, \widetilde{w}, \widetilde{\boldsymbol{\psi}}) \in C_0^\infty(\Omega)^5$. This yields

$$\int_\Omega (\sigma_{ij} \varepsilon_{ij}(\widetilde{\boldsymbol{W}}) + m_{ij} \varepsilon_{ij}(\widetilde{\boldsymbol{\psi}}) + q_i(\widetilde{w},_i + \widetilde{\psi}_i))$$

$$= \int_\Omega (f_i \widetilde{w}_i + f_3 \widetilde{w} + \mu_i \widetilde{\psi}_i), \quad \tilde{\boldsymbol{\eta}} \in C_0^\infty(\Omega)^5.$$

Here and henceforth

$$m_{ij} = m_{ij}(\boldsymbol{\phi}), \quad \sigma_{ij} = \sigma_{ij}(\boldsymbol{U}), \quad q_i = q_i(u, \boldsymbol{\phi}), \quad i, j = 1, 2.$$

Since $\widetilde w_1$, $\widetilde w_2$, $\widetilde w$, $\widetilde\psi_1$, and $\widetilde\psi_2$ are independent, we deduce the equilibrium equations

$$\sigma_{ij,j} = -f_i, \quad m_{ij,j} - q_i = -\mu_i, \quad i = 1,2,\ q_{i,i} = -f_3 \quad \text{in } \Omega. \tag{10}$$

Green's formula [9] says that

$$\int\limits_\Omega \sigma_{ij}\,\varepsilon_{ij}(\boldsymbol W) = -\int\limits_\Omega \sigma_{ij,j}w_i + \int\limits_\gamma (\sigma_{\boldsymbol\nu}\boldsymbol W_{\boldsymbol\nu} + \sigma_{\boldsymbol\tau i}w_{\boldsymbol\tau i}), \quad \boldsymbol W \in H^{1,0}_{\Gamma_0}(\Omega)^2, \tag{11}$$

where $\sigma_{\boldsymbol\nu}\boldsymbol\nu$ and $\sigma_{\boldsymbol\tau} = (\sigma_{\boldsymbol\tau 1}, \sigma_{\boldsymbol\tau 2})$ are the normal and tangential components of the vectors ($\sigma_{1j}\nu_j$ and $\sigma_{2j}\nu_j$), $\sigma_{ij}\nu_j = \sigma_{\boldsymbol\nu}\nu_i + \sigma_{\boldsymbol\tau i}$, and $\sigma_{\boldsymbol\nu} = \sigma_{ij}\nu_j\nu_i$, while $\boldsymbol\tau = (-\nu_2, \nu_1)$ and $w_{\boldsymbol\tau i} = w_i - \boldsymbol W_{\boldsymbol\nu}\nu_i$ for $i = 1,2$. Recall also (see [9]) that

$$\int\limits_\Omega m_{ij}\varepsilon_{ij}(\boldsymbol\psi) = -\int\limits_\Omega m_{ij,j}\psi_i + \int\limits_\gamma (m_{\boldsymbol\nu}\boldsymbol\psi_{\boldsymbol\nu} + m_{\boldsymbol\tau i}\psi_{\boldsymbol\tau i}), \quad \boldsymbol\psi \in H^{1,0}_{\Gamma_0}(\Omega)^2, \tag{12}$$

$$\int\limits_\Omega \nabla u \nabla w = \int\limits_\gamma \frac{\partial u}{\partial\boldsymbol\nu}w - \int\limits_\Omega w\Delta u, \quad w \in H^{1,0}_{\Gamma_0}(\Omega), \tag{13}$$

$$\int\limits_\Omega \boldsymbol\phi\nabla w = \int\limits_\gamma \boldsymbol\phi_{\boldsymbol\nu}w - \int\limits_\Omega w\,\mathrm{div}\,\boldsymbol\phi, \quad w \in H^{1,0}_{\Gamma_0}(\Omega). \tag{14}$$

where $m_{\boldsymbol\nu}$ and $m_{\boldsymbol\tau i}$ for $i = 1,2$ are defined by analogy with the previous formula written down for $\sigma_{\boldsymbol\nu}$ and $\sigma_{\boldsymbol\tau i}$ for $i = 1,2$ (see (11)).

Inserting $\boldsymbol\eta = 0$ and $\boldsymbol\eta = 2\boldsymbol\xi$ into (9), we infer that

$$B(\boldsymbol\xi,\boldsymbol\xi) = \int\limits_\Omega \mathbf F\boldsymbol\xi, \quad B(\boldsymbol\xi,\boldsymbol\eta) \ge \int\limits_\Omega \mathbf F\boldsymbol\eta, \quad \boldsymbol\eta \in K. \tag{15}$$

In the second relation here we apply (11)–(14) to integrate by parts and, taking the equilibrium equations (10) into account, obtain

$$\int\limits_\gamma (\sigma_{\boldsymbol\nu}\boldsymbol W_{\boldsymbol\nu} + \sigma_{\boldsymbol\tau i}w_{\boldsymbol\tau i}) + (m_{\boldsymbol\nu}\boldsymbol\psi_{\boldsymbol\nu} + m_{\boldsymbol\tau i}\psi_{\boldsymbol\tau i}) + \Lambda\left(\frac{\partial u}{\partial\boldsymbol\nu} + \boldsymbol\phi_{\boldsymbol\nu}\right)w \ge 0, \quad \boldsymbol\eta \in K. \tag{16}$$

Choose test functions so that $\boldsymbol W_{\boldsymbol\nu} = 0$, $\boldsymbol\psi_{\boldsymbol\nu} = 0$, and $w = 0$ on γ. Then, varying the values of $w_{\boldsymbol\tau i}$ and $\psi_{\boldsymbol\tau i}$ for $i = 1,2$ in (16) arbitrarily, we obtain

$$\sigma_{\boldsymbol\tau i} = m_{\boldsymbol\tau i} = 0,\ i = 1,2, \quad \text{on } \gamma. \tag{17}$$

Introduce the auxiliary vector function $\boldsymbol p = (p_1, p_2, p_3)$ consisting of sufficiently smooth functions p_i, for $i = 1,2,3$, on γ with $\operatorname{supp} p_i \subset \gamma$ for $i = 1,2,3$. It is known (see [6] for instance) that there exists a function $\tilde{\boldsymbol\eta} \in H(\Omega)$ such that

$$(p_1\nu_1, p_1\nu_2) = [\widetilde{\boldsymbol W}], \quad p_2 = [\widetilde w], \quad (p_3\nu_1, p_3\nu_2) = [\widetilde{\boldsymbol\psi}] \quad \text{on } \gamma. \tag{18}$$

Furthermore, it is obvious that $[\widetilde{\boldsymbol W}_{\boldsymbol\nu}] = p_1$ and $[\widetilde{\boldsymbol\psi}_{\boldsymbol\nu}] = p_3$ on γ; hence, on assuming that $p_1 + p_2\tan\alpha \le -h|p_3|$ on γ, we insert into (16) a function $\tilde{\boldsymbol\eta} \in H(\Omega)$, satisfying (18), and then by (17) see that

$$\int\limits_\gamma (\sigma_{\boldsymbol\nu}p_1 + m_{\boldsymbol\nu}p_3 + \boldsymbol q_{\boldsymbol\nu}p_2) \ge 0. \tag{19}$$

Consider the expansion

$$\sigma_{\boldsymbol{\nu}} p_1 + m_{\boldsymbol{\nu}} p_3 + \boldsymbol{q}_{\boldsymbol{\nu}} p_2 = \frac{1}{2}\left(\sigma_{\boldsymbol{\nu}} + \frac{1}{h} m_{\boldsymbol{\nu}}\right)(p_1 + p_2 \tan\alpha + h p_3)$$

$$+\frac{1}{2}\left(\sigma_{\boldsymbol{\nu}} - \frac{1}{h} m_{\boldsymbol{\nu}}\right)(p_1 + p_2 \tan\alpha - h p_3) + (-\sigma_{\boldsymbol{\nu}} \tan\alpha + \boldsymbol{q}_{\boldsymbol{\nu}}) p_2. \tag{20}$$

Choose p_1, p_2, and p_3 satisfying $p_1 = -p_2 \tan\alpha$, $p_3 = 0$ from (19). By (20), we find that

$$\sigma_{\boldsymbol{\nu}} \tan\alpha = \boldsymbol{q}_{\boldsymbol{\nu}} \quad \text{on } \gamma. \tag{21}$$

Using (20) and (21), rearrange (19) as

$$\int\limits_{\gamma} \left(\left(\sigma_{\boldsymbol{\nu}} + \frac{1}{h} m_{\boldsymbol{\nu}}\right)(p_1 + p_2 \tan\alpha + h p_3) + \left(\sigma_{\boldsymbol{\nu}} - \frac{1}{h} m_{\boldsymbol{\nu}}\right)(p_1 + p_2 \tan\alpha - h p_3)\right) \geq 0 \tag{22}$$

provided that $p_1 + p_2 \tan\alpha \leq -h|p_3|$ on γ. This implies that $h\sigma_{\boldsymbol{\nu}} + m_{\boldsymbol{\nu}} \leq 0$ and $h\sigma_{\boldsymbol{\nu}} - m_{\boldsymbol{\nu}} \leq 0$, or ? $- h\sigma_{\boldsymbol{\nu}} \geq |m_{\boldsymbol{\nu}}|$ on γ. ?!

By (17) and (21), integrating by parts the first relation in (15) yields

$$\int\limits_{\gamma} \left(\left(\sigma_{\boldsymbol{\nu}} + \frac{1}{h} m_{\boldsymbol{\nu}}\right)([\boldsymbol{U}_{\boldsymbol{\nu}}] + [u] \tan\alpha + h[\boldsymbol{\phi}_{\boldsymbol{\nu}}])\right.$$

$$\left. + \left(\sigma_{\boldsymbol{\nu}} - \frac{1}{h} m_{\boldsymbol{\nu}}\right)([\boldsymbol{U}_{\boldsymbol{\nu}}] + [u] \tan\alpha - h[\boldsymbol{\phi}_{\boldsymbol{\nu}}])\right) = 0.$$

All terms in the integrand of the last equality are nonnegative, and so

$$\big(h\sigma_{\boldsymbol{\nu}} + m_{\boldsymbol{\nu}}\big)([\boldsymbol{U}_{\boldsymbol{\nu}}] + [u] \tan\alpha + h[\boldsymbol{\phi}_{\boldsymbol{\nu}}]) = \big(h\sigma_{\boldsymbol{\nu}} - m_{\boldsymbol{\nu}}\big)([\boldsymbol{U}_{\boldsymbol{\nu}}] + [u] \tan\alpha - h[\boldsymbol{\phi}_{\boldsymbol{\nu}}]) = 0$$

on γ. This justifies the following statement.

Theorem 1. *If the solution* $\boldsymbol{\xi} = (\boldsymbol{U}, u, \boldsymbol{\phi})$ *to* (8) *is sufficiently smooth then* $\boldsymbol{\xi} = (\boldsymbol{U}, u, \boldsymbol{\phi})$ *is also a solution to the boundary value problem* (2), (3), (10) *with the boundary conditions*

$$\boldsymbol{U} = \boldsymbol{\phi} = (0,0), \quad u = 0 \quad \textit{on } \Gamma_0, \tag{23}$$

$$\sigma_{\boldsymbol{\tau}} = m_{\boldsymbol{\tau}} = (0,0), \quad \sigma_{\boldsymbol{\nu}} \tan\alpha = \boldsymbol{q}_{\boldsymbol{\nu}} \quad \textit{on } \gamma, \tag{24}$$

$$\sigma_{\boldsymbol{\tau}} = m_{\boldsymbol{\tau}} = (0,0), \quad [\boldsymbol{U}_{\boldsymbol{\nu}}] + [u] \tan\alpha \leq h|[\boldsymbol{\phi}_{\boldsymbol{\nu}}]|, \quad -h\sigma_{\boldsymbol{\nu}} \geq |m_{\boldsymbol{\nu}}| \quad \textit{on } \gamma, \tag{25}$$

$$\left(\sigma_{\boldsymbol{\nu}} + \frac{1}{h} m_{\boldsymbol{\nu}}\right)([\boldsymbol{U}_{\boldsymbol{\nu}}] + [u] \tan\alpha + h[\boldsymbol{\phi}_{\boldsymbol{\nu}}]]) = 0 \quad \textit{on } \gamma, \tag{26}$$

$$\left(\sigma_{\boldsymbol{\nu}} - \frac{1}{h} m_{\boldsymbol{\nu}}\right)([\boldsymbol{U}_{\boldsymbol{\nu}}] + [u] \tan\alpha - h[\boldsymbol{\phi}_{\boldsymbol{\nu}}]) = 0 \quad \textit{on } \gamma. \tag{27}$$

Remark 1. The converse is also true: a smooth function $\boldsymbol{\xi} \in K$ satisfying the equilibrium equations (10) and the relations (2), (3), (24)–(27) is a solution to problem (8). We can verify this by analogy with the arguments of [9, 12].

Remark 2. With $\alpha \equiv 0$ in conditions (24)–(27) we recover the boundary conditions corresponding to the boundary value problem for a plate with vertical edges in contact with a rigid obstacle [13].

REFERENCES

1. *Ambartsumyan S. A* Theory of Anisotropic Shells. Washington: NASA, 1964.

2. *Panasyuk V. V., Savruk M. P., and Datsyshin A. P.* Stress Distribution Near Cracks in Plates and Shells [in Russian]. Kiev: Naukova Dumka, 1976.

3. *Osadchuk V. A.* Stress-Strain State and Limiting Equilibrium of Shells with Cuts [in Russian]. Kiev: Naukova Dumka, 1985.

4. *Morozov N. F.* Mathematical Questions of the Theory of Cracks [in Russian]. Moscow: Nauka, 1984.

5. *Shatskiĭ I. P. and Makoviĭchuk N. V.* Effect of closure of collinear cracks on the stress-strain state and the limiting equilibrium of bent shallow shells // Prikl. Mekh. i Tekh. Fiz. 2011. V. 52, N 3. P. 159–166.

6. *Khludnev A. M.* Elasticity Problems in Nonsmooth Domains [in Russian]. Moscow: Fizmatlit, 2010.

7. *Khludnev A. M.* Equilibrium problem of an elastic plate with an oblique cut // J. Appl. Mech. Techn. Phys. 1997. V. 38, N 5. P. 757–761.

8. *Kovtunenko V. A., Leont′ev A. N., and Khludnev A. M.* Equilibrium problem of a plate with an oblique cut // J. Appl. Mech. Techn. Phys. 1998. V. 39, N 2. P. 302–311.

9. *Lazarev N. P.* The equilibrium problem of a shallow Timoshenko-type shell with a through crack // Sibirsk. Zh. Industr. Mat. 2012. V. 15, N 3. P. 58–69.

10. *Lazarev N. P.* An iterative penalty method for a nonlinear equilibrium problem of a Timoshenko plate with crack // Sibirsk. Zh. Industr. Mat. 2011. V. 14, N 4. P. 381–392.

11. *Rudoĭ E. M.* Invariant integrals in the plane elasticity problem for bodies with rigid inclusions and cracks // Sibirsk. Zh. Industr. Mat. 2012. V. 15, N 1. P. 99–109.

12. *Lazarev N. P.* Problem of equilibrium of the Timoshenko plate containing a crack on the boundary of an elastic inclusion with an infinite shear rigidity // J. Appl. Mech. Techn. Phys. 2013. V. 54, N 2. P. 322–330.

13. *Lazarev N. P.* The method of fictitious domains in the equilibrium problem of a Timoshenko plate contacting with a rigid obstacle // Vestnik NGU Ser. Mat. Mekh. Informat. 2013. V. 13, N 1. P. 91–104.

14. *Pelekh B. L.* Theory of Shells with a Finite Shear Rigidity [in Russian]. Kiev: Naukova Dumka, 1973.

November 24, 2014

N. P. Lazarev
Institute of Mathematics of North-Eastern Federal University
Yakutsk, Russia
nyurgun@ngs.ru

Yakutian Mathematical Journal
July—September, 2014. Vol. 21, No. 3

UDC 517.956

ON A CERTAIN INVERSE LINEAR PROBLEM FOR A PARABOLIC SYSTEM OF EQUATIONS

S. G. Pyatkov and E. M. Korotkova

Abstract. We consider well-posedness in Sobolev spaces of an inverse linear problem of determining the right-hand side of a parabolic system. The overdetermination conditions are given on some collection of surfaces. The well-posedness of the problem is proven under some conditions of the boundary operators.

Keywords: parabolic system, inverse problem, boundary value problem, overdetermination condition

§ 1. Introduction

Let G be a bounded domain in $\mathbb{R}^n$ with boundary Γ of class C^{2m} and let $Q = (0,T) \times G$. The parabolic system is of the form

$$u_t + A(t,x,D)u = \sum_{i=1}^{r} b_i(t,x)q_i(t,x') + f, \quad (t,x) \in Q,\ x = (x', x''), \tag{1}$$

where $x' = (x_1, x_2, \dots, x_k)$ and $x'' = (x_{k+1}, x_{k+2}, \dots, x_n)$, b_i $(i = 1,2,\dots,r)$ and f are given vector-functions with the vanishing components b_i beginning with the number r_0+1 $(r_0 < h)$, and A is a matrix elliptic operator of order $2m$ and dimension $h \times h$ representable as

$$A(t,x,D) = \sum_{|\alpha| \le 2m} a_\alpha(t,x)D^\alpha, \quad D = (\partial_{x_1}, \partial_{x_2}, \dots, \partial_{x_n}).$$

The unknowns in (1) are a solution u and functions $q_i(t,x')$, $i = 1,2,\dots,r = sr_0$, occurring on the left-hand side of (1). Equation (1) is complemented with the initial and boundary conditions

$$u|_{t=0} = u_0, \quad B_j u|_S = \sum_{|\beta| \le m_j} b_{j\beta}(t,x)D^\beta u|_S = g_j(t,x), \tag{2}$$

where $m_j < 2m$, $j = 1,2,\dots,m$, and $S = (0,T) \times \Gamma$. Denote by $P_0 a$ the vector of length $r_0 < h$ whose coordinates agrees with the first r_0 coordinates of the vector a of length h. The overdetermination conditions for determining the functions q_i are of the form

$$P_0 u|_{S_i} = \psi_i(t,x'), \quad S_i = (0,T) \times \Gamma_i,\ i = 1,2,\dots,s, \tag{3}$$

where $\{\Gamma_i\}$ is a collection of smooth k-dimensional surfaces lying in G and ψ_i, $i = 1,2,\dots,s$, are given vector-functions.

The problems of this type arise in describing heat and mass transfer, diffusion and filtration, and so on. Many inverse coefficient problems with overdetermination conditions of the form (3) and $k = n - 1$ for second order parabolic equations are addressed in the articles by Yu. Ya. Belov, Yu. E. Anikonov, and some other authors (see [1]). In the case of $n = 1$ ($k = 0$) the unknowns q_i depend only on t and the surfaces S_i are just points. The problems of this type are examined, for instance, in [2]. We can refer also to [3, 4], where the problems of the form (1), (2) are considered in the general setting. Most of the articles is devoted to different model problems. One of the models describing heat and mass transfer is the Navier–Stokes system complemented with equations for the temperature and the concentrations of transferrable substances. Given the data of measurements on the cross-sections of the channel, the problem is to define the unknown parameters (the coefficients of the equations) or the densities of sources (the right-hand side) (see, e.g., [5–8]).

