

ARCHIVE B

It contains some information on initial versions of older working papers, technical reports, and other miscellanea.

This file is rarely updated.

For the up-to-date information, visit <http://m3ai.wlu.ca> or contact us.

The Luttinger-Kohn theory for multiband Hamiltonians: A revision of ellipticity requirements

Dmytro Sytnyk*

*Numerical Mathematics Department, Institute of Mathematics, National Academy of Sciences, Ukraine,
M²NeT Laboratory, Wilfrid Laurier University, 75 University Avenue West, Waterloo, ON, Canada, N2L 3C5.*

Roderick Melnik†

M²NeT Laboratory, Wilfrid Laurier University, 75 University Avenue West, Waterloo, ON, Canada, N2L 3C5.

Modern applications require a robust and theoretically solid tool for the realistic modeling of electronic states in low dimensional nanostructures. The $k \cdot p$ theory has fruitfully served this role for the long time since its establishment. During the last three decades several problems have been detected in connection with the application of the $k \cdot p$ approach to such nanostructures. These problems are closely related to the violation of the ellipticity conditions for the underlying model, the fact that has been largely overlooked in the literature. We derive ellipticity conditions for 6×6 , 8×8 and 14×14 Hamiltonians obtained by the application of Luttinger-Kohn theory to the bulk zinc blende (ZB) crystals, and demonstrate that the corresponding models are non-elliptic for many common crystalline materials. With the aim to obtain the admissible (in terms of ellipticity) parameters, we further develop and justify a parameter rescaling procedure for 8×8 Hamiltonians. This allows us to calculate the admissible parameter sets for GaAs, AlAs, InAs, GaP, AlP, InP, GaSb, AlSb, InSb, GaN, AlN, InN. The newly obtained parameters are then optimized in terms of the bandstructure fit by changing the value of the inversion asymmetry parameter B that is proved to be essential for ellipticity of 8×8 Hamiltonian. The consecutive analysis, performed here for all mentioned $k \cdot p$ Hamiltonians, indicates the connection between the lack of ellipticity and perturbative terms describing the influence of out-of-basis bands on the structure of the Hamiltonian. This enables us to quantify the limits of models' applicability material-wise and to suggest a possible unification of two different 14×14 models, analysed in this work.

PACS numbers: 71.20.-b, 71.20.Nr, 73.22.-f, 31.15.xp, 02.30.Jr

I. INTRODUCTION

The collection of methods known as an effective mass theory is one of the fundamental topics in the physics of nanostructures. The theory has been used to describe a wide variety of physical phenomena ranging from the formation of electronic bands in periodic solids to the realistic field-mater interaction in modern semiconductor materials. Furthermore, the theory establishes a robust computational framework for simulating observable quantum-mechanical states and corresponding energies in the low-dimensional nanoscale systems, including quantum wells, wires and dots.

In the original Luttinger-Kohn work¹ authors applied the perturbation theory to the Schrödinger equation with a smooth potential and constructed a representation for valence bands Hamiltonian near the high symmetry point Γ of the first Brillouin zone in bulk zinc blende (ZB) crystals with large fundamental bandgap. Soon after that, Kane showed how to extend the model to the narrow gap materials such as InSb and Ge for instance, where one can also account for the influence of the conduction bands². One of the advantages of the $k \cdot p$ theory is in its universality and flexibility when it comes to simulation of electronic transport phenomena in the presence of electromagnetic and/or thermoelastic fields^{3,4}. Indeed, the theory had also been extended to cover Wurtzite (WZ) type of crystals, materials with inclusions, heterostruc-

ture materials and superlattices^{5–7}. Another advantage of the effective mass theory is its flexibility, as one can easily adjust the models to include additional effects like strain⁸, piezoelectricity, magnetic field, and respective nonlinear effects. These inbuilt multiscale effects are crucial for such applications as light-emission diodes, lasers, high precision sensors, photo-galvanic elements, hybrid bio-nanodevices, and many others⁹.

For a wide range of applications the Luttinger-Kohn models have provided good, computationally feasible and efficient approximations that agree well with experimental results^{10,11}. However, for some types of crystal materials band structure calculations based on such multiband models lead to the solutions with unphysical properties^{12,13} or so called spurious solutions^{14–19}.