Many inverse and extremal problems in the stationary case are exposed in the articles by G. V. Alekseev (see, for instance, [9–11]). We should note also the articles [12–16]. In particular, the authors of [13] considered the question of solvability of (1)–(3) in the Hölder classes for a second order elliptic operator A and the authors of [3, 4], the solvability questions for (1)–(3) in both linear and nonlinear cases with $r_0 = h$. Our results are rather similar to those in [3]. We also point out the monographs [17–19] devoted to inverse problems for parabolic and elliptic equations and systems and the recent results [20–22].

In the present article we prove existence and uniqueness of solutions to (1)–(3) in a Sobolev space and establish continuous dependence of solutions on the data of the problem.

§ 2. Definitions, Notations, and Statements of the Main Results

Let E be a Banach space. Denote by $L_p(G;E)$ (with G a domain in $\mathbb{R}^n$) the space of strongly measurable functions on G with values in E and the finite norm $\|\|u(x)\|_E\|_{L_p(G)}$. We also employ the spaces $C^k(\overline{G};E)$ comprising functions having continuous and bounded partial derivatives up to the order k in G which admit continuous extensions on the closure $\overline{G}$. The Sobolev spaces $W_p^s(G;E)$ and $W_p^s(Q;E)$ are defined conventionally (see [23–26]). For fractional s the Sobolev space $W_p^s(G;E)$ coincides with the Besov space $B_{p,p}^s(G;E)$. If $E = \mathbb{C}$ or $E = \mathbb{C}^n$ then we use the notation $B_{p,p}^s(G)$. Similarly, we involve the notations $W_p^s(G)$ and $C^k(\overline{G})$ rather than $W_p^s(G;E)$ and $C^k(\overline{G};E)$ in this case.

The containment $u \in W_p^s(G)$ (or $u \in C^k(\overline{G})$) for a given vector-function $u = (u_1, u_2, \dots, u_k)$ means that every component u_i belongs to $W_p^s(G)$ (or $C^k(\overline{G})$). The norm on the corresponding space is the sum of the norms of the coordinates unless otherwise stated. The similar convention is adopted for matrices as well, i.e., $a \in W_p^s(G)$, $a = \{a_{ij}\}_{j,i=1}^k$, means that $a_{ij}(x) \in W_p^s(G)$ for all i and j. Given an interval $J = (0,T)$, put

$$W_p^{s,r}(Q) = W_p^r(J; L_p(G)) \cap L_p\big(J; W_p^s(G)\big),$$
$$W_p^{s,r}(S) = W_p^r(J; L_p(\Gamma)) \cap L_p\big(J; W_p^s(\Gamma)\big).$$

Let us describe the class of domains G. We say that the boundary $\Gamma = \partial G$ *belongs to* C^{2m} if, for every point $x_0 \in \Gamma$, there exist a neighborhood U (the coordinate neighborhood) and a coordinate system y (the local coordinate system) obtained from the initial by rotation and translation of the origin in which

$$\overline{U}\cap G=\{y\in\mathbb{R}^n : y'\in\overline{B}_r,\ \omega(y')<y_n\le\omega(y')+d\},\quad y'=(y_1,\dots,y_{n-1}),$$
$$\overline{U}\cap(\mathbb{R}^n\setminus\overline{G})=\{y\in\mathbb{R}^n : \omega(y')-d\le y_n<\omega(y')\},$$
$$\Gamma\cap\overline{U}=\{y\in\mathbb{R}^n : y'\in\overline{B}_r,\ y_n=\omega(y')\},$$

where $y'=(y_1,y_2,\dots,y_{n-1})$, $B_r=\{y' : |y'|<r\}$, $\delta>0$ is a constant, and $\omega\in C^{2m}(\overline{B}_r)$. Without loss of generality, we assume that the y_n-axis of the local coordinate system is directed along the normal to Γ at x_0.

Fix a parameter $p>n+2m$. Let $B_r(x_0)$ be the ball of radius r centered at x_0. Specify the conditions on G and the surfaces Γ_i.

(A) (a) Case of $k\ge 1$. There exists a domain $\Omega\subset\mathbb{R}^k$ with boundary of class C^{2m} such that $G\subset\Omega\times\mathbb{R}^{n-k}$,

$$\Gamma_i=\bigl\{x\in\mathbb{R}^n : x''=\varphi^i(x')=\bigl(\varphi^i_{k+1}(x'),\varphi^i_{k+2}(x'),\dots,\varphi^i_n(x')\bigr),\ x'\in\Omega\bigr\},$$

$\varphi^i(x')\in C^{2m}(\overline{\Omega})$ for all $i=1,2,\dots,s$ and a constant $\delta>0$ such that

$$U_{\delta i}=\{(x',\varphi^i(x')+\eta) : x'\in\Omega,\ \eta\in\mathbb{R}^{n-k},\ |\eta|<\delta\}\subset G,\quad U_{\delta i}\cap U_{\delta j}=\varnothing,$$

for $i\ne j$, $i,j=1,2,\dots,s$.

(b) Case of $k=0$. We take some interior points $\{x_i\}_{i=1}^s$ of G as $\{\Gamma_i\}_{i=1}^s$. Let $U_{\delta i}=B_\delta(x_i)$ and choose a number $\delta>0$ such that $\overline{U}_{\delta i}\subset G$ and $U_{\delta i}\cap U_{\delta j}=\varnothing$ for $i\ne j$, $i,j=1,2,\dots,s$.

The requirement (A) is a geometric condition; it is used in all articles on the inverse problems in question. It is fulfilled, for example, if $G=\Omega\times\mathbb{R}^{n-k}$, where Ω is a bounded domain of class C^{2m}.

Below we use the notations: $Q^\tau=(0,\tau)\times G$, $Q_0=(0,T)\times\Omega$, $Q_\tau=(0,\tau)\times\Omega$, $S_0=(0,T)\times\partial\Omega$, $G_\delta=\bigcup_i U_{\delta i}$, $\Gamma_\delta=\Gamma\cap\partial G_\delta$, $S_\delta=(0,T)\times\Gamma_\delta$, $Q_{\delta i}=(0,T)\times U_{\delta i}$, $Q_\delta=(0,T)\times G_\delta$, and $Q^\tau_\delta=(0,\tau)\times G_\delta$.

Further, we assume that the following conditions hold.

Consistency condition and smoothness of the data.

$$\exists\Phi(t,x)\in W_p^{1,2m}(Q) : \Phi|_{t=0}=u_0(x),\quad B_j\Phi|_S=g_j,\ j=1,\dots,m, \tag{4}$$

$$\partial_{x_i}\Phi\in W_p^{1,2m}(Q_\delta),\quad P_0\Phi|_{S_j}=\psi_j(t,x')\in W_p^{1,2m}(Q_0), \tag{5}$$

$$f\in L_p(Q),\quad \partial_{x_i}f\in L_p(Q_\delta),\quad f|_{S_j}\in C(\overline{Q}_0), \tag{6}$$

where $j=1,2,\dots,s$, $i=k+1,\dots,n$, and δ is a constant from (A).

As a consequence of (4)–(6) and the embedding theorems, we have

$$u_0(x)\in B_{p,p}^{2m-2m/p}(G),\quad g_j\in W_p^{k_j,2mk_j}(S),\quad k_j=\frac{2m-m_j-1/p}{2m}, \tag{7}$$

$$\partial_{x_i}g_j\in W_p^{k_j,2mk_j}(S_\delta),\quad \partial_{x_i}u_0(x)\in B_{p,p}^{2m-2m/p}(G_\delta), \tag{8}$$
$$j=1,2,\dots,m,\ i=k+1,\dots,n;$$

the latter conditions along with the agreement conditions ensure the existence of a function Φ satisfying (4) and (5). To justify this fact, we can apply the theorems on extension of functions on the boundary of a domain to a domain while preserving the necessary function classes. To simplify the arguments, we however use the conditions on the data written in the form (4), (5).

The conditions on the coefficients of A and B_j are more or less conventional. We assume that

$$a_\alpha(t,x) \in L_\infty(Q),\ |\alpha| < 2m, \quad a_\alpha(t,x) \in C(\overline{Q}),\ |\alpha| = 2m, \\ b_{j\beta} \in C^{2m-m_j}(\overline{S}), \quad j = 1,\dots,m,\ |\beta| \le m_j, \tag{9}$$

$$b_j(t,x) \in L_\infty(Q), \quad \partial_{x_i} b_j(t,x) \in L_\infty(Q_\delta), \quad j = 1,2,\dots,s,\ i \ge k+1, \tag{10}$$

$$\text{for all } i = k+1, k+2, \dots, n \text{ the functions } \partial_{x_i} a_\alpha(t,x) \text{ and } \partial_{x_i} b_{j\beta}(t,x) \\ \text{satisfy (9), where } Q \text{ is replaced with } Q_\delta, \text{ and } S \text{ with } S_\delta. \tag{11}$$

Consider the matrix $B(t,x')$ of order sr_0 whose rows with the numbers from $(j-1)r_0+1$ to jr_0 are occupied by the columns

$$(-P_0 b_1(t,x), -P_0 b_2(t,x), \dots, -P_0 b_r(t,x))|_{x''=\varphi^j(x')}.$$

The entries of this matrix are bounded on Q_0 almost everywhere. We require that there exists a constant $\delta_0 > 0$ such that

$$|\det B(t,x')| \ge \delta_0 \quad \text{almost everywhere in } \Omega \times (0,T). \tag{12}$$

Examine the system of equations

$$B(t,x')\vec{q}^{\,0} = \vec{g}, \quad \vec{q}^{\,0} = \big(q_1^0, q_2^0, \dots, q_{sr_0}^0\big), \tag{13}$$

where $\vec{g}$ is the column whose coordinates with the numbers from $(j-1)r_0+1$ till jr_0 coincide with the coordinates of the vector

$$P_0(f_0(t,x',\varphi^j(x')) - A\Phi(t,x',\varphi^j(x')) - \Phi_t(t,x',\varphi^j(x'))).$$

Under condition (12), the system (13) has the unique solution $\vec{q}^{\,0} = \big(q_1^0, \dots, q_{sr_0}^0\big) = (B(t,x'))^{-1}\vec{g}(t,x')$. The above conditions on the data ensure that $\vec{q}^{\,0} \in L_p(Q_0)$.

Consider the operator $A_0(t,x,D) = \sum_{|\alpha|=2m} a_\alpha D^\alpha$ and assume that the operator $\partial_t + A_0$ is parabolic, i.e., there exists a constant $\delta_1 > 0$ such that every root p of the polynomial $\det(A_0(t,x,i\xi) + pE) = 0$ (E is the identity matrix) satisfies the inequality

$$\operatorname{Re} p \le -\delta_1 |\xi|^{2m}, \quad \xi \in \mathbb{R}^n,\ (x,t) \in Q. \tag{14}$$

The Lopatinskiĭ condition is written in the form: given a point $(t_0,x_0) \in S$, rewrite the operators A_0 and B_{j0} at (t_0,x_0) ($B_{j0} = \sum_{|\beta|=m_j} b_{j\beta} D^\beta$) in the local coordinate system y and assume that the system

$$(\lambda E + A_0(i\xi', \partial_{y_n}))v(y_n) = 0, \quad B_{j0}(i\xi', \partial_{y_n})v(0) = h_j \in \mathbb{C}^h, \tag{15}$$

$\xi' = (\xi_1,\dots,\xi_{n-1})$, $y_n \in \mathbb{R}^+$, $j = 1,2,\dots,m$, has a unique solution in $C(\overline{\mathbb{R}}^+; \mathbb{C}^h)$ decreasing at infinity for all $\xi' \in \mathbb{R}^{n-1}$, $|\arg\lambda| \le \pi/2$, and $h_j \in \mathbb{C}^h$ such that $|\xi'| + |\lambda| \ne 0$.

Under conditions (4), (9), (14), and (15) the following theorem holds (see [27, Chapter 7, Theorem 10.4]).

Theorem 1. *Let G be a bounded domain with boundary of class C^{2m}. Then, for $g \in L_p(Q)$, there exists a unique solution $u \in W_p^{1,2m}(Q)$ to the problem*

$$u_t + A(t,x,D_x)u = g,\ (t,x) \in Q, \quad u|_{t=0} = u_0(x), \quad B_j u|_S = g_j,$$

satisfying

$$\|u\|_{W_p^{1,2m}(Q)} \le c\Big[\|g\|_{L_p(Q)} + \sum_{j=1}^{m}\|g_j\|_{W_p^{k_j,2mk_j}(S)} + \|u_0\|_{B_{pp}^{2m-2m/p}(G)}\Big],$$

where c is a constant independent of the data of the problem g, g_j, and u_0 and a solution u.

Fix $i \in \{1,2,\dots,s\}$ and make the change of variables $y' = x'$, $y'' = x'' - \varphi^i(x')$, $t = t$ in $Q_{\delta_1 i}$, $\delta_1 < \delta$. Denote by $A^i(t,y,D_y)$ and $B^i_j(t,y,D_y)$ the operators A and B_j written in new variables. Let $A^i_{y'}$ and $B^i_{jy'}$ designate the parts of the operators A^i and B^i_j not containing derivatives with respect to $y_{k+1}, y_{k+2},\dots,y_n$; in this case $A^i_{y''}$ and $B^i_{jy''}$ stand for the remaining parts of the operators. Similar meaning have the notations $A_{x'}$, $B_{jx'}$, $A_{x''}$,$B_{jx''}$, $A_{0x'}$, and $A_{0x''}$.

Describe the connections between the derivatives in new and old variables. We have

$$\partial_{x_j} = \partial_{y_j} - \sum_{r=k+1}^{n}\varphi^i_{ry_j}(y')\partial_{y_r},\ j\le k,\quad \partial_{x_j} = \partial_{y_j},\ j>k,$$

$$\partial_{y_j} = \partial_{x_j} + \sum_{r=k+1}^{n}\varphi^i_{rx_j}(x')\partial_{x_r},\ j\le k,\quad \partial_{y_j} = \partial_{xj},\quad j>k.$$

Thus, we infer

$$A^i_{y'}(t,y,D_{y'}) = A_{x'}(t,y',y''+\varphi^i(y'),D_{y'}),$$

$$B^i_{jy'}(t,y,D_{y'}) = B_{jx'}(t,y',y''+\varphi^i(y'),D_{y'}).$$

We can see that in the new variables $A_{x'}$ and $B_{jx'}$ have the same form. Let $\{a_{lm}(t,x',\varphi^i(x'),D_{x'})\}_{l,m=1}^h$ and $\{b_{lm}(t,x',\varphi^i(x'),D_{x'})\}_{l,m=1}^h$ be the entries of $A_{x'}(t, x',\varphi^i(x'),D_{x'})$ and $B_{jx'}(t,x',\varphi^i(x'),D_{x'})$, respectively. Denote by $\widetilde{A}_{x'}$ ($\widetilde{A}^i_{y'}$) the matrix operator with entries $\{a_{lm}\}_{l,m=1}^{r_0}$; and by $\widetilde{B}_{jx'}$ ($\widetilde{B}^i_{jy'}$), the matrix operator with entries $\{b_{lm}\}_{l,m=1}^{r_0}$. Given an arbitrary vector a of length h, denote by P_1a the vector of length h obtained from the vector a of length h by replacing with zeros its first r_0 coordinates. Assume that the following condition holds:

(B) For every $j=1,\dots,s$, the operator $\widetilde{A}^j_{y'}$ is parabolic in Q_0, $\deg P_0A^j_{y'}P_1v < 2m$, $j=1,2,\dots,s$, and the Lopatinskiĭ condition holds for the operators $\widetilde{A}^j_{y'}$ and $\widetilde{B}^j_{iy'}$, $i=1,\dots,m$, in Q_0.

In this case we can consider the auxiliary problems

$$\psi^j_t + \widetilde{A}^j_{y'}(t,y',D_{y'})\psi^j = 0,\quad (t,y')\in Q_0, \tag{16}$$

$$\psi^j(0,y') = 0, \tag{17}$$

$$\widetilde{B}^j_{iy'}\psi^j|_{S_0} = 0,\quad j=1,2,\dots,s,\ i=1,2,\dots,m, \tag{18}$$

where ψ^j is a vector of length r_0.

The following uniqueness problem holds (see [27, Chapter 7, Theorem 10.4]).

Theorem 2. *Assume that Ω is a bounded domain with boundary of class C^{2m} and the conditions* (9) *and* (B) *hold. Then solutions $\psi^j \in W_p^{1,2m}(Q_0)$, $j = 1, \dots, s$, to* (16)–(18) *are identically zero.*

Theorems 1 and 2 are the main statements used in the proof of our basic results. They are valid for a wide classes of domains (see [27]).

Assume that Ψ_0 is a class of vector-functions $\vec{\psi} = (\psi^1, \psi^2, \dots, \psi^s) \in W_p^{1,2m}(Q_0)$ of length r_0 such that (16) and (17) hold and there exists a function Φ satisfying (4), (5), with $u_0 = 0$, $g_j \equiv 0$, $j = 1, \dots, m$, such that

$$P_0(B_{ix'}(t, x', \varphi^j(x'), D_{x'})\widetilde{\psi^j})|_{S_0} = P_0 B_{ix'}(t, x', \varphi^j(x'), D_{x'})(I - P_1)\Phi|_{S_0}, \qquad (19)$$

where $i = 1, 2, \dots, m$, $j = 1, \dots, s$, and $\widetilde{\psi^j}$ is a vector of length h, whose first r_0 coordinates coincide with the coordinates of ψ^j, and the remaining coordinates are zero. We say that the equalities (3) hold in a generalized sense if there exists a vector-function $\vec{\psi} = (\psi^1, \psi^2, \dots, \psi^s) \in \Psi_0$ such that

$$P_0 u|_{S_i} = \psi_i(t, x') + \psi^i(t, x'), \quad (t, x') \in Q_0, \ i = 1, 2, \dots, s. \qquad (20)$$

The fulfilment of (3) in a generalized sense means that this equality holds in the quotient space $\big(W_p^{1,2m}(Q_0)\big)^s / \Psi_0$, where $\big(W_p^{1,2m}(Q_0)\big)^s$ is the space of vector-functions $\vec{\psi} = (\psi^1, \psi^2, \dots, \psi^s)$ whose components $\psi^i \in W_p^{1,2m}(Q_0)$ are vectors of length r_0.

Theorem 3. *Let conditions* (A), (B), (4)–(6), (9)–(12), (14), *and* (15) *hold. Fix $\delta_1 < \delta$. Then the following are valid:*

1. *There exists a constant $c > 0$ such that a solution $u, q_1, \dots, q_r$ to* (1)–(3) *from the class*

$$u \in W_p^{1,2m}(Q), \quad \nabla_{x''} u \in W_p^{1,2m}(Q_{\delta_2}), \quad \delta_2 < \delta, \quad q_j \in L_p(Q_0), \ j = 1, 2, \dots, r,$$

satisfies the estimate

$$\|u\|_{W_p^{1,2m}(Q)} + \|\nabla_{x''} u\|_{W_p^{1,2m}(Q_{\delta_1})} + \sum_{j=1}^{r} \|q_j\|_{L_p(Q_0)} \le c\bigg(\|\Phi\|_{W_p^{1,2m}(Q)}$$
$$+ \|\nabla_{x''}\Phi\|_{W_p^{1,2m}(Q_\delta)} + \|f\|_{L_p(Q)} + \|\nabla_{x''} f\|_{L_p(Q_\delta)} + \sum_{j=1}^{r_0} \|\psi_j\|_{W_p^{1,2m}(Q_0)} \bigg). \qquad (21)$$

2. *There exists a unique solution $(u, q_1, \dots, q_r)$ to* (1)–(3), *with* (3) *fulfilled in a generalized sense, from the class*

$$u \in W_p^{1,2m}(Q), \quad \nabla_{x''} u \in W_p^{1,2m}(Q_{\delta_1}), \quad \delta_1 < \delta, \quad q_j \in L_p(Q_0), \ j = 1, 2, \dots, r.$$

3. *A solution $(u, q_1, \dots, q_r)$ to* (1)–(3), *where $u_0 \equiv 0$, $f \equiv 0$, $g_j \equiv 0$, and $\vec{\psi} = (\psi_1, \psi_2, \dots, \psi_s) \in \Psi_0$, from the class*

$$u \in W_p^{1,2m}(Q), \quad \nabla_{x''} u \in W_p^{1,2m}(Q_{\delta_1}), \quad \delta_1 < \delta,$$

does not exists if $\vec{\psi} \not\equiv 0$.