As a result, there have been various attempts to explain the origin of the spurious solutions and develop some reliable procedures on how to avoid them^{16,19–21}. These approaches rely on three main ideas: (a) to modify the original Hamiltonian and remove the terms responsible for the spurious solutions^{15,22}, (b) to change bandstructure parameters^{14,20,23}, and (c) to identify and exclude from simulations the physically inadequate observable states^{11,24} or change the numerical scheme to avoid such states altogether^{21,25}. All mentioned approaches suffer from the common weakness – the lack of clear justification of the underlying theoretical procedure and thus from limitations in their applicability^{16,17}.

In this work we show that spurious solutions are just a consequence of a more fundamental problem in applications of the effective mass theory: the non-ellipticity of the multiband Hamiltonian in the position representation.

The systematic study of connection between the structure of 6×6 , 8×8 and 14×14 Hamiltonians, their ellipticity in the position representation and the material parameters for ZB crystals allow us to conclude that the widely adopted $k \cdot p$ models turn out to be non-elliptic (hyperbolic) for a broad class of known material parameters. The phase space of the hyperbolic model is wider than the spaces of norm-bounded observable states. Such models, therefore, are susceptible to unphysical solutions, even in the bulk case. Meanwhile, the corresponding time-dependent Schrödinger equation loses the fundamental property of state conservation²⁶.

These facts lead to an important assertion. Since any qualitative multiband approximation of Schrödinger Hamiltonian must preserve its core physical properties, such as ellipticity and, as a consequence, semi-boundedness of set of energy states; the lack of ellipticity for certain materials implies that the usage of multiband Hamiltonian for such materials is fundamentally incorrect. This results in substantial ramifications for the applications of effective-mass theory to bulk solids and heterostructures.

The whole procedure of obtaining the materials parameters from experiment and their incorporation into mathematical models of effective mass theory needs to be revisited, taking into account the general ellipticity constraints derived in the present work. Before this is done, we propose here the sets of elliptic Hamiltonian parameters for GaAs, AlAs, InAs, GaP, AlP, InP, GaSb, AlSb, InSb, GaN, AlN, InN, optimized in terms of the bandstructure fit. We also supply a parameter rescaling procedure used to obtain these sets from the available non-elliptic parameters.

The paper is organized as follows. First, we revise basic properties of the Schrödinger equation and its approximations represented by $k \cdot p$ models. In section III we outline a mathematical procedure to obtain the ellipticity constraints for a Hamiltonian in the position representation. For the $k \cdot p$ Hamiltonians the constraints are comprised of the set of linear material-dependent inequalities²⁷. In sections III and IV we present a direct evaluation of ellipticity constraints for 6×6 and 8×8 ZB Hamiltonians based on parameter sets gathered from major material-data sources^{28–32}. Most of the 53 analyzed parameter sets lead to the failure of the Hamiltonians' ellipticity. That is why the main part of section IV is devoted to a parameter rescaling procedure aimed at correcting the Hamiltonian's ellipticity. As a result of the procedure we propose elliptic parameter sets for all analyzed materials. The newly obtained sets are, then, compared to the original materials parameters by means of the differences in the associated bandstructures of 8×8 Hamiltonian. In this section, we further extend

the ellipticity conditions of 8×8 model³³ to the case of nonzero inversion-asymmetry parameter B . Afterwards, B is used to improve the bandstructure fit of the proposed elliptic parameter set for indium nitride. Section V is devoted to the ellipticity analysis of two existing 14×14 models^{34,35}.

The summary of results together with discussions on applicability and future directions are given in the concluding section.