4. *If $P_0\big(B^j_{iy''} v\big)\big|_{S_0} = 0$ and $P_0\big(B^j_{iy'} P_1 v\big)\big|_{S_0} = 0$ for every v and $i = 1, 2, \dots, m$, $j = 1, 2, \dots, s$, then $\Psi_0 = \{0\}$ and there exists a unique solution $(u, q_1, \dots, q_r)$ to* (1)–(3), *with* (3) *understood in the usual sense, from the class*

$$u \in W_p^{1,2m}(Q), \quad \nabla_{x''} u \in W_p^{1,2m}(Q_{\delta_1}), \quad \delta_1 < \delta, \quad q_j \in C(\overline{Q}_0), \ j = 1, 2, \dots, r.$$

§ 3. Proof of the Main Results

We need the auxiliary statement which results from the embedding theorems.

Lemma 1. *If $u \in W_p^{1,2m}(Q^\tau)$, $\tau > 0$, $p > n + 2m$, then the derivative $D_x^\alpha u$ for $|\alpha| \le 2m - 1$ belongs to $C(\overline{Q^\tau})$ after possible modification of the set of measure zero and if $u(0, x) = 0$ then*

$$\|D^\alpha u\|_{C(\overline{Q^\tau})} \le c\|u\|_{W_p^{1,2m}(Q^\tau)}\tau^\beta$$

for $|\alpha| < 2m$, where β and c are some positive constants independent of u.

Proof. This can be found in [3].

Proof of Theorem 3. First we prove claim 4. Let u be a solution to (1), (2). We now establish some estimates. Make the change of variables $q_i(t, x') = \mu_i(t, x') + q_i^0(t, x')$ and $u = v + \Phi$, where the function Φ is taken from (4). We have that

$$v_t + A(t, x, D)v = g + \sum_{i=1}^r b_i(t, x)\mu_i(t, x'), \quad (t, x) \in Q, \tag{22}$$

where

$$g = f - \Phi_t - A(x, t, D)\Phi + \sum_{i=1}^r b_i(t, x)q_i^0(x'),$$

$$v|_{t=0} = 0, \quad B_j v|_S = 0, \ j = 1, 2, \dots, m, \tag{23}$$

$$v|_{S_i} = v(t, x', \varphi^i(x')) = 0, \quad i = 1, 2, \dots, s. \tag{24}$$

Thus, we have reduced (1)–(3) to the simpler equivalent problem (22)–(24) which is studied below. Fixing $\mu_j \in L_p(Q_\tau)$ and determining a solution v to (22), (23) on the interval $(0, \tau)$, we specify some mapping $v = v(\vec{\mu})$, $\vec{\mu} = (\mu_1, \dots, \mu_r)$. Study its properties. Assume that $\|\vec{\mu}\|_{L_p(Q_\tau)} = \sum_{i=1}^r \|\mu_i\|_{L_p(Q_{\tau_0})}$ and $\|\vec{\mu}\|_{L_p(Q_\tau)} \le R_0$. The constant R_0 is defined below. By Theorem 1, we can express $v = v(\vec{\mu})$ through the functions μ_i as follows:

$$v = (\partial_t + A)^{-1}g + (\partial_t + A)^{-1}\sum_{i=1}^r b_i(t, x)\mu_i(t, x'). \tag{25}$$

Moreover,

$$\|v\|_{W_p^{1,2m}(Q^\tau)} \le c\|g\|_{L_p(Q)} + c\|\vec{\mu}\|_{L_p(Q_\tau)} \le c\|g\|_{L_p(Q)} + cR_0 = c_1. \tag{26}$$

Here and in what follows, the symbols c_i denote constants independent of the data f, g_j, u_0, and ψ_j of the problem. Demonstrate that this solution has an additional smoothness in $Q_{\delta j}^\tau$. Consider the domain $Q_{\delta j}^\tau$. Fix $\delta_2 < \delta_1 < \delta$ (the constant δ is that of (A)). Construct a function $\psi_0(x'') \in C_0^\infty(\mathbb{R}^{n-k}) : \psi_0 = 0$ for $|x''| \ge \delta_1$ and $\psi_0 = 1$ for $|x''| \le \delta_2$. In this case the function $\psi = \psi_0(x'' - \varphi^j(x'))$ is supported in $\overline{U}_{\delta j}$. Put $\Delta_i v = (v(x + e_i\eta) - v(x))/\eta$, where e_i is the ith coordinate vector, $i > k$, and $|\eta| < \delta - \delta_1$. The function $\tilde{v} = \psi(x)\Delta_i v$ is a solution to the problem

$$\tilde{v}_t + A(t, x, D)\tilde{v} = \psi[A, \Delta_i]v + \psi\Delta_i g + [A, \psi]\Delta_i v + \psi\Delta_i\bigg(\sum_{j=1}^r b_j(t, x)\mu_j(t, x')\bigg),$$

$$B_r\tilde{v}|_S = \psi[B_r, \Delta_i]v + [B_r, \psi]\Delta_i v, \quad r = 1, 2, \dots, m. \tag{27}$$

Here $[A,\Delta_i]=A\Delta_i-\Delta_iA$, $[A,\psi]=A\psi-\psi A$, and so on (thus the square brackets denote the corresponding commutator). The right-hand side is supported in $\overline{Q}^{\tau}_{\delta j}$. Moreover, by the properties of the finite differences (see [28, Chapter 2, Lemma 4.6]) and smoothness assumptions for the coefficients, we can say that the norm of the right-hand side of (27) in $L_p(Q^\tau)$ is bounded uniformly in η by

$$c_2(\|\partial_{x_i}g\|_{L_p(Q_{\delta j})}+\|g\|_{L_p(Q)})+c_3\|\vec{\mu}\|_{L_p(Q_{\tau_0})}.$$

Theorem 1, the estimates (26), and embedding theorems validate the inequality

$$\|\tilde{v}\|_{W_p^{1,2m}(Q^\tau)}\le c_4(\|\nabla_{x''}g\|_{L_p(Q_{\delta j})}+\|g\|_{L_p(Q)})+c_5R_0=c_6. \tag{28}$$

The constants on the right-hand side of (28) are independent of η and τ. In view of Lemma 4.6 in [28, Chapter 2], the generalized derivative $\partial_{x_i}v$ belongs to $W_p^{1,2m}\big(Q^\tau_{\delta_2 j}\big)$ and meets the estimate

$$\|v_{x_i}\|_{W_p^{1,2m}(Q^\tau_{\delta_2 j})}\le c_6. \tag{29}$$

Since the constants $\delta_2<\delta$ and j are arbitrary, we conclude that a solution v possesses the property $v_{x_i}\in W_p^{1,2m}\big(Q^\tau_{\delta_2 j}\big)$ for every $\delta_2<\delta$ and $j=1,2,\dots,s$. Without loss of generality, we can assume that the constant c_6 in (29) is also independent of $i=k+1,\dots,n$. Using (29), we can justify that

$$\|\nabla_{x''}v\|_{W_p^{1,2m}(Q^\tau_{\delta_2})}\le c_7(\|\nabla_{x''}g\|_{L_p(Q_\delta)}+\|g\|_{L_p(Q)})+c_8\|\vec{\mu}\|_{L_p(Q_\tau)}. \tag{30}$$

Prove solvability of the problem. Let v and $\vec{\mu}$ be a solution to (22)–(24); thus, $v=v(\vec{\mu})$. Passing in $Q_{\delta_1 j}$, $\delta_1<\delta$, to the variables $y'=x'$, $y''=x''-\varphi^j(x')$, $t=t$, we see that $Q_1=(0,T)\times\Omega\times B_{\delta_1}$, $B_{\delta_1}=\{y'':|y''|<\delta_1\}$. Consider the operators $A^j_{y'}$, $B^j_{iy'}$, $A^j_{y''}$, and $B^j_{iy''}$. The operators with the index y' are the parts of the corresponding operators not containing the derivatives with respect to the variables y''. Putting $y''=0$ and using the equalities $A^j_{iy'}(I-P_1)v|_{y''=0}=0$, $j=1,2,\dots,s$, $i=r+1,\dots,r$, we arrive at the relation

$$\begin{aligned}P_0\big(A^j_{y''}v\big)|_{y''=0}-P_0g(t,y',\varphi^j(y'))+P_0\big(A^j_{y'}P_1v\big)|_{y''=0}\\ =\sum_{i=1}^{r}P_0b_i(t,y',\varphi^j(y'))\mu_i(t,y').\end{aligned} \tag{31}$$

The right-hand side of (31) is written as $B(t,y')\vec{\mu}$, where the matrix B is defined in Section 2 and $\vec{\mu}=(\mu_1,\mu_2,\dots,\mu_r)$. The left-hand side is representable as the sum $\vec{g_0}(t,y')+H(\vec{\mu})$, with $\vec{g_0}$ and $H(\vec{\mu})$ the columns whose coordinates with the numbers from $(j-1)r_0+1$ till jr_0, $j=1,2,\dots,s$, are the vectors

$$P_0f(t,y',\varphi_j(y'))-P_0(A\Phi(t,y',\varphi_j(y')))-P_0\Phi_t(t,y')+\sum_{i=1}^{r}P_0b_i(t,y',\varphi_i(y'))q_i^0(t,x')$$

and, respectively,

$$-P_0\big(A^j_{y''}v(t,y',0)\big)-P_0\big(A^j_{y'}P_1v(t,y',0)\big).$$

By the definition of q_i^0, the first of these vectors vanishes. In this case

$$\vec{\mu}(t,y')=B^{-1}H(\vec{\mu})(t,y')=R(\vec{\mu})=R(0)+R_0(\vec{\mu}),\quad R_0(\vec{\mu})=R(\vec{\mu})-R(0). \tag{32}$$

We obtain a system of equations for the quantities μ_i. The right-hand side is the operator taking the vector-function $\vec{\mu}$ into the quantity

$$-P_0\big(A^j_{y''}v(t,y',0) + A^j_{y'}P_1 v(t,y',0)\big)|_{y''=0},$$

where v is a solution to (23)–(25). If $v \in W_p^{1,2m}(Q^\tau)$ and $\nabla_{x''}v \in W_p^{1,2m}\big(Q^\tau_{\delta_1}\big)$ for all $\delta_1 < \delta$ then the conditions on the coefficients and Lemma 1 ensure that all of the derivatives $D^\alpha v$ and $D^\alpha P_1 v$ occurring into the operators $P_0\big(A^j_{y''}v\big)\big|_{y''=0}$ and $A^j_{y'}P_1 v(t,y',0)$ is continuous on Q_τ (after a possible modification on a set of measure zero). Let $R_0 = 2\|R(0)\|_{L_p(Q_0)}$. Obviously, (in view of (26) and (29) with $R_0 = 0$) we have that

$$R_0 \le c(\|g\|_{W_p^{1,2m}(Q)} + \|\nabla_{x''}g\|_{W_p^{1,2m}(Q_\delta)}).$$

Show that there exists $\tau_1 \le T$ for which the operator

$$R(\vec{\mu}) = B^{-1}H(\vec{\mu})(t,y'), \quad R : L_p(Q_{\tau_1}) \to L_p(Q_{\tau_1}),$$

is defined, takes the ball $B_{R_0}(\tau_1) = \{\vec{\mu} \in L_p(Q_{\tau_1}) : \|\vec{\mu}\|_{L_p(Q_{\tau_1})} \le R_0\}$ of $L_p(Q_{\tau_1})$ into itself, and contractive in this ball. Note that the containment $\nabla_{x''}v \in W_p^{1,2m}(Q^\tau_{\delta_1})$ implies that $\nabla_{y''}v \in W_p^{1,2m}(\widetilde{Q}^\tau_{\delta_1})$ with $\widetilde{Q}^\tau_{\delta_1} = (0,\tau)\times\Omega\times B_{\delta_1}$, $B_{\delta_1} = \{y'' : \ |y''| < \delta_1\}$ and on the contrary. Fix $\delta_1 < \delta$. Lemma 1 and the estimates (26) and (30) yield

$$\|R(\vec{\mu}) - R(0)\|_{L_p(Q_\tau)} \le c_0\tau^\beta(\|\nabla_{x''}v_1\|_{W_p^{1,2m}(Q^\tau_{\delta_1})} + \|v_1\|_{W_p^{1,2m}(Q^\tau)}) \le \tau^\beta c_1. \quad (33)$$

By the definition of R_0, we have

$$\|B^{-1}H(\vec{\mu})(t,y')\|_{L_p(Q_\tau)} \le \frac{R_0}{2} + c_0\tau^\beta(\|\nabla_{x''}v_1\|_{W_p^{1,2m}(Q^\tau_{\delta_1})} + \|v_1\|_{W_p^{1,2m}(Q^\tau)}), \quad (34)$$

where v_1 is a solution to (22), (23) with $g = 0$. In view of (26) and (30) we have the estimate

$$\|R(\vec{\mu})\|_{L_p(Q_\tau)} \le \frac{R_0}{2} + \tau^\beta c_1. \quad (35)$$

Choose τ_1 so that $c_1\tau_1^\beta \le R_0$. By (35) the operator R is defined, takes the ball $B_{R_0}(\tau_1)$ into itself, and contractive in this ball. Applying the fixed point theorem we can state that there exist a unique solution to (32) in the ball $B_{R_0}(\tau_1)$. Let $v = v(\vec{\mu})$. Demonstrate that this function satisfies (24). By construction, v is a solution to (22), (23). Passing in $Q_{\delta_1 j}$, $\delta_1 < \delta$, to the new variables $y' = x'$, $y'' = x'' - \varphi^j(x')$, $t = t$, taking $y'' = 0$, and applying P_0, we obtain the equalities

$$\psi^j_t + \widetilde{A}^j_{y'}(t,y',D_{y'})\psi^j + \big(P_0A^j_{y'}P_1v + P_0A^j_{y''}v\big)\big|_{y''=0} = \sum_{i=1}^r P_0 b_i(t,y',0)\mu_j(t,y'),$$

$$B^j_i v\big|_{S_0^{\tau_1}} = 0, \quad S_0^{\tau_1} = (0,\tau_1)\times\Omega, \quad i = 1,\dots,m,\ j = 1,2,\dots,s, \quad (36)$$

rather than (22), (23), where $\psi^j = P_0 v|_{y''=0}$. In view of (31) and the conditions $P_0\big(B^j_{iy''}v\big)|_{S_0} = 0$ and $P_0\big(B^j_{iy'}P_1v\big)|_{S_0} = 0$ for every v and all $i = 1,2,\dots,m$, $j = 1,2,\dots,s$, we obtain

$$\psi^j_t + \widetilde{A}^j_{y'}(t,y',D_{y'})\psi^j = 0, \quad (37)$$

$$\widetilde{B}^j_{iy'}\psi^j|_{S_0^{\tau_1}} = P_0(B_{iy'}(I-P_1)v(t,y',0))|_{S_0^{\tau_1}} = 0, \quad i = 1,\dots,m,\ j = 1,2,\dots,s. \quad (38)$$

Since the claim of Theorem 2 is valid, we establish that $\psi^j \equiv 0$; hence, v satisfies (24).

Claims 1 and 4 of Theorem 3 are proven locally in time. But we can repeat the argument on $[\tau_0, 2\tau_0]$, $[2\tau_0, 3\tau_0]$, and so on noting that the solvability interval remains the same due to the fact that the problem is linear. Hence, we can justify solvability of the problem on the whole segment $[0,T]$.

Prove claim 2. Consider the integral equation (32) which, as we have demonstrated, is solvable locally in time. The function $v = u - \Phi$ satisfies (23). Moreover, the functions $\psi^j = P_0 v(t, y', 0)$ meet (37). Obviously,

$$\widetilde{B}_{ix'}\psi^j(t,x',\varphi^j(x')) = P_0(B_{ix'}(I-P_1)v)|_{x''=\varphi^j(x')}, \quad i=1,2,\dots,m,\ j=1,\dots,s. \tag{39}$$

Thus the collection $(\psi^1, \psi^2, \dots, \psi^s)$ belongs to the class Ψ_0. The equality $u = v + \Phi$ implies that the boundary conditions (3) are fulfilled in the generalized sense. Uniqueness of solutions satisfying (3) in the generalized sense is obvious, since the corresponding vectors $\vec{\mu}^1$ and $\vec{\mu}^2$ satisfy the same system (37).

Prove claim 3 of the theorem. Let u $(q_1, q_2, \dots, q_r)$ be a solution to (1)–(3) with data of the statement of Theorem 3. We have

$$u_t + A(t,x,D)u = \sum_{i=1}^{r} b_i(t,x)q_i(t,x'), \quad (t,x) \in Q, \tag{40}$$

$$u|_{t=0} = 0, \quad B_j u|_S = 0, \quad j = 1,2,\dots,m, \tag{41}$$

$$u|_{S_i} = \psi_i, \quad i = 1,2,\dots,s. \tag{42}$$

Passing in $Q_{\delta_1 j}$, $\delta_1 < \delta$, to the new variables $y' = x'$, $y'' = x'' - \varphi^j(x')$, $t = t$, putting $y'' = 0$, and involving the definition of class Ψ_0, we infer

$$P_0\big(A^j_{y''}u\big)|_{y''=0} + P_0 A^j_{y'} P_1 u = \sum_{i=1}^{r} P_0 b_i(t,y',\varphi^j(y'))q_i(t,y'),$$

where u is a solution to (40), (41). We arrive at an analog of (32) which is actually (32) with zero data. As we have proven (32) locally has the unique solution and so $\vec{q} \equiv 0$. From (40), (41) it follows that $u \equiv 0$; a contradiction.

REFERENCES

1. *Belov Yu. Ya.* Inverse Problems for Partial Differential Equations. Utrecht: VSP, 2002.
2. *Ivanchov M.,* Inverse Problems for Equations of Parabolic Type, WNTL Publ., Lviv (2003) (Math. Studies. Monograph Ser.; 10).
3. *Pyatkov S. G. and Samkov M. L.* On some classes of coefficient inverse problems for parabolic systems of equations // Siberian Adv. in Math. 2012. V. 22, N 4. P. 287–302.
4. *Pyatkov S. G.* On some classes of inverse problems for parabolic equations // J. Inverse Ill-Posed Probl. 2011. V. 18, N 8. P. 917–934.
5. *Capatina A. and Stavre R.* A control problem in biconvective flow // J. Math. Kyoto Univ. 1997. V. 37, N 4. P. 585–595.
6. *Babeshko O. M., Evdokimova O. V., and Evdokimov S. M.* On taking into account the types of sources and settling zones of pollutants // Dokl. Math. 2000. V. 61, N 2. P. 283–285.
7. *Kalinina E. A.* The numerical study of the inverse problem of source reconstruction for a two-dimensional nonstationary convection-diffusion equation // Dal′nevostochn. Mat. Sb. 2004. V. 5, N 1. P. 89–99.
8. *Kriksin Yu. A., Plyushchev S. N., Samarskaya E. A., and Tishkin V. F.* The inverse problem of source reconstruction for a convection-diffusion equation // Mat. Model. 1995. V. 7, N 11. P. 95–108.
9. *Alekseev G. V.* Solvability of control problems for stationary equations of magnetohydrodynamics of a viscous fluid // Siberian Math. J. 2004. V. 45, N 2. P. 197–213.

10. *Alekseev G. V. and Kalinina E. A.* Identification of the coefficient of the lowest order term for the stationary convection-diffusion-reaction equation // Sibirsk. Zh. Industr. Mat. 2007. V. 10, N 1. P. 3–16.

11. *Alekseev G. V.* Coefficient inverse extremum problems for stationary heat and mass transfer equations // Comp. Math. Math. Phys. 2007. V. 47, N 6. P. 1007–1028.

12. *Efremenkova O. V.* Solvability of a parabolic inverse problem for determining an absorption coefficient of special type // Mat. Zametki YaGU. 2006. V. 13, N 1. P. 72–79.

13. *Pyatkov S. G. and Tsybikov B. N.* On evolutionary inverse problems for parabolic equations // Dokl. Math. 2008. V. 77, N 1. P. 111–113.

14. *Sergienko I. V. and Deineka V. S.* Solution of inverse boundary-value problems for multicomponent parabolic distributed systems // Cybern. Syst. Anal. 2007. V. 43, N 4. P. 507–526.

15. *Farcas A. and Lesnic D.* The boundary-element method for the determination of a heat source dependent on one variable // J. Eng. Math. 2006. V. 54. P. 375–388.

16. *Iskenderova A. D. and Akhundov A. Ya.* Inverse problem for a linear system of parabolic equations // Dokl. Math. 2009. V. 79, N 1. P. 73–75.

17. *Kozhanov A. I.* Composite Type Equations and Inverse Problems. Utrecht: VSP, 1999.

18. *Isakov V.,* Inverse Problems for Partial Differential Equations, Springer-Verlag, Berlin (2006) (Appl. Math. Sci.; V. 127).

19. *Prilepko A. I., Orlovsky D. G., and Vasin I. A.* Methods for Solving Inverse Problems in Mathematical Physics. New York: Marcel Dekker, Inc., 1999.

20. *Pyatkov S. G. and Safonov E. I.* On some classes of linear inverse problems for parabolic systems of equations // Siberian Electronic Math. Reports. 2014. V. 11. P. 777–799.

21. *Pyatkov S. G. and Safonov E. I.* Some inverse problems for convection-diffusion equations // Vestnik YuUrGU Ser. Mat. Model. Progr. 2014. V. 7, N 4. P. 36–50.

22. *Pyatkov S. G. and Safonov E. I.* Determination of the source function in the mathematical models of convection-diffusion // Yakutian Math. J. 2014. V. 21, N 2. P. 107–118.

23. *Triebel H.* Interpolation Theory; Function Spaces; Differential Operators. Berlin: VEB Deutscher Verl. Wiss., 1978.

24. *Amann H.* Nonhomogeneous linear and quasilinear elliptic and parabolic boundary value problems // Function Spaces, Differential Operators and Nonlinear Analysis. Stuttgart; Leipzig: Teubner, 1993. P. 9–126. (Teubner-Texte Math.; V. 133).

25. *Amann H.* Compact embeddings of vector-valued Sobolev and Besov spaces // Glasnik Mat. 2000. V. 35. P. 161–177.

26. *Denk R., Hieber M., and Prüss J.* Optimal L_p-L_q-estimates for parabolic boundary value problems with inhomogeneous data // Math. Z. 2007. V. 257. P. 193–224.

27. *Ladyzhenskaya O. A., Solonnikov V. A., and Ural′tseva N. N.,* Linear and Quasilinear Equations of Parabolic Type, Amer. Math. Soc., Providence (1968) (Transl. Math. Monogr.; V. 23).

28. *Ladyzhenskaya O. A. and Ural′tseva N. N.* Linear and Quasilinear Elliptic Equations. New York and London: Academic Press, 1968.

December 5, 2014

S. G. Pyatkov
Sobolev Institute of Mathematics, Novosibirsk, Russia
Yugra State University, Khanty-Mansiisk, Russia
`s-pyatkov@ugrasu.ru;  pyatkov@math.nsc.ru`

E. M. Korotkova
Yugra State University, Yugra Scientific Institute
of Informational Technologies, Khanty-Mansiisk, Russia
`kem@uriit.ru`

Yakutian Mathematical Journal
July—September, 2014. Vol. 21, No. 3

UDC 517.926; 517.27

A NUMERICAL METHOD FOR SOLVING THE MINIMIZATION PROBLEM OF RESOURCE CONSUMPTION FOR LINEAR SYSTEMS WITH CONSTANT DELAYS IN THE STATE AND CONTROL

G. V. Shevchenko

Abstract. We propose a numerical method for solving the minimization problem of resource consumption for linear systems with constant time delay in both the phase states and the control. This method is a generalization of the method for solving the problem with delay only in the phase state. Global convergence of the method to a ε-optimal solution is proven. By an ε-optimal solution we mean an feasible control $u(t)$, $t \in [0, T]$, transferring the system into the ε-neighborhood of the origin on which the functional of the problem differs from the optimal value at most by ε.

Keywords: delay differential equation, functional, optimal control, numerical method

Let a controlled object be described by a system of linear ordinary differential equations with constant delays $h_1 > 0$ and $h_2 > 0$ in the phase state and the control, respectively, of the form

$$\begin{gathered}\dot{x}(t) = A(t)x(t) + C(t)x(t-h_1) + B(t)u(t) + D(t)u(t-h_2), \quad t \in [0,T], \\ x(t) = \varphi(t), \quad t \in [-h_1, 0],\end{gathered} \tag{1}$$

where x is the phase vector of the state of the object, $\varphi(t)$ is a given continuous function, $A(t)$, $C(t)$, and $B(t)$ are continuous matrices of sizes $n \times n$, $n \times n$, and $n \times s$, respectively, $u(t)$ is a measurable control obeying the constraint

$$|u_j(t)| \le 1, \quad j = \overline{1,s}, \ t \in [0,T], \tag{2}$$

and $u(t) \equiv 0$ on the interval $[-h_2, 0]$.

We assume that (1) is transferable for a given function φ on $[-h_1, 0]$ and the zero control on $[-h_2, 0]$ to the origin by feasible controls.

Problem 1. *Find a feasible control $u^0(t)$, $t \in [0,T]$, transferring* (1) *in a fixed time T to the finite zero state $x(T) = 0$ and minimizing the functional*

$$\mathscr{F}(u) = \int_0^T \sum_{j=1}^{s} \alpha_j |u_j(t)| \, dt, \tag{3}$$

where $\sum_{j=1}^{s} \alpha_j \ne 0$, $\alpha_j \ge 0$.

Obviously, Problem 1 has a solution only if $T \ge T_{\text{opt}}$, where T_{opt} is the time for transferring by time-optimal control (1) to the origin. It is naturally to assume that $T > T_{\text{opt}}$.

The author was supported by the Russian Foundation for Basic Research (Grant 13–01–00329) and the Siberian Division of the Russian Academy of Sciences (Grant 80).

By the Pontryagin maximum principle in the problem with delay [1, Chapter 4, Section 27; 2], we introduce the adjoint system

$$\dot{\psi} = \begin{cases} -A^*(t)\psi(t) - C^*(t+h_1)\psi(t+h_1), & t \in [0, T-h_1], \\ -A^*(t)\psi, & t \in [T-h_1, T], \end{cases} \tag{4}$$

and write out the Pontryagin function for the above problem as follows:

$$H(\psi(t), x(t), x(t-h_1), u(t), u(t-h_2)) = -\sum_{j=1}^{s} \alpha_j |u_j(t)|$$
$$+\langle \psi(t), A(t)x(t) + C(t)x(t-h_1)\rangle + \langle \psi(t), B(t)u(t) + D(t)u(t-h_2)\rangle, \tag{5}$$

where $\langle \cdot, \cdot \rangle$ is the inner product of vectors.

In this case, for optimality of $u^0(t)$ and $x^0(t)$, $t \in [0, T]$, it is necessary that a nonzero vector-function $\psi^0(t)$ be a solution to the adjoint problem (4) for some boundary condition $\psi(T) = c^0$ and

$$H(\psi^0(t), x^0(t), x^0(t-h_1), u^0(t), u^0(t-h_2))$$
$$= \begin{cases} \max\limits_{u \in U} [H(\psi^0(t), x^0(t), x^0(t-h_1), u, u^0(t-h_2)) + H(\psi^0(t+h_2), \\ \qquad x^0(t+h_2), x^0(t-h_1+h_2), u^0(t+h_2), u)], & t \in [0, T-h_2], \\ \max\limits_{u \in U} H(\psi^0(t), x^0(t), x^0(t-h_1), u, u^0(t-h_2)), & t \in [T-h_2, T], \end{cases} \tag{6}$$

where $U = \{u \in \mathbb{R}^s \mid |u_j| \le 1,\ j = \overline{1, s}\}$. Note that $x^0(t) = \varphi(t)$ for $t \in [-h_1, 0]$ and $u^0(t) = 0$ for $t \in [-h_2, 0]$.

The relations (6) and (5) yield

$$u^0(t) = \begin{cases} \arg\max\limits_{u \in U} \left[-2 \sum\limits_{j=1}^{s} \alpha_j |u_j| + \langle \psi^0(t), B(t)u \rangle + \langle \psi^0(t+h_2), D(t+h_2)u \rangle \right], \\ \qquad t \in [0, T-h_2], \\ \arg\max\limits_{u \in U} \left[-\sum\limits_{j=1}^{s} \alpha_j |u_j| + \langle \psi^0(t), B(t)u \rangle \right], \quad t \in [T-h_2, T]. \end{cases}$$

The last relation can be rewritten as

$$u_j(t) = \begin{cases} -1, & \Psi_j(t) < -2\alpha_j, \\ 0, & -2\alpha_j \le \Psi_j(t) \le 2\alpha_j, \\ 1, & \Psi_j(t) > 2\alpha_j, \end{cases} \quad j = \overline{1, s},\ t \in [0, T-h_2], \tag{7}$$

$$u_j(t) = \begin{cases} -1, & \langle \psi^0(t), B_j(t) \rangle < -\alpha_j, \\ 0, & -\alpha_j \le \langle \psi^0(t), B_j(t) \rangle \le \alpha_j, \\ 1, & \langle \psi^0(t), B_j(t) \rangle > \alpha_j, \end{cases} \quad j = \overline{1, s},\ t \in [T-h_2, T], \tag{8}$$

where $\Psi_j(t) = \langle \psi^0(t), B_j(t) \rangle + \langle \psi^0(t+h_2), D_j(t+h_2) \rangle$; $B_j(t)$ and $D_j(t+h_2)$ are the jth columns of the matrices $B(t)$ and $D(t+h_2)$, respectively, $j = \overline{1, s}$; and $\psi(t)$ is a solution to (4) with the boundary condition

$$\psi(T) = c^0. \tag{9}$$

Here c^0 ia a nonzero vector in $\mathbb{R}^n$.

By the homogeneity of (4), we can restrict demonstration to the consideration of the boundary conditions (9) with unit norm, $\|c\| = 1$, replacing (7) and (8) by the expressions

$$u_j(t) = \begin{cases} -1, & \Psi_j(t) < -2\mu\alpha_j, \\ 0, & -2\mu\alpha_j \le \Psi_j(t) \le 2\mu\alpha_j, \\ 1, & \langle\Psi_j(t) > 2\mu\alpha_j, \end{cases} \quad j = \overline{1, s},\ t \in [0, T - h_2], \tag{10}$$

$$u_j(t) = \begin{cases} -1, & \langle\psi^0(t), B_j(t)\rangle < -\mu\alpha_j, \\ 0, & -\mu\alpha_j \le \langle\psi^0(t), B_j(t)\rangle \le \mu\alpha_j, \\ 1, & \langle\psi^0(t), B_j(t)\rangle > \mu\alpha_j, \end{cases} \quad j = \overline{1, s},\ t \in [T - h_2, T], \tag{11}$$

where μ is a nonnegative real.

Let $S \subset \mathbb{R}^n$ be the unit sphere centered at the origin. Denote by $u(t, c, \mu)$ the control whose components satisfy (10) and (11) for some vector $c \in S$ and a nonnegative μ. It follows from (10) and (11) that if $\alpha_j > 0$ and

$$\mu \ge \mu_j(c) = \frac{1}{\alpha_j}\big\{ \max_{0 \le t \le T - h_2} |0{,}5\Psi_j(t)|, \max_{T - h_2 \le t \le T} |\langle\psi^0(t), B_j(t)\rangle| \big\}$$

then $u_j(t, c, \mu) \equiv 0$, $t \in [0, T]$, and for

$$\mu \ge \mu(c) = \max_{j \in \{i = \overline{1, s} | \alpha_i > 0\}} \mu_j(c)$$

all components of $u(t, c, \mu)$ with the corresponding numbers $\alpha_j > 0$ identically equal to zero.

Consider the function

$$\mathscr{G}(c, \mu) = \mathscr{F}(u(c, \mu)) = \int_0^T \sum_{j=1}^{s} \alpha_j |u_j(t, c, \mu)| \, dt$$

for a fixed c in S, where $u(c, \mu) = u(t, c, \mu)$, $t \in [0, T]$. The function $\mathscr{G}(c, \mu)$ on $[0, \mu(c)]$ is continuous and decreases monotonically in μ in view (10) and (11). Moreover, there exists a unique $\bar{\mu}(c)$, $0 \le \bar{\mu}(c) < \mu(c)$, such that $\mathscr{G}(c, \mu)$ on $[0, \bar{\mu}(c)]$ is a constant function

$$\mathscr{I}_{\max} \triangleq \sum_{j=1}^{s} \alpha_j T$$

strictly decreasing in μ on $[\bar{\mu}(c), \mu(c)]$. Hence, for every fixed c in S and a positive $\mathscr{I} \le \mathscr{I}_{\max}$ there exists a unique $\mu_{\mathscr{I}}(c) \in [\bar{\mu}(c), \mu(c)]$ such that $\mathscr{G}(c, \mu_{\mathscr{I}}(c)) = \mathscr{I}$.

Denote by $\Omega(\mathscr{I})$ the collection of the right endpoints of the trajectories of (1) for all feasible controls such that the value of (3) is less than or equal to $\mathscr{I}$; i.e.,

$$\Omega(\mathscr{I}) = \{x = x(T, u) \mid u(t) \in U,\ t \in [0, T],\ \mathscr{F}(u) \le \mathscr{I}\},$$

where $x(T, u)$ is a solution to (1) for a feasible control u at $t = T$. The boundary of $\Omega(\mathscr{I})$ is the set

$$\partial\Omega(\mathscr{I}) = \{x = x(T, u_{\mu_{\mathscr{I}}(c)}) \mid u_{\mu_{\mathscr{I}}(c)} = u_{\mu_{\mathscr{I}}(c)}(t) = u(t, c, \mu_{\mathscr{I}}(c)),\ t \in [0, T],\ c \in S\}.$$

The set $\Omega(\mathscr{I}_{\max})$ coincides with the attainability domain $\mathfrak{R}(T)$ of (1), since the boundaries coincide and they are both solid.

Since $\mathscr{G}(c,\mu)$ is strictly decreasing in μ on $[\bar{\mu}(c),\mu(c)]$ for every $c \in S$, we have

$$\Omega(\mathscr{I}_2) \subset \Omega(\mathscr{I}_1) \quad \text{for all } \mathscr{I}_1 > \mathscr{I}_2 \geq 0. \tag{12}$$

The fact that the set $\mathfrak{R}(T)$ is solid and (2) imply that $\Omega(\mathscr{I})$ is a body for every $0 \leq \mathscr{I} < \mathscr{I}_{\max}$. Moreover, in view of (12) we have the least $\mathscr{I}_{\min}$ such that the origin lies on the boundary of the set $\Omega(\mathscr{I}_{\min})$ and in the interior of $\Omega(\mathscr{I})$ for every $\mathscr{I} \in (\mathscr{I}_{\min}, \mathscr{I}_{\max}]$. Thus, there exist $c^* \in S$ and $\tilde{\mu}^*$ such that $x(T, u(c^*, \tilde{\mu}^*)) = 0$.

We propose to look for such $c^* \in S$ and $\tilde{\mu}^*$ with the use of simplices in simplex coverings of $\Omega(\mathscr{I})$ [3]. The description and justification of this method requires some new notions.

Let $z^1, \dots, z^{n+1} \in \mathbb{R}^n$ be distinct points such that their convex hull $\sigma = [z^1, \dots, z^{n+1}]$ is a body in $\mathbb{R}^n$. The set σ is called an *n-dimensional simplex with vertices $z^1, \dots, z^{n+1}$*. Two n-dimensional simplices σ^1 and σ^2 are referred to as *adjacent* if they have n common vertices and their intersection is an $(n-1)$-dimensional simplex. The definition implies that the intersection is a common facet (= face of maximal dimension).

Assume that $\mathfrak{B} \subset \mathbb{R}^n$ is a compact body, $\sigma^0 = [z_0^1, \dots, z_0^{n+1}]$ is an n-dimensional simplex with vertices on the boundary of $\mathfrak{B}$. Given its face $\sigma_j^0 = [z_0^1, \dots, z_0^{j-1}, z_0^{j+1}, \dots, z_0^{n+1}]$, $j = \overline{1, n+1}$, construct the adjacent simplex with vertices $z_0^1, \dots, z_0^{j-1}, \tilde{z}^j, z_0^{j+1}, \dots, z_0^{n+1}$, whose "new" vertex $\tilde{z}^j$ is a boundary point of $\mathfrak{B}$ of the maximal distance from the hyperplane passing through the remaining vertices and the point z_0^j and this vertex lie on the different sides of the hyperplane.

The simplices constructed are called *simplices of the first layer* and the simplex σ^0 is referred to as a *simplex of the zero layer*. Applying the same scheme, for every simplex of the first layer we can construct the adjacent simplices by using its $(n-1)$-dimensional faces not common with the $(n-1)$-dimensional faces of σ^0. The simplices constructed from the *second layer*. Clearly, the second layer contains exactly $n(n+1)$ simplices. Similarly, for every simplex of the kth layer ($k \geq 2$) we can construct the adjacent simplices of the $(k+1)$th layer.

Denote by $\mathfrak{S}_k$ the union of all simplices of the kth layer. Obviously, the kth layer contains $n^{k-1}(n+1)$ simplices. We call the set $\bigcup_{k=0}^{\infty} \mathfrak{S}_k$ a *convex covering* of $\mathfrak{B}$ and denote it by $\Pi_{\mathfrak{B}}$.

By construction, since $\mathfrak{B}$ is compact, $\operatorname{co} \mathfrak{B} = \overline{\Pi_{\mathfrak{B}}}$, where $\operatorname{co} \mathfrak{B}$ is the convex hull of $\mathfrak{B}$ and $\overline{\mathfrak{D}}$ is the closure of $\mathfrak{D}$.

Thus, the interior of $\mathfrak{B}$ lies in the convex covering $\Pi_{\mathfrak{B}}$ and its boundary in the boundary of the covering $\Pi_{\mathfrak{B}}$. (In what follows, by a covering by n-dimensional simplices we mean the above covering.) Hence, the following theorem holds.

Theorem 1 (on a covering). *The convex hull of every compact body coincides with the closure of a convex covering of a body.*

Corollary 1. *Let $\mathfrak{B}$ be a compact body in $\mathbb{R}^n$ and $z^0 \in \operatorname{int} \mathfrak{B}$. Then every covering $\Pi_{\mathfrak{B}}$ of a body $\mathfrak{B}$ contains a finite $k_0 \geq 0$ and an n-dimensional simplex $\sigma \in \mathfrak{S}_{k_0}$ such that $z^0 \in \sigma$.*

As easily seen, a simplex covering of every compact body $\mathfrak{B}$ in $\mathbb{R}^n$ is defined by a choice of (unique) simplex of the zero layer. Without loss of generality, we assume that this simplex $\sigma^0 = [z^1, \dots, z^{n+1}]$ is constructed in accord with the following scheme.