II. OVERVIEW OF LUTTINGER-KOHN BANDSTRUCTURE THEORY

The material properties (such as fundamental bandgaps and spin-orbit splitting energies) obtained experimentally, represent real quantum phenomena, whereas models based on multiband Hamiltonians are meant to approximate them. As such these models are derived from the stationary Schrödinger equation that represents an averaged charge carrier interactions in the crystalline structure^{36,37}. The derivation scheme involves the application of Bloch wave representation and the projection of the original Hamiltonian to the orthogonal subspace of the reduced phase space^{1,38}. The projective part of Hamiltonian is then adjusted with help of perturbation theory^{1,38,39} to account for the influence of outer bands. However, this last step lacks a rigorous theoretical foundation as it does not guarantee the convergence of the perturbative expansion^{40,41}. The result is that the derived Hamiltonian, although directly based on the experimental parameters (Tables I, II, V), represents a totally different mathematical object compared to its origin. The physical evidence, to support this claim has been already known for GaAs⁴² and for Si⁴³.

We start with the Schrödinger equation

$$H_0\psi(x) = \left(\frac{\mathbf{p}^2}{2m_0} + V(x) + H_{SO} \right) \psi(x) = E_n\psi(x), \quad (1)$$

where $\mathbf{p} = i\hbar\nabla$ is a momentum operator of charge carrier with the mass m_0 , $V(x)$ is the effective potential, $x \in \Omega \subset \mathbb{R}^3$. The unknown E_n stands for the eigenenergy of the system and the function $\psi(x)$ is the corresponding eigenstate. The Hamiltonian H_{SO} accounts for relativistic effects of spin.

In the finite domain Ω we supplement (1) by the boundary conditions

$$\psi(x) = f(x), \quad x \in \partial\Omega, \quad (2)$$

assuming that the combination of given Ω and $f(x)$ endows operator H_0 with all necessary properties, postulated by the standard axiomatic approach to quantum mechanics⁴⁴. The operator H_0 is an elliptic partial differential operator. It is symmetric over its domain of definition $D(H_0) \subset \mathcal{H}^3(\Omega)$. Furthermore we require that the boundary $\partial\Omega$ is sufficiently smooth, so that a self-adjoint extension of H_0 exists and possesses the property of the

probability current conservation^{44,45}. All mentioned assumptions can be satisfied in the bulk case⁴⁶, which will be our main focus throughout the work.

If $V(x)$ is a gently varying function over the unit cell¹, the original operator H_0 can be approximated by another operator H (using Bloch theorem), determined by the projection P of H_0 on the considered eigenspace and Löwdin perturbation theory^{1,38}. The last step in this approximation procedure accounts for the influence of the elements from the space (so-called class of states B) complement to the chosen eigenspace (so-called class of states A) by the formula

$$H = PH_0 + \sum_{i=1}^r \delta^r H^{(r)} \quad (3)$$

up to the order r . Setting $\delta = 1$ leads one to the final approximation, under the assumption that the series (3) is convergent for such δ . Despite wide applicability of such

approximations, the intrinsic ellipticity requirements for the realizations of H have not been explicitly verified in a systematic manner (see 1, 2, 20, 38, and 47, as well as more recent works^{14,15,18,22,25,34,35}). The only known to us work where it has been done for the case of InAs, GaAs and Al_{0.3}Ga_{0.7}As is [16]. Hence, in what follows we analyze such requirements systematically for all common 6×6 , 8×8 and 14×14 ZB Hamiltonians.

III. SIX-BANDS HAMILTONIAN ANALYSIS

This section is devoted the ellipticity analysis of the classical 6×6 Hamiltonian for ZB¹ type of crystals, demonstrating our approach in detail. In this work we use the Luttinger parameter notation which is common in recent works on the subject. When necessary, the parameters will be converted from other parameter notations³⁷.