Let c^1 be an arbitrary point in $\mathbb{R}^n$. As the first vertex of a simplex σ^0, we choose a point z^1 from $\mathfrak{B}$ such that $\langle c^1, z^1 \rangle = \max_{x \in \mathfrak{B}} \langle c^1, x \rangle$. Obviously, z^1 is a boundary point of $\mathfrak{B}$.

The vertices z^j, $j = \overline{2, n+1}$, are chosen as follows. First, we find a solution c^j to the simultaneous linear algebraic equations

$$\langle c, z^i \rangle = -1, \quad i = \overline{1, j-1}, \tag{13}$$

which is normalized. Next, we choose a point z^j as the jth vertex such that

$$\langle c^j, z^j \rangle = \max_{x \in \mathfrak{B}} \langle c^j, x \rangle, \quad j = \overline{2, n+1}. \tag{14}$$

Obviously, z^j, $j = \overline{2, n+1}$, are boundary points of $\mathfrak{B}$.

Constructing the simplex σ^0 we have a possibility of various choices of the vectors c^j, $j = \overline{2, n}$, since (13) for $j < n+1$ is underdetermined and has a nonunique solution. We can exclude a possibility of various choices as follows: Let Z be a square matrix of order $j-1$ composed from the first $j-1$ linearly independent columns of the matrix of (13) and let $x = (x_1, \dots, x_{j-1})$ be a solution to the simultaneous linear algebraic equations $Zx^T = b^T$, where $b = (-1, \dots, -1)$. Take $c^j = (c^j_1, \dots, c^j_n)$ such that

$$c^j_i = \begin{cases} x_i, & \text{if the } i\text{th column of the matrix of (13) enters } Z,\ i = \overline{1, n}, \\ 0 & \text{otherwise,} \end{cases}$$

and normalize it.

If a solution to (13) is unique then it is possible that the choice of the vertices of the zero layer is not unique, since in the general case $\mathfrak{B}$ can be nonconvex and the maximum in (14) can be attained at several points. If it is true then we choose among them the jth vertex as the vertex with the minimal norm.

The attainability domain $\mathfrak{R}(T)$ (or what is the same $\Omega(\mathscr{I}_{\max})$) is a compact body in $\mathbb{R}^n$. Hence, we can apply the covering theorem and its corollary. Moreover, since the right-hand side of (1) is linear in phase states and control, $\mathfrak{R}(T)$ is a strictly convex set.

Section 2 presents some method and numerical algorithm for solving the control problem of Section 1.

2. The Method and Numerical Algorithm

The method proposed below relies on constructing a sequence of simplices $\{\sigma^k\}$ with vertices on the boundaries of $\Omega(\mathscr{I})$. Before we present a detailed numerical scheme of the method, we set forth its brief formal description.

The first stage. Construct a sequence of adjacent simplices $\{\sigma^k\}$ whose vertices are boundary points of $\mathfrak{R}(T)$ such that $\rho(\sigma^k) \geq \rho(\sigma^{k+1})$ for every $k \geq 0$, where $\rho(\sigma)$ designates the distance from the origin to the simplex σ. Note that this sequence is finite in view of the above corollary.

The simplex σ^0 is constructed in accord with the above scheme, where the attainability set $\mathfrak{R}(T)$ is taken as $\mathfrak{B}$. We terminate this construction when the simplex σ^{k_0} absorbs the origin. The simplex σ^k, $k \geq 1$, is constructed as follows: Let $z^* \in \sigma^{k-1}$ be such that $\|z^*\| = \min_{z \in \sigma^{k-1}} \|z\|$. Since $0 \notin \sigma^{k-1}$ (otherwise, constructing stops for $k_0 = k-1$) and z^* lies on the boundary of the simplex σ^{k-1} in some of its facets. Let this facet contain the vertices $z^1, \dots, z^{i_0-1}, z^{i_0+1}, \dots, z^{n+1}$. These vertices are the first n vertices of σ^k. We can find a solution $\hat{c}$ to the simultaneous linear algebraic equation:

$$\langle c, z^i \rangle = -1, \quad i = 1, \dots, i_0 - 1, i_0 + 1, \dots, n+1,$$

and normalize it. Take the end of the trajectory $\hat{x}(T)$ of (1) under the control

$$u(t) = \arg\max_{u\in U}\langle \hat{c}, x(t)\rangle, \quad t \in [0,T], \tag{15}$$

as the $(n+1)$th vertex. Under this control we have the equality $\langle \hat{c}, \hat{x}(T)\rangle = \max_{x\in\mathfrak{R}(T)}\langle \hat{c}, x\rangle$.

Assume that z^i is a vertex of the simplex σ^{k_0}, u^i and c^i are the control and adjoint parameters corresponding to this vertex, $i = \overline{1, n+1}$, and $\lambda_1^0, \dots, \lambda_{n+1}^0$ is a solution to the simultaneous linear algebraic equations:

$$\sum_{i=1}^{n+1} \lambda_i z^i = 0, \quad \sum_{i=1}^{n+1} \lambda_i = 1. \tag{16}$$

Since $0 \in \sigma^{k_0}$, we have $\lambda_i^0 \geq 0$ for all $i = \overline{1, n+1}$. Assign $\mathscr{I} = \mathscr{I}_{\max}$.

The second stage. If $\Lambda = \{i \mid \lambda_i^0 = 0\} \neq \varnothing$ then pass to Step 2.

Step 1. Find γ, $0 < \gamma < 1$, such that $0 \in [x(T, \gamma u^1), \dots, x(T, \gamma u^{n+1})]$ and $\langle c^i, x(T, \gamma u^i)\rangle > 0$ and put

$$c^* := \frac{\sum\limits_{i=1}^{n+1} \lambda_i^0 c^i}{\|\sum\limits_{i=1}^{n+1} \lambda_i^0 c^i\|}, \quad \mathscr{I} := \gamma\mathscr{I}, \quad z^i = x(T, \gamma u^i),\ i = \overline{1, n+1}.$$

Next we can find a solution $\lambda_1^0, \dots, \lambda_{n+1}^0$ to (13) and pass to Step 2.

Step 2. Put $i_0 := \min_{i\in\Lambda} i$, find a solution $\tilde{c}$ to the system of linear algebraic equations $\langle c, \tilde{x} + 0{,}5(z^i - \tilde{x})\rangle = -1$, $i \neq i_0, 1 \leq i \leq n+1$, where $\tilde{x}$ is the right end of the trajectory of a free motion of (1), (under the control $u = u(t) \equiv 0$, $t \in [0,T]$) and normalize it. If $\langle \tilde{c}, z^{i_0}\rangle < 0$, then we change the sign of $\tilde{c}$. Put $c^* := \tilde{c}$ and pass to the next step.

Step 3. Find $\mu_0(c^*) > 0$ such that $\mathscr{F}(u(c^*, \mu_0(c^*))) = \mathscr{I}$.

Let $z^* = x(T, u(c^*, \mu_0(c^*)))$. If $\|z^*\| \leq \varepsilon$, where $\varepsilon > 0$ is a necessary accuracy at the hit at the origin then the calculations stop. The control $u(c^*, \mu_0(c^*)) = u(t, c^*, \mu_0(c^*))$, $t \in [0,T]$, is approximately optimal and $\mathscr{F}(u(c^*, \mu_0(c^*)))$ is an approximately optimal value of (3).

If $\|z^*\| > \varepsilon$ then among the points z^i, $i = \overline{1, n+1}$, we choose n points $z^{i_1}, \dots, z^{i_n}$ such that $\sigma^* = [z^{i_1}, \dots, z^{i_n}, z^*]$ contains the origin. The points $z^{i_1}, \dots, z^{i_n}, z^*$ and the corresponding parameters $c^{i_1}, \dots, c^{i_n}, c^{i_{n+1}} = c^*$ are enumerated one after another. Find a solution $(\lambda_1^0, \dots, \lambda_{n+1}^0)$ to (16). If $|\Lambda| = 1$ then we pass to Step 1; otherwise, to Step 2.

Before describing the numerical algorithm of solving the problem of Section 1, we introduce the notations:

$c^{(k)}$ is the kth approximation of optimal values of the boundary conditions for the adjoint system (4);

ψ^k is a solution to (4) with the boundary condition $\psi(T) = c^{(k)}$;

$\mathscr{I}_k$ is the kth approximation of the optimal value $\mathscr{I}_{\min}$ of (3);

μ^1 and μ^2 are the lower and upper bounds of localization of a solution μ to the equation

$$\mathscr{F}(u(c^*, \mu)) = \mathscr{I}_k; \tag{17}$$

k is the number of an iteration.

The numerical algorithm of solving the above problem to within the notations coincides with the description of the numerical algorithm of solving the minimization problem of resource consumption with delay only in the phase state of the system [4]. The proof of convergence is similar to that in [4].

REFERENCES

1. *Pontryagin L. S., Boltyanskiĭ V. G., Gamkrelidze R. V., and Mishchenko E. F.* Mathematical Theory of Optimal Processes [in Russian]. Moscow: Nauka, 1983.

2. *Bokov G. V.* Pontryagin's maximum principle of optimal control problems with time-delay // J. Math. Sci. 2011. V. 172, N 5. P. 623–634.

3. *Shevchenko G. V.* Method to determine an optimal control in the minimum of resource consumption for the nonlinear stationary systems // Automation and Remote Control. 2009. V. 70, N 4. P. 672–682.

4. *Shevchenko G. V.* A numerical method to minimize resource consumption by linear systems with constant delay // Automation and Remote Control. 2014. V. 75, N 10. P. 1732–1742.

November 15, 2014

G. V. Shevchenko
Sobolev Institute of Mathematics, Novosibirsk, Russia
`shevch@math.nsc.ru`

Yakutian Mathematical Journal
July—September, 2014. Vol. 21, No. 3

UDC 51.76

ESTIMATING THE PROFILE OF THE MEAN FORCE POTENTIAL FOR TRANSMEMBRANE TRANSPORT OF A WATER MOLECULE BY THE UMBRELLA SAMPLING METHOD

M. Yu. Antonov, T. V. Naumenkova, A. V. Popinako, I. N. Nikolaev, and K. V. Shaĭtan

Abstract. Modeling the passive transmembrane transport of small molecules by molecular modeling methods faces a series of difficulties because in the available modeling time it is practically impossible to observe the process of spontaneous diffusion and measure its averaged macroscopic characteristics. Therefore, of interest are those approaches and methods that we can use to get in reasonable time the results enabling us to make comparative studies or comparisons with natural experimental data.

We use the methods of controlled molecular dynamics and umbrella sampling to study the process of transmembrane transport of water molecules through a lipid bilayer. We use a bilayer of 1,2-dimyristoyl-sn-glycero-3-phosphatidylcholine (DMPC) in the force field OPLS-AA. We construct the middle force profiles from the calculated data with trajectories of length 6–10 ns. We compare the resulting profiles for various spatial positions of the test water molecule (permeating through the bilayer), for different steps of the choice of original configurations, and for various initial values of the velocity of atoms in the system. The total modeling time was more than 10 ms. We demonstrate a significant dependence of the results of calculations on initial data and the simulation protocol. In the framework of this approach we estimate the kinetic permeability parameters and the potential of an average force, as well as study the dependence of the calculated values on the initial data of the simulation. We estimate the accuracy of the calculation basing on the analysis of a series of simulations.

Keywords: molecular modeling, umbrella sampling method, biomembranes

One of the main functions of the cell membrane is that of a barrier: to control the exchange of substances between the cell and intercellular medium. The structural basis of biological membranes are lipid bilayers. For a membrane to fulfill its barrier function, it is crucial that the lipid bilayer is semipermeable by some molecules like that of water.

Molecular modeling methods showed their efficiency in studying simulated biological systems a long time ago. It stands to reason to apply them to diffusion in lipid bilayers [1].

The authors were supported by the Russian Foundation for Basic Research (Grant 12–04–31934\13), the Sakha Scientific Educational Support Foundation for the Junior Scientists (Grant 2014–01–0008), and the Russian Science Foundation (Grant 14–24–00172).

In experimental studies of the permeability of lipid bilayers we measure the permeability coefficient P (in cm/sec) [2, 3]. However, modeling the passive transmembrane transport of small molecules on a computer with the use of molecular modeling methods faces a series of difficulties because in the available modeling time it is practically impossible to observe the process of spontaneous diffusion and measure its averaged macroscopic characteristics. This leads us to those approaches and methods that can bring in reasonable time the results enabling us to make comparative studies or draw comparisons with natural experimental data.

A widely used approach is to study transmembrane transport by the methods of controlled molecular dynamics. In the framework of this approach, an additional potential enables us to stimulate the system with the desired degrees of freedom. In particular, the method enables us to obtain the values of the "effective" microviscosity as well as to estimate the diffusion coefficient, which we can use for comparative analysis. A drawback of this approach is the observed dependence of the results on the simulation protocol, as well as the absence of estimate for the partition coefficient between the solvent and membrane [4, 5].

Another approach is to use the umbrella sampling method (US). Its use enables us to undertake statistical analysis of the high-energy regions of the configuration space, which are practically unpopulated in the equilibrium molecular dynamics simulations.

In this article we use the methods of controlled molecular dynamics and umbrella sampling to study the process of transmembrane transport of water molecules through a lipid bilayer. We use a bilayer of 1,2-dimyristoyl-sn-glycero-3-phosphatidylcholine (DMPC) in the force field OPLS-AA. We compare the resulting profiles for various spatial positions of the test water molecule (permeating through the bilayer), for various steps of the choice of original configurations, and for various initial values of the velocity of atoms in the system.

1. Materials and Methods

In essence, the molecular dynamics method reduces to representing the system under study as a set of material points whose interaction is described by the laws of classical mechanics. Each point is an atom or a group of atoms. Fig. 1 shows a system under study, as well as the shape and chemical structure of the DMPC molecule.

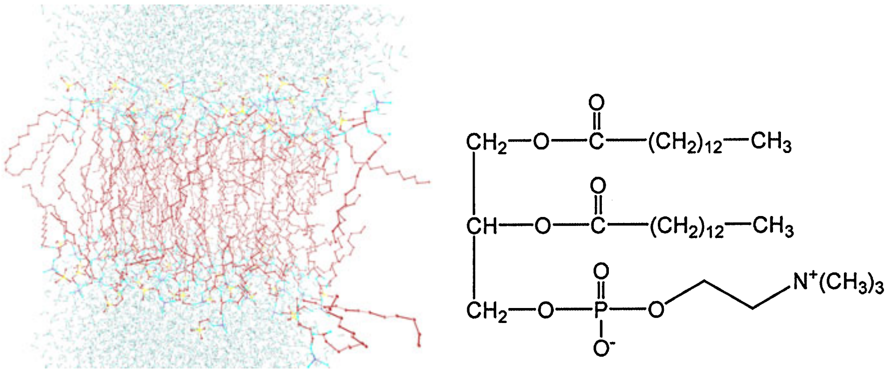
The sum of partial potentials describes interaction in the system:

$$\begin{aligned} U(r) = {} & \sum_{i,j} U^{v}_{ij}(b_{i,j}) + \sum_{i,j,k} U^{\theta}_{ijk}(\theta_{i,j,k}) \\ & + \sum_{i,j,k,l} U^{\phi}_{ijkl}(\phi_{i,j,k,l}) + \sum_{i \neq j} \big[U^{c}_{ij}(r_{i,j}) + U^{vdw}_{ij}(r_{i,j})\big], \end{aligned} \tag{1}$$

where the first three terms represent interaction between chemically bonded atoms (energies of valence bonds, valence angles, torsion angles), and the last two terms represent the electrostatic and van der Waals forces between the pairs of atoms which are not bonded.

As a rule, the energies of valence bonds and valence angles are described as the harmonic potential with a prescribed rigidity constant:

$$U^{v}_{ij}(b_{i,j}) = \frac{k_{ij}}{2}\big(b_{ij} - b^{0}_{ij}\big)^2. \tag{2}$$



(a) (b)

Fig. 1. (a) Form of the system under study; (b) shape and structure of the DMPC molecule

The value of the rigidity constant is chosen to match the geometry and oscillation frequencies of model compounds.

The torsion angle of four neighboring atoms amounts to the angle between the planes of the extreme triples of atoms. The corresponding potential is described as

$$U^{\phi}_{ijkl}(\phi_{i,j,k,l}) = \sum_{n} k^{\phi}_{ijkln}(1 + \cos(n\phi - \phi^0)). \tag{3}$$

In the molecular dynamics method the electrons are not explicitly accounted for. Each atom possesses a partial charge, while the Coulomb energy of interaction between the atoms is described by the standard expression

$$U^{C}_{ij}(r_{i,j}) = \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}. \tag{4}$$

Partial charges are assigned to describe the distribution of electron density in the system as well as it is possible.

Van der Waals interactions amount to repulsion and attraction of electrodynamic nature described by the Lennard–Jones potential

$$U^{LJ}_{ij}(r_{i,j}) = 4\varepsilon_{i,j}\left(\left(\frac{\sigma_{i,j}}{r_{i,j}}\right)^{12} - \left(\frac{\sigma_{i,j}}{r_{i,j}}\right)^{6}\right). \tag{5}$$

The general form of the potential energy function is essentially unchanged in various force fields, but the parametrization of the forces may differ. In this article we use the force field OPLS-AA [6, 7].

Therefore, the motion of atoms in the potential field is described by the system

$$m_i \ddot{\overrightarrow{r_i}} = -\left(\frac{\partial U}{\partial x_i}, \frac{\partial U}{\partial y_i}, \frac{\partial U}{\partial z_i}\right) \tag{6}$$

of second order ordinary differential equations, which is to be solved numerically. The applicable numerical methods differ in accuracy and computational complexity. The Verlet method [1], widely used in molecular dynamics, and its variants find a compromise between the accuracy of the procedure and the calculation speed. The coordinates of atoms in the new temporal layer are calculated as

$$\begin{aligned} r_i(t + \Delta t) &= 2r_i(t) - r_i(t - \Delta t) + \frac{F_i(t)}{m_i}\Delta t^2, \\ v_i(t) &= \frac{1}{2\Delta t}(r_i(t + \Delta t) - r_i(t - \Delta t)). \end{aligned} \tag{7}$$

It is inexpedient to use more accurate schemes in the molecular dynamics method due to high computational complexity, which cannot be compensated for by the use of larger integration meshsizes. Apart from that, the Verlet scheme conserves some integrals of motion and in general better conserves the energy of the system.

The study of systems in constant energy ensembles is usually of little interest because the majority of experimental data is obtained at constant temperature and/or pressure. Thus, to maintain constant temperature and pressure in molecular dynamics simulations, we use the special algorithms: thermostats and barostats. In this article we use a stochastic dynamics thermostat whose essence is that to the atoms of the system we apply an additional random (stochastic) force that describes the interaction of the system with an external medium of a specified temperature [8]. To control constant pressure, we use the Berendsen barostat [9]. In this approach the system volume can vary: We complement the main equation of motion with the equations of motion of the boundary of the periodic cell.

In experiments with the passive transmembrane transport of small molecules it is possible to measure the permeability coefficient P of the membrane and the partition coefficient K_p between water and a hydrophobic medium [2, 3, 10, 11]. Their values are related as

$$P = \frac{K_p \cdot D}{\Delta x}, \tag{8}$$

where D is the diffusion coefficient in the membrane, K_p is the interphase partition coefficient of the substance, and Δx is the thickness of the membrane. In molecular dynamics it is possible [4] in various ways to measure directly the diffusion coefficient D, whereas it impossible to measure directly the coefficient K_p of the interphase partition.

To describe the transition of the system between the states of the system, it is convenient to introduce the reaction coordinate ξ, which enables us to describe the changing system continuously. The Gibbs free energy G satisfies

$$G(\xi) = -kT \cdot \log(p(\xi)), \tag{9}$$

where $p(\xi)$ is the density of the probability of finding the system in a state with this value of ξ. In modeling transmembrane transport, we usually choose as ξ the coordinate z of the barycenter of the small molecule, where the z-axis is orthogonal to the lateral plane of the membrane (Fig. 2a).

Therefore, from an equilibrium molecular dynamics trajectory of the system under study we can obtain the profile of the free energy G by estimating the density of probability with the use of statistical methods. The problem is that in the reasonable modeling time it is practically impossible to observe spontaneous diffusion of small molecules through a lipid bilayer and it is impossible to gather adequate statistics concerning the probability of a certain phase because the regions of the conformation space with high values of G are either absent from the sample or insufficiently well presented.

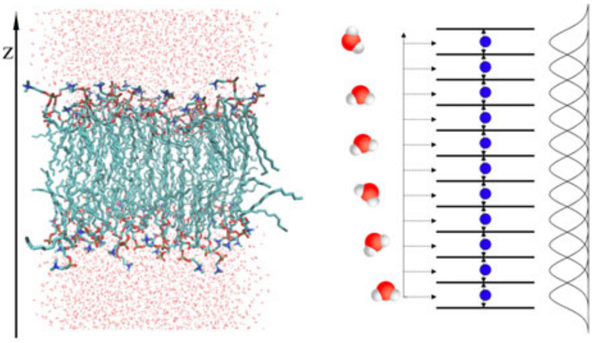
The umbrella sampling method is a solution here. It enables us to estimate in the reasonable modeling time the partition coefficient between water and membrane. In this approach, we apply to the system an additional potential capable of holding the molecule in question inside a potentially unprofitable region [12–14] (Fig. 2b). The modification of the original potential energy function amounts to the addition of an external potential; as a rule, we use the harmonic potential applied to the reaction coordinate:

$$\dot{U}(r) = U(r) - W(r), \quad \text{where } W(r) = k_w(\xi(r) - \xi_0)^2. \tag{10}$$

Since the form of $W(r)$ is known, we evaluate the free energy G of the unperturbed system as

$$G(\xi) = -kT \cdot \log(\dot{p}(\xi)) - W(\xi) + \text{const}, \tag{11}$$

where $\dot{p}(\xi)$ is the density of the probability of finding the system in a state with this value of ξ in the perturbed system.



(a) (b)

Fig. 2. (a) Axis of the reaction coordinate; (b) shape of harmonic potentials for holding a water molecule inside the membrane (regions of high values of free energy)

Using this approach, we can reliably reconstruct the profile of free energy only in a small neighborhood of ξ_0; hence, we choose a set of initial points along the reaction coordinate so that the distribution functions of the system near each initial point overlap with the distribution functions associated to the neighboring points.

There are several methods for reconstructing the original profile of free energy from the set of distributions found near each initial point. In this article we use the weighted histogram analysis method (WHAM, [15]).

2. A Molecular Dynamics Protocol

We used the model of water TIP4P/2005 which sufficiently well describes the liquid and crystal states, diffusion and viscosity properties, as well as surface tension effects [16–18]. To estimate the partial charges of atoms in the lipids, we used the unrestricted Hartree–Fock method, the basis 6-31G*, and the Mulliken method. The cutoff radius for nonvalent interaction was 1.8 nm. The calculations ran as an NPT ensemble. To maintain constant temperature and pressure, we used the stochastic dynamics thermostat and Berendsen barostat [9]. To account for the surface tension effects in the bilayer, we chose the negative pressure in the lateral plane sufficient to keep the specific area per lipid within experimentally observed limits [19]. The barostat pressure was 1 bar along the membrane normal, and 50 bar in the orthogonal direction. As the model membrane of an eukaryotic cell we used a bilayer of 80 DMPC molecules. To construct the bilayer, we used five conformations of molecules of each type. We placed the lipid molecules at the nodes of a hexagonal lattice and rotated them about their axes through random angles. The surface area per one lipid molecule was 0.6 nm^2. To assemble the membrane systems, we used the original software package *BiLayer*.

The simulations ran in the software package *Gromacs 4.6.7* [20]. Prior to the umbrella sampling procedure we performed the relaxation of the model membrane for 20 ns in the NPT ensemble at 300K. We specified the initial velocities of atoms using a random number generator with the Maxwell distribution. The integration meshsize was 1 fs. For the umbrella sampling procedure, we chose the starting configurations for the water molecule with the meshsizes of 0.05 and 0.1 nm along the z-axis (the reaction coordinate). We used the harmonic potential with the force constant 1000 kJ/(mol·nm^2). The trajectories to gather statistics were 6–10 ns long. We analyzed the dependence of the calculated energy profiles on the initial data, coordinates, the length of trajectories, as well as the dependence on the initial velocities of atoms in the system.

To process and inspect trajectories, we used the embedded modules of *Gromacs* [20], and for visualization, the *VMD* package [21].

3. The Results

Table 1 shows the main structural characteristics of bilayers after relaxation for 20 ns. The values were averaged over the last 2 ns of trajectories. For comparison we include the experimental data (E) of [22].

Table 1. Main structural characteristics of modeled bilayers

Parameter	DMPC bilayer	
	E	MD
Bilayer thickness, E	33.8	37.1±0.2
Surface area per lipid, E2	65.4	55.2±0.2
Order parameter	0.184	0.218

It is clear that the molecular dynamics protocol enables us to model membrane systems whose structural characteristics are close to those observed experimentally. The resulting membranes and the developed molecular dynamics protocol were used for further calculations.

We analyzed the dependence of the calculated profile of free energy on initial data. We considered the dependence on the length of modeling trajectories of the perturbed systems, the meshsize of sampling the starting configurations along the reaction coordinate, the initial velocities of atoms in the system, as well as the choice of the trajectory of the small molecule through the lipid bilayer.

Table 2. The calculated barrier height in experiments with meshsize 0.1 nm and $K_{\text{harm}} = 1000$ kJ/(mol·nm^2).

Simulation number	1	2	3	4	5	6	7	8	9
barrier height, kJ/mol	23.5	19.6	25.0	22.8	32.0	27.7	35.5	25.0	36.0

We studied the dependence of the calculated results on the position of initial coordinates of the water molecule in the lipid bilayer. To this end, we chose four possible trajectories of the water molecular through the lipid bilayer, and chose initial points on these trajectories with the meshsize 0.1 nm; the lengths of trajectories were 6–10 ns. We made nine calculations with different initial values of the

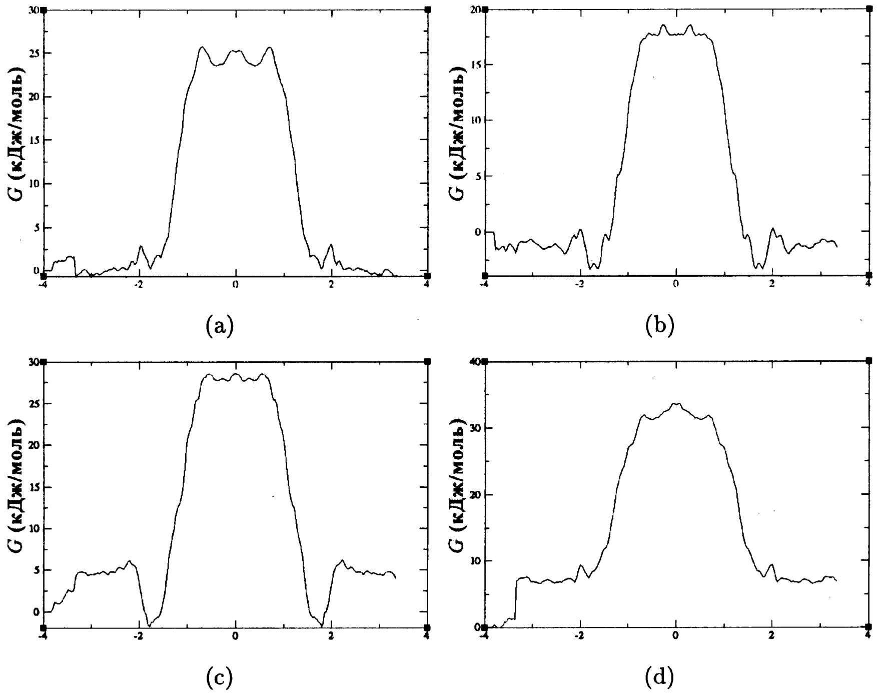


Fig. 3. Profiles of free energy in four experiments with various initial data

velocities of atoms in the system (Fig. 3). The calculated heights of the potential barrier presented in Table 2 were between 20 and 36 kJ/mol with the average calculated value of 27 kJ/mol and the root-mean-square deviation of 5.5 kJ/mol. The available experimental estimates for the interphase partition coefficient K_p between water and a hydrophobic solvent for water lie between $4.2{\cdot}10^{-5}$ (water and hexadecane [2]) and $1.4{\cdot}10^{-3}$ (water and olive oil [2]). The average barrier height 27 kJ/mol obtained from nine measurements corresponds to the calculated value $K_p^{\text{wat/membrane}} = 1.6 \cdot 10^{-4}$ and does not contradict the experimental estimates. At the same time, rather large root-mean-square deviation could indicate the stochastic character of the process under study and possible errors of the chosen measurement procedure.

The histograms in Fig. 4 characterize the statistical covering of the conformation space along the reaction coordinate in the simulations. Note that for the choice of meshsize 0.1 nm and the force constant K_{harm} of the harmonic potential equal to 1000 kJ/(mol·nm^2) the histogram reveals regions with low population.

To study the influence of the quality of choice of the conformation space, we ran additional simulations with meshsize 0.05 nm and the same value of the force constant K_{harm}. We made four measurements with different initial velocities of atoms. It is clear from Table 3 that the decrease to meshsize 0.05 nm does not lead to considerable changes in the root-mean-square deviation (the calculated value is 4.64 kJ/mol). The histograms characterizing the statistical covering of the conformation space along the reaction coordinate (not shown) fail to indicate the appearance of the regions weakly represented in the final sample.

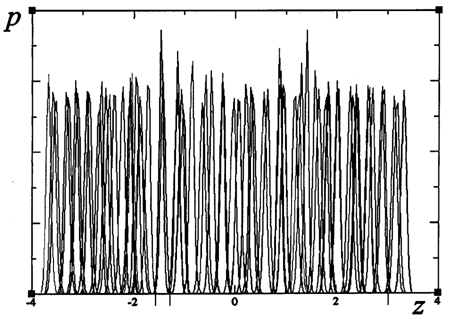


Fig. 4. The histogram for one simulation with meshsize 0.1 nm. Low population regions are noticeable (marked on the z-axis)

Table 3. The calculated barrier height in simulations with meshsize 0.05 nm and $K_{\text{harm}} = 1000$ kJ/(mol·nm^2)

Simulation number	1	2	3	4
Barrier height, kJ/mol	32.0	19.5	28.0	24.0

4. Conclusions

In the framework of our approach we estimated the kinetic permeability parameters and the profile of the mean force potential, as well as studied the dependence of the calculated values on the initial data of the simulation.

We showed that the umbrella sampling method enables us to study the transmembrane transport of water molecules through DMPC bilayer and estimate the profiles of the mean force. The calculated the average height of the barrier energy about 27 kJ/mol lies within the limits of admissible values known from experimental measurements of the coefficient K_p^{wat} of interphase partition between water and a hydrophobic solvent.

Sampling with meshsize 0.1 nm and the length of trajectories 6–10 ns can create regions in the conformation space with low population along the reaction coordinate, but for meshsize 0.05 nm we did not observe this. However, a small variation in the meshsize insignificantly influenced the measured value of the root-mean-square deviation, which may suggest in this case the stochastic nature of the process of transmembrane transport.

We can characterize the accuracy in nine simulations with meshsize 0.1 nm as conditionally sufficient for approximate estimates of the barrier height, but insufficient for estimating the interphase partition coefficient.

The simulations were made using the resources of the MSU supercomputer [23] and the NEFU supercomputer *Arian Kuz′min*.

REFERENCES

1. *Verlet L.* Computer experiments on classical fluids. I. Thermodynamical properties of Lennard–Jones molecules // Phys. Rev. 1967. V. 159, N 1. P. 98–103.
2. *Walter A. and Gutknecht J.* Permeability of small nonelectrolytes through lipid bilayer membranes // J. Membr. Biol. 1986. V. 90, N 3. P. 207–217.
3. *Diamond J. M. and Katz Y.* Interpretation of nonelectrolyte partition coefficients between dimyristoyl lecithin and water // J. Membr. Biol. 1974. V. 17, N 2. P. 121–154.
4. *Shaitan K. V., Antonov M. Y., Tourleigh Y. V., Levtsova O. V., Tereshkina K. B., Nikolaev I. N., and Kirpichnikov M. P.* Comparative study of molecular dynamics, diffusion, and

permeability for ligands in biomembranes of different lipid composition // Biochem. Suppl. Ser. A Membr. Cell Biol. 2008. N 2. P. 73–81.

5. *Antonov M. Yu., Naumenkova T. V., Levtsova O. V., and Shaĭtan K. V.* Study of dynamics and transmembrane diffusion by the methods of computer modeling // Supercomputer Technologies of Mathematical Modeling: Proc. of the II Intern. Conf. 2014. P. 16–29.

6. *Jorgensen W. L., Chandrasekhar J., Madura J. D., Impey R. W., and Klein M. L.* Comparison of simple potential functions for simulating liquid water // J. Chem. Phys. 1983. V. 79, N 2. P. 926.

7. *Jorgensen W. L., Maxwell D. S., and Tirado-Rives J.* Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids // J. Amer. Chem. Soc. 1996. V. 118, N 45. P. 11225–11236.

8. *Holm C. and Kremer K.* Advanced Computer Simulation. V. 173. Berlin and Heidelberg: Springer-Verlag, 2005.

9. *Berendsen H. J. C., Postma J. P. M., van Gunsteren W. F., DiNola A., and Haak J. R.* Molecular dynamics with coupling to an external bath // J. Chem. Phys. 1984. V. 81, N 8. P. 3684.

10. *Olbrich K., Rawicz W., Needham D., and Evans E.* Water permeability and mechanical strength of polyunsaturated lipid bilayers // Biophys. J. 2000. V. 79, N 1. P. 321–327.

11. *Bean R. C., Shepherd W. C., and Chan H.* Permeability of lipid bilayer membranes to organic solutes // J. Gen. Physiol. 1968. V. 52, N 3. P. 495–508.

12. *Buch I., Sadiq S. K., and de Fabritiis G.* Optimized potential of mean force calculations for standard binding free energies // J. Chem. Theory Comput. 2011. V. 7, N 6. P. 1765–1772.

13. *Kastner J.* Umbrella sampling // Wiley Interdiscip. Rev. Comput. Mol. Sci. 2011. V. 1, N 6. P. 932–942.

14. *Torrie G. M. and Valleau J. P.* Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling // J. Comput. Phys. 1977. V. 23, N 2. P. 187–199.

15. *Kumar S., Rosenberg J. M., Bouzida D., Swendsen R. H., and Kollman P. A.* The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method // J. Comput. Chem. 1992. V. 13, N 8. P. 1011–1021.

16. *Vega C. and de Miguel E.* Surface tension of the most popular models of water by using the test-area simulation method // J. Chem. Phys. 2007. V. 126, N 15. P. 154707.

17. *Abascal J. L. F. and Vega C.* A general purpose model for the condensed phases of water: TIP4P/2005 // J. Chem. Phys. 2005. V. 123, N 23. P. 234–505.

18. *Conde M. M., Gonzalez M. A., Abascal J. L. F., and Vega C.* Determining the phase diagram of water from direct coexistence simulations: the phase diagram of the TIP4P/2005 model revisited // J. Chem. Phys. 2013. V. 139, N 15. P. 154–505.

19. *Chiu S. W., Clark M., Balaji V., Subramaniam S., Scott H. L., and Jakobsson E.* Incorporation of surface tension into molecular dynamics simulation of an interface: a fluid phase lipid bilayer membrane // Biophys. J. 1995. V. 69, N 4. P. 1230–1245.

20. *Hess B., Kutzner C., van der Spoel D., and Lindahl E.* GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation // J. Chem. Theory Comput. 2008. V. 4, N 3. P. 435–447.

21. *Humphrey W., Dalke A., and Schulten K.* VMD: visual molecular dynamics // J. Mol. Graph. 1996. V. 14, N 1. P. 33-38.

22. *Petrache H. I., Dodd S. W., and Brown M. F.* Area per lipid and acyl length distributions in fluid phosphatidylcholines determined by (2)H NMR spectroscopy // Biophys. J. 2000. V. 79, N 6. P. 3172–3192.

23. *Sadovnichy V., Tikhonravov A., Voevodin V., and Opanasenko V.* ‘Lomonosov’: Supercomputing at Moscow State University // Contemporary high performance computing: From petascale toward exascale. 2013. P. 283–307.

November 28, 2014

M. Yu. Antonov; I. N. Nikolaev
North-Eastern Federal University, Yakutsk, Russia
mikhail@s-vfu.ru; n_ivan_n@mail.ru

T. V. Naumenkova; K. V. Shaĭtan
Lomonosov Moscow State University, Moscow, Russia
tnaumenkova@gmail.com; shaytan49@yandex.ru

A. V. Popinako
A. N. Bach Institute of Biochemistry, Moscow, Russia
popinakoav@gmail.com

Yakutian Mathematical Journal
July—September, 2014. Vol. 21, No. 3

UDC 517.977.58

MODIFICATION OF THE NEUSTADT—EATON METHOD

V. G. Starov

Abstract. In the Neustadt–Eaton method we need to solve a discrete equation in order to find an initial value of the adjoint system. Solving the equation, on every step we repeatedly determine a solution to a differential equation for an optimal trajectory. We offer to use a “sliding” approximation of a solution to the discrete equation for every component with the help of an algebraic function with its successive extrapolation. The extrapolation allows us to essentially decrease numerical costs.

Keywords: admissible control, optimal control, adjoint system

1. Statement of the Problem and the Brief Summary of the Neustadt–Eaton Method

Let a controllable system be described by the linear differential equation

$$\dot{x} = Ax + Bu, \quad x(t_0) = x_0. \tag{1}$$

Here x is an n-dimensional vector of the phase state, A and B are matrices of dimensions $n \times n$ and $n \times m$, respectively, and u is an m-dimensional control whose components belong to the class of piecewise continuous functions satisfying the constraints $|u_j| \leqslant M_j$, $M_j > 0$, $j = \overline{1, m}$.

It is assumed that the linear system is completely controllable and can be transferred into the origin by a bounded control; i.e., x_0 belongs to the controllability domain V.

Problem. *Find a control u transferring* (1) *from the initial state x_0 to the origin $x(t_k) = 0$ in minimum time $T = t_k - t_0$.*

The above-mentioned method is presented in [1]. In our case the Hamiltonian H takes the form

$$H(\psi, x, u) = \psi(Ax + Bu),$$

and the adjoint system of equations is written as

$$\dot{\psi} = -A'\psi. \tag{2}$$

Let $p = (p_1, \dots, p_n)$ be a nonzero vector. Denote by $\psi(t, p)$ a solution to (2) with the initial data $\psi(0, p) = p$ and by $u(t, p)$, a control corresponding to $\psi(t, p)$; the maximum principle yields

$$\psi(\tau)Bu(\tau) = \max_{u \in U} \psi(\tau)Bu.$$

This control is considered on the segment $0 \leqslant t \leqslant T$. Denote by $x(t)$ the trajectory corresponding to the control $u(t, p)$ and satisfying the terminal condition $x(T) = 0$. The initial point $x(0)$ of this trajectory is denoted by $\xi_T(p) =$

$\big(\xi_T^1(p), \dots, \xi_T^n(p)\big)$ and $x(t)$ is an optimal trajectory transferring the point $x_0 = \xi_T(p)$ to the origin in time T. Write out a solution to the differential equation (1) at time T, taking it into account that $x(t) = 0$. We infer

$$\sum_{i=1}^{n} \varphi_i(T) \left[\xi_T^i(p) + \int\limits_0^T (\psi^i(\tau) B u(\tau, p))\, d\tau \right] = 0.$$

The expressions in the square brackets vanish due to the linear independence of the vectors $\varphi_1(T), \dots, \varphi_n(T)$ and so

$$\xi_T^i(p) = -\int\limits_0^T (\psi^i(\tau) B u(\tau, p))\, d\tau, \quad i = 1, 2, \dots, n. \tag{3}$$

So, for given an n-dimensional (nonzero) vector p, we can define some point $\xi_T(p)$ by (3).

The Neustadt–Eaton method allows us to recover a vector p that solves (1).

The function

$$f(t, p) = p(x_0 - \xi_t(p))$$

is introduced and its continuity in t and p is demonstrated. For a fixed p, the function increases and there exists a unique t, $0 \leqslant t \leqslant t_k$, such that $f(t, p) = 0$. Denote by $F(p)$ this value t.

Examine the differential equation

$$\frac{dp}{d\tau} = -[x_0 - \xi_{F(p)}(p)],$$

where $\xi_{F(p)}(p)$ is a solution to (3) at time $t = F(p)$. It can be proven that a solution $p(\tau)$ as $\tau \to \infty$ solves (1).

Eaton replaces the differential equation with the difference equation

$$\frac{\Delta p_k}{\Delta \tau_k} = \frac{p_{k+1} - p_k}{\Delta \tau_k} = -[x_0 - \xi_{F(p)}(p)].$$

Hence,

$$p_{k+1} = p_k - \Delta\tau_k [x_0 - \xi_{F(p)}(p)];$$

next, the normalization procedure gives that

$$p_{k+1} = \frac{q_{k+1}}{|q_{k+1}|}, \quad \text{where } q_{k+1} = p_k - \Delta\tau_k [x_0 - \xi_{F(p)}(p)].$$

In the final form the algorithm for calculating p is as follows: Put

$$q_{k+1}^{(m)} = p_k - 2^{-m}(x_0 - \xi_{F(p_k)}(p_k)) \tag{4}$$

and choose the least nonnegative integer m such that the vector $p_{k+1} = \frac{q_{k+1}^m}{|q_{k+1}^m|}$ obeys the inequality

$$p_{k+1}(x_0 - \xi_{F(p_k)}(p_{k+1})) < -2^{-(m+1)} |(x_0 - \xi_{F(p_k)}(p_k))|^2.$$

Take the vector $-x_0/|x_0|$ as p_0.

The weak convergence of the method is pointed out in [2].

2. The Method of Solution

We propose to approximate a solution to (4) obtained by the Neustadt–Eaton method for every component by an algebraic dependence (a function).

2.1. Quadratic approximation of functions. Under the point quadratic approximation [3], the measure of a deviation of a polynomial

$$Q_c(z) = a_0 + a_1 z + \cdots + a_c z^c$$

from a given function $y = f(z)$ on the point set $z_0, z_1, \ldots, z_l$ is the quantity

$$S = \sum_{i=1}^{l} [Q_c(z_i) - f(z_i)]^2,$$

called a *quadratic error*. To construct an approximating polynomial, we need to choose the coefficients $a_0, a_1, \ldots, a_c$ minimizing S.

To solve the approximation problem we employ the general methods of differential calculus. Namely, find the derivatives of S with respect to the variables $a_0, a_1, \ldots, a_c$. Equating these partial derivatives to zero, we obtain a system of $c+1$ equations with $c+1$ unknowns $a_0, a_1, \ldots, a_c$.

We can take the linear model

$$y = a_0 + a_1 z. \tag{5}$$

The coefficients a_0 and a_1 are determined from the system

$$a_0 s_0 + a_1 s_1 = t_0,$$

$$a_0 s_1 + a_1 s_2 = t_1,$$

where $s_0 = \sum_{i=0}^{l} 1$, $s_1 = \sum_{i=0}^{l} z_i$, $s_2 = \sum_{i=0}^{l} z_i^2$, $t_0 = \sum_{i=0}^{l} y_i$, $t_1 = \sum_{i=0}^{l} z_i y_i$, and l is the number of points.

By numerical studies, we elaborate the following algorithm for solving (1).

2.2. An algorithm of a modified method. We are given the vector $p_0 = -\frac{x_0}{|x_0|}$.

Step 1. Make l steps of the Neustadt–Eaton method and store in memory the values of all components of the vector p at every step. The time $\Delta t = \sum_{k=1}^{l} \Delta\tau_k$ is also stored.

Step 2. Given components of the vector p, define an approximating dependence by (5) and extrapolate it on the time Δt beyond the lth point.

Step 3. Taking the extrapolating values as the vector p_0, we can repeat the calculations of Step 1.

Next, we can pass to Step 2 and so on.

In this way we can realize a “sliding” approximation and extrapolation of a solution to (4).

To confirm efficiency of this modification of the method, we have carried out numerical experiments.

3. Numerical Simulation

We consider systems of linear differential equations of the third and fourth orders.

Solving (4), by efficiency we mean the percentage of the total time of integrations in all iterations of the modified algorithm in that for the Neustadt–Eaton method.

In the examples below $l = 4$, $b = 4$, and $M = 5$. The quantities x_{10}, x_{20}, x_{30}, and x_{40} in tables are the initial values of the phase coordinates of (1), $T_{\Sigma N}$ stands for the total time of integration on all iterations of the Neustadt–Eaton method, and $T_{\Sigma M}$ stands for the total time of integration on all iterations of the modified method.

Example 1.

$$\dot{x_1} = x_2, \quad \dot{x_2} = x_3, \quad \dot{x_3} = bu, \ |u| \leqslant M,$$

$$x_1(t_0) = x_{10}, \quad x_2(t_0) = x_{20}, \quad x_3(t_0) = x_{30}.$$

Table 1. Example 1

x_{10}	x_{20}	x_{30}	$T_{\Sigma N}$	$T_{\Sigma M}$	%
5.0	5.0	5.0	3537	2027	42.7
0.0	0.0	4.0	1612	907	43.7
6.0	10.0	0.0	3051	1745	42.8
70.0	0.0	0.0	12577	6023	52.1
0.0	50.0	0.0	19134	10546	44.9
0.0	5.0	13.0	3782	2260	40.2
5.0	0.0	10.0	3586	1652	53.9

Example 2.

$$\dot{x_1} = x_2, \quad \dot{x_2} = x_3, \quad \dot{x_3} = a_1x_1 + a_2x_2 + a_3x_3 + bu, \ |u| \leqslant M,$$

$$x_1(t_0) = x_{10}, \quad x_2(t_0) = x_{20}, \quad x_3(t_0) = x_{30},$$

$$\lambda_1 = -2.704, \quad \lambda_{23} = -0.198 \pm 0.830.$$

Table 2. Example 2

x_{10}	x_{20}	x_{30}	$T_{\Sigma N}$	$T_{\Sigma M}$	%
1.0	1.0	0.0	41707	18433	55.8
0.0	1.0	0.0	15504	7901	49.0
0.0	0.0	6.0	26832	9246	65.5
1.0	0.0	0.0	23913	7936	66.8
0.0	1.0	1.0	13377	4008	70.0

Example 3.

$$\dot{x_1} = x_2, \quad \dot{x_2} = x_3, \quad \dot{x_3} = x_4, \quad \dot{x_4} = bu, \ |u| \leqslant M,$$

$$x_1(t_0) = x_{10}, \quad x_2(t_0) = x_{20}, \quad x_3(t_0) = x_{30}, \quad x_4(t_0) = x_{40}.$$

Table 3. Example 3

x_{10}	x_{20}	x_{30}	x_{40}	$T_{\Sigma N}$	$T_{\Sigma M}$	%
1.0	1.0	0.0	0.0	20966	9159	56.3
0.0	0.0	2.0	0.0	16404	7916	51.7
1.0	1.0	1.0	1.0	15122	9170	39.4
0.0	0.0	0.0	7.0	42216	19501	53.8
0.0	0.0	2.0	2.0	25198	9998	60.3
1.0	2.0	3.0	4.0	127285	13484	89.4
1.0	2.0	2.0	1.0	34436	12487	63.7
8.0	0.0	0.0	0.0	39160	14897	62.0
0.0	1.0	0.0	0.0	16728	5998	64.1
1.0	0.0	0.0	5.0	19635	9484	51.7
0.0	1.0	1.0	0.0	12354	6654	46.1

Example 4.

$$\dot{x_1} = x_2, \quad \dot{x_2} = x_3, \quad \dot{x_3} = x_4, \quad \dot{x_4} = a_{41}x_1 + a_{42}x_2 + a_{43}x_3 + a_{44}x_4 + bu, \ |u| \leqslant M,$$

$$x_1(t_0) = x_{10}, \quad x_2(t_0) = x_{20}, \quad x_3(t_0) = x_{30}, \quad x_4(t_0) = x_{40},$$

$$\lambda_{12} = -0.784 \pm 0.986, \quad \lambda_{34} = -1.016 \pm 0.916.$$

Table 4. Example 4

x_{10}	x_{20}	x_{30}	x_{40}	$T_{\Sigma N}$	$T_{\Sigma M}$	%
1.0	1.0	1.0	1.0	83317	32491	61.0
2.0	3.0	0.0	0.0	97357	41457	57.4
4.0	0.0	0.0	0.0	77071	46003	40.3
1.0	2.0	3.0	4.0	107298	60854	43.3
0.0	5.0	0.0	0.0	100502	52527	47.7
2.0	0.0	0.0	20.0	92762	40371	56.5
0.0	3.0	3.0	0.0	81823	38515	52.9
630.839	−614.358	74.716	1191.669	34356	15779	54.1
−82.290	155.603	−222.679	187.20	11560	6753	41.6
−93.192	180.223	−260.808	213.634	23517	10956	53.4

Conclusion

The modification proposed allows us to essentially reduce numerical costs. In the above examples the costs are fifty percent less than those in the Neustadt–Eaton method.

REFERENCES

1. *Boltyanskiĭ V. G.* Mathematical Methods of Optimal Controls [in Russian]. Moscow: Nauka, 1969.

2. *Fedorenko R. P.* Approximate Solution of Optimal Control Problems [in Russian]. Moscow: Nauka, 1978.

3. *Demidovich V. B., Maron I. A., and Shuvalova E. Z.* Numerical Methods in Analysis [in Russian]. Moscow: Nauka, 1967.

December 5, 2014

V. G. Starov
Sobolev Institute of Mathematics, Novosibirsk, Russia
`vladalex@math.nsc.ru`

Yakutian Mathematical Journal
July—September, 2014. Vol. 21, No. 3

UDC 621.89:536.24

NUMERICAL DETERMINATION OF THE TEMPERATURE FIELD IN A SYSTEM OF BEARINGS ON THE SAME SHAFT WITH ACCOUNT FOR ITS SPEED

R. S. Tikhonov and N. P. Starostin

Abstract. We propose a quasi-three-dimensional mathematical model of thermal process in a system of bearings on the same shaft with account for its speed. We give the results of determining the time step in a numerical solution of the problem by the finite difference method, studying the mutual influence of temperature fields in the system of bearings, and finding conditions under which the mathematical model could simplify.

Keywords: plain bearing, mathematical model, friction, heat conductivity, temperature, heat equation, numerical solution

Introduction

The mathematical modeling of nonstationary thermal processes in systems of bearings on the same shaft is used to solve various applied problems. This includes the method of thermal diagnostics of friction, which enables us to determine the heat emission caused by friction and, accordingly, the friction torques from temperature measurements by solving an inverse boundary value problem [1]. The temperature field in a system of bearings was previously determined on assuming the uniform distribution of temperature across the shaft as a consequence of sufficiently fast rotation (more 5 rad/s) and treating the shaft as a one-dimensional rod [1–4]. The angular velocities of the shaft in plain bearings for which calculating the temperature field requires for the motion of the shaft was determined in [5]. The influence of the rotation of the shaft on the temperature field was factored for one plain bearing in the flat case in [6], and in the three-dimensional case in [7].

In this article we consider the problem of determining the nonstationary temperature field in a system of bearings on the same slowly rotating shaft, for which we should account for the motion of the shaft (Fig. 1).

The mathematical model of thermal process in a system of bearings on the same shaft is a generalization of the model for one bearing. A planar mathematical model for a plain bearing taking into account the clearance between the shaft and the hub and an algorithm for a numerical solution of the problem of nonstationary heat exchange are proposed in [8]. We can only obtain a realistic picture of heat propagation in a bearing when accounting for the spatial distribution of heat arising from friction. In thermal diagnostics of friction with the use of a complete three-dimensional model of the thermal process it is necessary to specify temperature on some plane in a neighborhood of the friction zone, which is impossible in practice. For this reason, we construct a quasi-three-dimensional mathematical model.

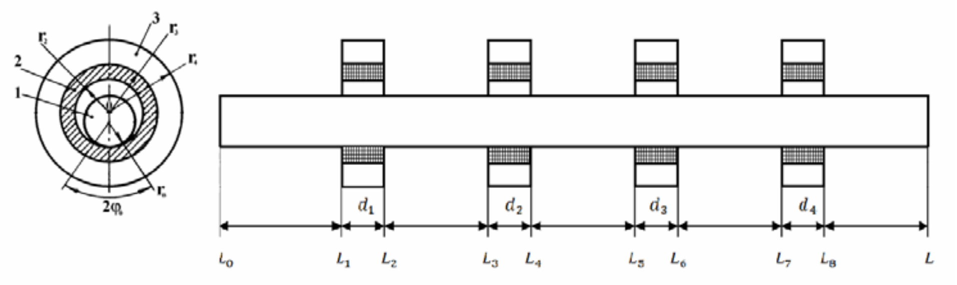


Fig. 1. The scheme of calculation for a system of plain bearings: shaft 1; hub 2; race 3

Assume that the temperature distribution is uniform along the length of the bearing and cage since the heat transfer at their end surfaces is insignificant. Indeed, calculations using the full three-dimensional model of the thermal process in a plain bearing show that the convective heat transfer from the end surfaces of the hub and cage amounts to less than 0.5% of thermal friction. Therefore, we may regard the hub and cage as flat, while the shaft as three-dimensional.

The nonstationary temperature field in a bearing is described by the two-dimensional quasi-linear heat equation for hubs with cages:

$$C_{ik}\frac{\partial T}{\partial t} = \frac{1}{r}\frac{\partial}{\partial r}\left(r\lambda_{ik}\frac{\partial T_k}{\partial r}\right) + \frac{1}{r^2}\frac{\partial}{\partial \varphi}\left(\lambda_{ik}\frac{\partial T}{\partial \varphi}\right),$$
$$R_{2k} < r < R_{4k}, \quad -\pi < \varphi < \pi, \quad 0 < t \le t_m, \quad k = 1, \dots, N, \ i = 2, 3, \tag{1}$$

while for the shaft, by the three-dimensional equation with the convective term accounting for its rotation:

$$C_1\frac{\partial U}{\partial t} = \frac{1}{r}\frac{\partial}{\partial r}\left(r\lambda_1\frac{\partial U}{\partial r}\right) + \frac{1}{r^2}\frac{\partial}{\partial \varphi}\left(\lambda_1\frac{\partial U}{\partial \varphi}\right) + \Omega C_1\frac{\partial U}{\partial \varphi} + \frac{\partial}{\partial z}\left(\lambda_1\frac{\partial U}{\partial z}\right),$$
$$0 < r < R_1, \quad -\pi < \varphi < \pi, \quad 0 < t \le t_m. \tag{2}$$

In the zone of friction between the hubs and the shaft, impose the conditions of frictional heat emission:

$$\lambda_1\frac{\partial U(r,\varphi,z,t)}{\partial r}\bigg|_{r=R_1} - \lambda_2\frac{\partial T_k(r,\varphi,t)}{\partial r}\bigg|_{r=R_{2k}} = Q_k(\varphi,t), \quad |\varphi| \le \varphi_0, \tag{3}$$

$$\frac{1}{d_k}\int\limits_{z_{k-1}}^{z_k} U(R_1,\varphi,z,t)\,dz = T_k(R_{2k},\varphi,t). \tag{3$'$}$$

On the free surfaces of the shaft, hubs, and cages, impose the conditions of convective heat exchange:

$$\lambda_1\frac{\partial U(r,\varphi,z,t)}{\partial r}\bigg|_{r=R_1} = -\alpha_1(U(R_1,\varphi,z,t) - T_0), \tag{4}$$

$$\lambda_{2k}\frac{\partial T_k(r,\varphi,t)}{\partial r}\bigg|_{r=R_2} = \alpha_2(T_k(R_{2k},\varphi,t) - T_0), \quad |\varphi| > \varphi_0, \tag{5}$$

$$\lambda_{3k}\frac{\partial T_k(r,\varphi,t)}{\partial r}\bigg|_{r=R_{4k}} = \alpha_3(T_k(R_{4k},\varphi,t) - T_0), \quad -\pi < \varphi \le \pi. \tag{6}$$

On the ends of the shaft, impose conditions of the first and third kind:

$$\lambda_1 \frac{\partial U(r,\varphi,z,t)}{\partial z}\bigg|_{z=L} = -\alpha_1(U(R_1,\varphi,L,t) - T_0), \quad U(R_1,\varphi,0,t) = T_0. \tag{7}$$

In the center of the shaft, impose the condition that the thermal flow is bounded:

$$\lim_{r\to 0}\left(r\lambda_1 \frac{\partial U}{\partial r}\right) = 0. \tag{8}$$

There are periodicity conditions with respect to the angular coordinates:

$$\frac{\partial T_k(r,\varphi,t)}{\partial \varphi}\bigg|_{\varphi=-\pi} = \frac{\partial T_k(r,\varphi,t)}{\partial \varphi}\bigg|_{\varphi=\pi}, \quad T_k(r,-\pi,t) = T_k(r,\pi,t), \tag{9}$$

$$\frac{\partial U(r,\varphi,z,t)}{\partial \varphi}\bigg|_{\varphi=-\pi} = \frac{\partial U(r,\varphi,z,t)}{\partial \varphi}\bigg|_{\varphi=\pi}, \quad U(r,-\pi,z,t) = U(r,\pi,z,t). \tag{10}$$

Assume that the initial distribution of temperature in the elements of the friction assembly are equal and uniform:

$$T_k(r,\varphi,0) = U(r,\varphi,z,0) = T_0. \tag{11}$$

1. Numerical Solution

To solve problem (1)–(11), we used the finite difference method applicable to equations with variable coefficients, nonlinear equations, and widely used to solve various applied problems [9–15]. The presence of the convective term in the heat equation (2), for the rotating shaft, causes certain difficulties for the numerical solution of the stated problem. The use of monotone locally one-dimensional difference schemes of total approximation guarantees the fulfillment of the maximum principle, that is, for arbitrary steps τ and h_φ with respect to temporal and angular variables we can define an approximate solution. From the set of approximate solutions choose a solution satisfying the sticking condition as the temporal step tends to 0. We intend to use the developed algorithm for calculating the temperature field to solve boundary value problems and implement an algorithm for solving the inverse heat exchange problem of determining frictional heat emission. Therefore, the expenses of computer time to solve the inverse problem depend on the temporal step in the numerical solution of the direct problem. For an excessively small step the expenses of computer time can turn out prohibitive for practical calculations. Thus, we should choose the temporal step as large as possible for the specified sticking criterion.

Suppose that we have a sufficiently fine spatial mesh. For this mesh, run simulations to determined the temporal step which ensures that the solution converges. We varied the Courant number ($\gamma = \tau\upsilon/h_\varphi$, where $\upsilon = R_1\Omega$) characterizing the relation of the angular step to the angular velocity of the shaft and the temporal step. Since all plain bearings share the shaft, it suffices to consider the case of one bearing. Subsequently, we use the temporal step found this way to calculate temperature in a system of several bearings.

Fig. 2 shows the results of temperature calculations in plain bearings in dependence on the Courant number for the following geometric sizes: $R_{1k} = 12$ mm, $R_{2k} = 13$ mm, $R_{3k} = 16$ mm, $R_{4k} = 12$ mm, where $k = 1$. The shaft and cage are made from steel, while the hub is made of the filled fluoropolymer Φ4K20. The angular velocity of the shaft is π rad/s, and the angle of contact is 30°. The specific intensivity of heat emission is constant and equals $Q = 67\text{kW/m}^2$. Calculations show that for $\gamma < 1$ the solutions converge. For practical calculations, we can define the temporal step from the condition $\gamma = 2$ since for $\gamma < 2$ the values of temperature vary within one degree.

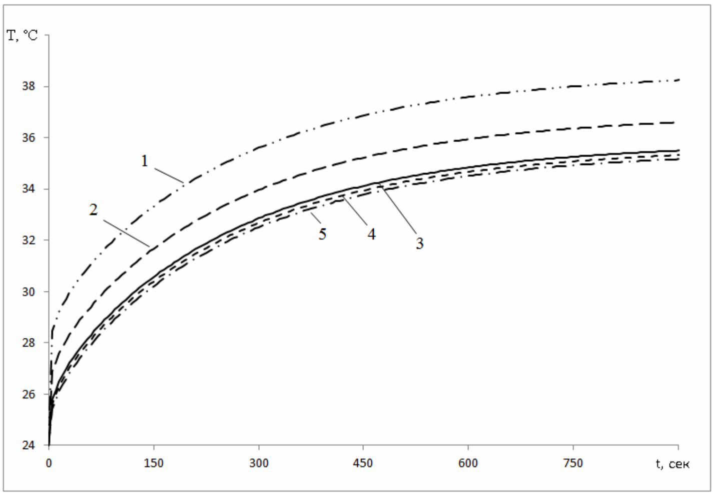


Fig. 2. Calculation dependencies of the maximal temperature in the friction zone for different Courant numbers γ: 1 is $\gamma = 36$; 2 is $\gamma = 12$; 3 is $\gamma = 2$; 4 is $\gamma = 1$; 5 is $\gamma = 1/8$

2. Comparison of Calculated and Experimental Temperature

To establish whether our mathematical model with two-dimensional and three-dimensional heat equations is adequate to the real thermal process in plain bearings, we compared the simulated and experimental temperature (Fig. 3).

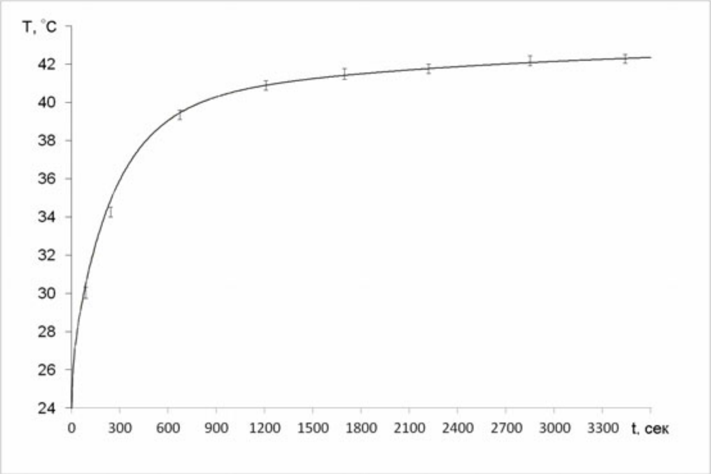


Fig. 3. Calculated dependence of temperature on time in an inner point of the hub at 0.5 mm from the friction zone. (The symbol I marks the confidence intervals of experimental temperature data)

We made experiments on a friction machine CMT-1 for one bearing, recording the temperature with copper-constantan thermocouples of diameter 0.1 mm and the use of multichannel “TERMODAT” device at five points of a hub at 0.5 mm away from the friction zone. The angle of contact was 60°.

Repeating the experiment three times, we determined a confidence interval for the dependence of temperature on time. Fig. 3 shows that the calculated dependence of temperature for $\varphi = 0$ lies within the range of experimental data. We obtained similar results for other points of temperature measurements, which confirms that our description of the thermal process in a plain bearing under consideration is adequate.

The calculations resulted in determining the effective coefficients of the heat equation ensuring that the simulated and experimental temperatures are close. The difference between the effective coefficients and the handbook values of thermophysical characteristics of materials is not greater than the range of properties of the polymer composition and steel. The effective heat conductivity coefficient λ_2 of the filled fluoropolymer Φ4K20 is 0.34 W/(m $\cdot\,{}^\circ C$), the spatial heat capacity C_2 equals $2.2{\cdot}10^6$ J/(m$^3\cdot{}^\circ C$), while for steel $\lambda_1 = 46$ W/(m$\cdot{}^\circ C$) and $C_1 = 3.48{\cdot}10^6$ J/(m$^3\cdot{}^\circ C$).

We calculated the heat exchange coefficient of the rotating shaft as [16]

$$\alpha_1 = Nu\frac{\lambda_{voz}}{2R_1}, \quad Nu = 0.95(2Re^2 + Cr)^{0,35}. \tag{12}$$

The comparison of temperature data indicates that the proposed mathematical model is applicable for determining the temperature field in plain bearings.

3. Modeling Thermal Processes in a System of Bearings

We generalized our algorithm to systems of plain bearings. As an example we took a system of four identical bearings made of the filled fluoropolymer Φ4K20. We specified heat emission intensivity in the bearings as functions of time:

$$\begin{gathered} Q_1(\varphi,t) = 3851,656(t+1)^{\frac{1}{4}}|\cos\varphi|, \quad Q_2(\varphi,t) = 2^{15,89-\frac{(t-600)^2}{90000}}|\cos\varphi|, \\ Q_3(\varphi,t) = 3851,656\pi|\cos\varphi|, \quad Q_4(\varphi,t) = 12^{4,3194-\frac{(t-600)^2}{360000}}|\cos\varphi|. \end{gathered} \tag{13}$$

These functions are chosen so that the first of them increases, the third is constant, while the second and fourth have a point of maximum on the tenth minute. The dynamics of temperature field in each bearing affects the dependence of temperature on time in neighboring bearings, attested by the change of temperature in time in the friction zone of bearings (Fig. 4). As heat emission in the first bearing increases, the temperature in the friction zone must rise. Under the influence of the decrease in heat emission in the second bearing after 10 minutes, the temperature in the first bearing drops after 15 minutes of operation. In the third bearing the temperature should settle due to constant heat emission, but under the influence of the decrease in temperature the fourth bearing it also begins to drop.

Studying temperature fields in plain bearings, we determined the distance between bearings of 10 cm which exclude the mutual influence on the change of temperature fields of the bearings in time. The results of calculations of this sort can be used, for instance, to develop multiposition station for testing friction and wear of materials.

In the thermal diagnostics of friction, based on solving the inverse boundary value problem of heat exchange, the expenses of computer time on the direct problem are most important. We can decrease computation time by simplifying the mathematical model. As for the mathematical model of thermal process in a system

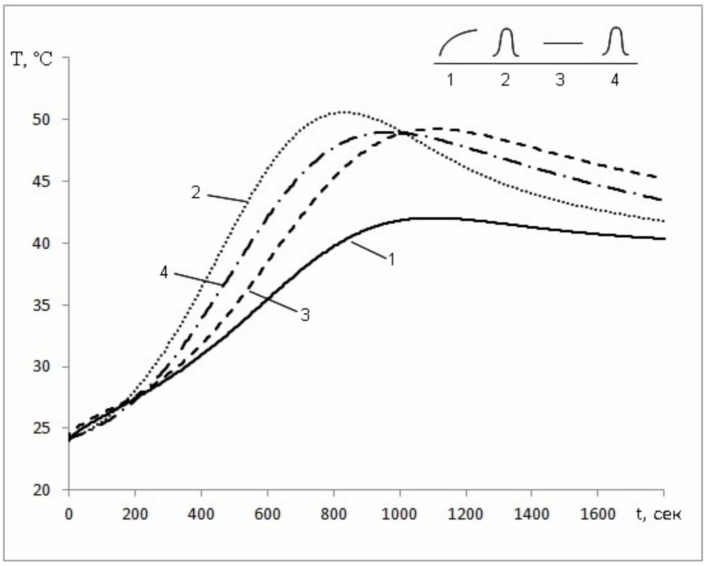


Fig. 4. Dependence of the maximal temperature in the contact zone of plain bearings on time: 1 is the temperature of the 1st bearing, 2 is the 2dn bearing, 3 is the 3rd bearing, 4 is the 4th bearing

of bearings, we may assume that the temperature is uniform across the shaft, which enables us to consider the shaft as one-dimensional and account for the rotation of the shaft only in the coefficients of heat exchange of its surface with the surrounding medium. To establish the validity of this assumption along with analyzing the temperature distribution in the radial variable, it is necessary to study the change of temperature in the shaft in circles. Suppose that the intensivity of heat emission, diameter, and thermophysical properties of the shaft allow us to assume that the temperature distribution in the radial variable is uniform. As the rotation of the shaft speeds up, the change of temperature along the circle tends to a uniform distribution. Let us determine the angular velocity of the shaft above which we can assume that the temperature distribution is uniform.

At a short distance from the plain bearing (2-4 mm) along the axial variable the temperature distribution in the shaft along the circle becomes uniform. For this reason, to study the uniformity of temperature in the circular variable in a system of plain bearings, we considered one bearing with the initial data chosen above. In the simulations for various angular velocities the function of specific intensivity of heat emission remained unchanged thanks to the condition $pR_1\Omega = pV = \text{const}$. Temperature reaches its maximal and minimal values in the friction zone of the shaft at the points where the surface of the shaft come into and out of contact. Fig. 5 shows the calculated dependence of the surface temperature of the shaft on the angular coordinate. Despite the increase of the heat exchange coefficient with faster rotation, after coming out of contact the surface of the shaft is cooled less and the minimal temperature of contact increases. In addition, owing to the shorter duration of contact, the maximal temperature on the surface of the shaft decreases. Thus, the increase of angular velocity leads to a more uniform temperature distribution

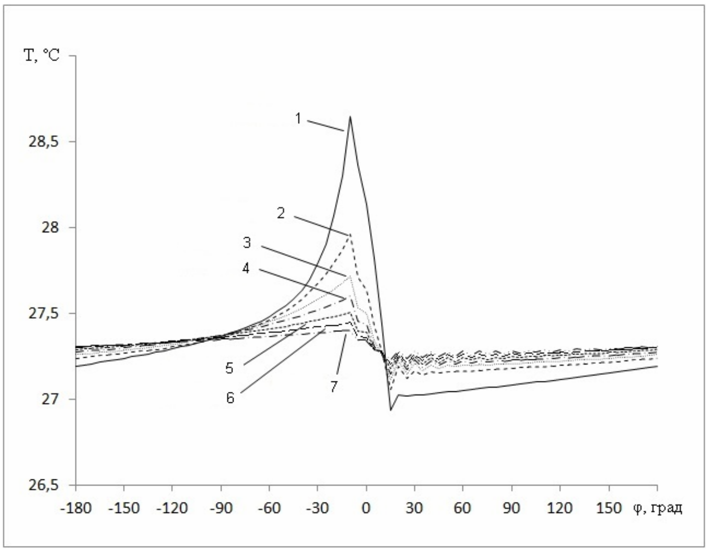


Fig. 5. Temperature distribution on the surface of the shaft for various angular velocities: 1 is 0,1π rad/s; 2 is 0,3π rad/s; 3 is 0,5π rad/s; 4 is 0,7π rad/s; 5 is π rad/s; 6 is 1,4π rad/s; 7 is 2π rad/s

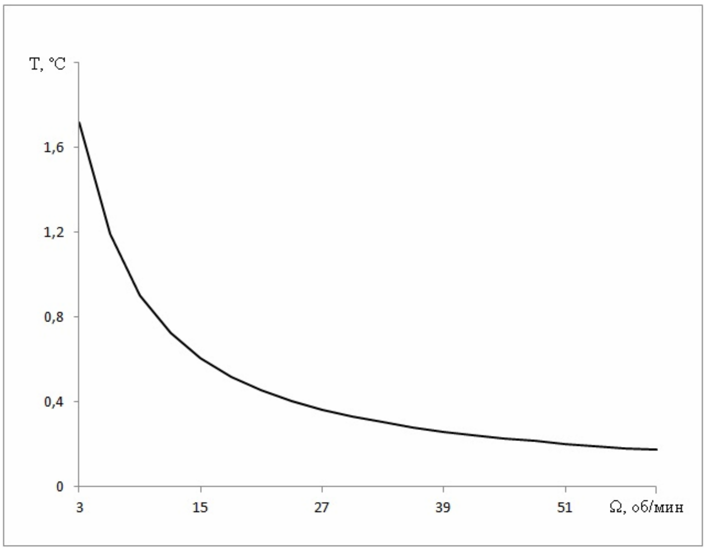


Fig. 6. Dependence of the difference between the extremal values of temperature on the surface of the shaft on the angular velocity of the shaft at time $t = 1$ min

along the circle.

We make the simplifying assumptions in the mathematical model of thermal process in a system of bearings basing on the accuracy of the description of the process required by the objectives of research. Fig. 6 presents the difference between the extremal values of temperature on the surface of the shaft in dependence on the angular velocity. This dependence enables us to determine the angular velocity of the shaft above which we may assume that the temperature distribution is uniform with certain accuracy.

4. Conclusions

We proposed a mathematical model of thermal process in a system of plain bearings for the rotation of the shaft and a technique enabling us to determine, basing on simulations, a temporal step suitable for practical calculations.

By comparing the simulated and experimental temperature, we established that this mathematical model is adequate to the real thermal process in a plain bearing.

By studying temperature fields in plain bearings, we determined the distance between bearings which excludes their mutual thermal influence and the angular velocity above which our mathematical model simplifies.

We can recommend the proposed mathematical model of thermal process for determining friction in each bearing of the system from temperature data.

Notation: Q_k is the specific intensivity of heat emission in the contact zone of bearing k; U is the temperature of the shaft; T_k is the temperature of bearing k; T_0 is the temperature of the surrounding medium; k and N are the index and the number of bearings; R_1 is the radius of the shaft, R_{2k} and R_{3k} are the inner and outer radii of the hub of bearing k; R_{4k} is the outer radius of the cage of bearing k; d_k is the length of bearing k; L is the length of the shaft; r, φ, z are cylindrical coordinates; z_{k-1} and z_k are axial coordinates of the ends of bearing k; $2\varphi_0$ is the contact angle; Ω is the angular velocity; t is time; t_m is the trial time; C_1 is the spatial heat capacity of the material of the shaft; C_{2k} and C_{3k} are the spatial heat capacities of the materials of the hub and cage of bearing k respectively; α_1, α_2, and α_3 are the heat transfer coefficients from the surfaces of the shaft, hub, and cage respectively; λ_1 is the heat conductivity coefficient of the material of the shaft; λ_{2k} and λ_{3k} are the heat conductivity coefficients of the materials of the hub and cage of bearing k respectively; λ_{voz} is the heat conductivity coefficient of the surrounding medium; Re is the Reynolds' criterion; Cr is the Grashof's criterion; p is pressure; V is linear velocity.

REFERENCES

1. *Starostin N. P., Tikhonov A. G., Morov V. A., and Kondakov A. S.* Calculation of the Tribotechnical Parameters with Sliding Support [in Russian]. Yakutsk: Izdat. YaNTs SO RAN, 1999.
2. *Cherskiĭ I. N., Bogatin O. B., and Starostin N. P.* Reconstruction of the friction moments in a system of nongreased bearings from measurements of temperature // Trenie Iznos. 1986. V. VII, N 5. P. 878–887.
3. *Bogatin O. B., Starostin N. P., Cherskiĭ I. N., et al.* Experimental evaluation of the efficiency of reconstructing the friction moment in a system of nongreased bearings from measurements of temperature // Trenie Iznos. 1991. V. 12, N 3. P. 442–445.
4. *Starostin N. P.* Mathematical modeling of the heat regime and temperature diagnostics of the friction in a system of cylindrical plain bearings // Mat. Zametki YaGU. 1997. V. 4, N 2. P. 161–170.
5. *Vasil′eva M. A., Kondakov A. S., and Starostin N. P.* Study of the applicability of simplified models of heat processes in radial plain bearings based on numerical experiments // Mat. Zametki YaGU. 2008. V. 15, N 2. P. 84–91.
6. *Starostin N. P., Kondakov A. S., and Vasil′eva M. A.* Thermal diagnostics of friction in plain bearings with account for speed and mode of shaft motion // J. Friction and Wear. 2012. V. 33, N 5. P. 330–337.

7. *Kondakov A. S., Starostin N. P., and Vasil′eva M. A.* A three-dimensional inverse boundary value problem of heat diagnostics of the friction of plain bearings // Mat. Zametki YaGU. 2012. V. 19, N 2. P. 187–195.

8. *Cherskiĭ I. N., Bogatin O. B., and Borisov A. Z.* Analysis of the temperature field of a polymer plain bearing in a nonstationary friction period // Trenie Iznos. 1981. V. 2, N 2. P. 231–238.

9. *Samarskiĭ A. A.* The Theory of Difference Schemes. Moscow: Nauka, 1983.

10. *Butt M. M. and Taj M. S. A.* Numerical methods for heat equation with variable coefficients // Int. J. Computer Math. 2009. V. 86, N 9. P. 1612–1623.

11. *Shih T.-M., Sung Ch.-H., and Yang B.* A numerical method for solving nonlinear heat transfer equations // Numerical Heat Transfer. Part B: Fundamentals. 2008. V. 54, N 4. P. 338–353.

12. *Sankar M., Park J., Kim D., and Do Y.* Numerical study of natural convection in a vertical porous annulus with an internal heat source: Effect of discrete heating // Numer. Heat Transfer. Part A: Applications. 2013. V. 63, N 9. P. 687–712.

13. *Fu W.-Sh. and Tong B.-H.* Numerical investigation of heat transfer of a heated channel with an oscillating cylinder // Numer. Heat Transfer. Part A: Applications. 2003. V. 43, N 6. P. 639–658.

14. *Liu Ch.-Sh.* An iterative method to recover the heat conductivity function of a nonlinear heat conduction equation // Numer. Heat Transfer. Part B: Fundamentals. 2014. V. 65, N 1. P. 80–101.

15. *Nabongo D. and Boni T. K.* Numerical quenching for a semilinear parabolic equation // Math. Modeling Anal. 2008. V. 13, N 4. P. 521–538.

16. *Dropkin D. and Karmi A.* Natural-convection heat transfer from a horizontal cylinder rotating in air // Trans. ASME. 1957. V. 79, N 4. P. 741–749.

November 14, 2014

R. S. Tikhonov; N. P. Starostin
Institute of Oil and Gas Problems, Yakutsk, Russia
roman_tikhon@mail.ru; nikstar56@mail.ru