The Luttinger-Kohn (LK) Hamiltonian is defined as follows^{38,39}

$$H^{LK} = \begin{pmatrix} P+Q & S & R & 0 & -\frac{1}{\sqrt{2}}S & -\sqrt{2}R \\ S^* & P-Q & 0 & R & \sqrt{2}Q & \sqrt{\frac{3}{2}}S \\ R^* & 0 & P-Q & -S & \sqrt{\frac{3}{2}}S^* & -\sqrt{2}Q \\ 0 & R^* & -S^* & P+Q & \sqrt{2}R^* & -\frac{1}{\sqrt{2}}S^* \\ -\frac{1}{\sqrt{2}}S^* & \sqrt{2}Q & \sqrt{\frac{3}{2}}S & \sqrt{2}R & P-\Delta_{SO} & 0 \\ -\sqrt{2}R^* & \sqrt{\frac{3}{2}}S^* & \sqrt{2}Q & -\frac{1}{\sqrt{2}}S & 0 & P-\Delta_{SO} \end{pmatrix} \quad \begin{aligned} P &= -\frac{\hbar^2}{2m_0}\gamma_1 \mathbf{k}^2, \\ Q &= -\frac{\hbar^2}{2m_0}\gamma_2(k_x^2 + k_y^2 - k_z^2), \\ R &= -\frac{\hbar^2}{2m_0}\frac{-\sqrt{3}}{2}[(\gamma_2 + \gamma_3)k_-^2 + (\gamma_2 - \gamma_3)k_+^2], \\ S &= -\frac{\hbar^2}{2m_0}(-2\sqrt{3})\gamma_3 k_- k_z, \end{aligned}$$

where $\mathbf{k}^2 = k_x^2 + k_y^2 + k_z^2$, $k_\pm = k_x \pm ik_y$. Each of the P, Q, R, S is a second order position dependent differential operator in the position representation or, equivalently, second order polynomial in the momentum representation¹.

Our aim is to check the type (elliptic, hyperbolic or essentially hyperbolic) of the H^{LK} as a partial-differential operator (PDO), keeping in mind that the Schrödinger operator from (1) is elliptic. Only the second order derivative terms are playing the dominant role in the following analysis because contributions from the terms linear in the components of \mathbf{k} as well as from the potential, are bounded in the domain $D(H^{LK})$ ⁴⁸. It means that the results for more complicated physical models with potential contributions from additional fields (e.g. strain, magnetic field, etc.) will stay the same as for the original H^{LK} , analyzed here. The fact that the Hamiltonian is a linear operator guarantees that it is also true for any other representation of H^{LK} obtained by linear (basis) transformations.

In a more general sense, for any m -dimensional matrix PDO $H = \{h_{ij}\}_{i,j=1}^m$, where each element h_{ij} is a second

order one-dimensional PDO^{49,50}

$$h_{ij} = \sum_{k,l=0}^n h_{ij}^{kl} \frac{\partial^2}{\partial x_k \partial x_l}, \quad (4)$$

the associated quadratic form (also known in the mathematical literature as a principal symbol) is defined by

$$G(\xi_1, \dots, \xi_m) = v M v^T, \quad v = (\xi_1, \dots, \xi_m), \quad (5)$$

where M is an $m \times m$ matrix composed from the elements h_{ij}^{kl} . The $\mathbf{k} \cdot \mathbf{p}$ Hamiltonians in \mathbb{R}^n are a special case of (4). They are symmetric as a matrix PDO so the associated quadratic form G will have M with only real eigenvalues λ_i (e. g. 16).

Using these notations, the procedure of obtaining the ellipticity condition for H is reduced to the question about the sign of λ_i for the associated M . More precisely, the matrix differential operator H will be elliptic if and only if all eigenvalues of the corresponding Hermitian M will have the same sign^{48,49}.

In general, it is a challenging task to calculate the eigenvalues of M explicitly, even for Hamiltonians with

dimension as small as 3×3 , but this has proved to be possible⁴¹ for highly symmetric and sparse band structure Hamiltonians like H^{LK} and several others considered here.

Taking into account the fact that the sequence of eigenenergies of H_0 is semi-bounded from below, for an approximation H^{LK} we obtain

$$\lambda_i < 0, \quad i = 0, 1, \dots, nm. \quad (6)$$

Constraints (6) guarantee the ellipticity (in strong sense⁵⁰) of Hamiltonian H . The operator H possesses a self-adjoint extension in $D(H) \subset \mathcal{H}^{n+2}(\Omega)$, $n > 0$, provided that the boundary $\partial\Omega$ is sufficiently smooth, as we have assumed in the previous section. Then it can be extended to a Hermitian operator by a closure in the norm [p. 113, 49] or via the Lax-Milgram procedure⁴⁸. From the physical point of view the smoothness characteristics of $D(H)$ fulfil the natural assumption of quantum theory that the state of the system must be a continuous function of spatial variables even when some coefficients of H have finite jumps, as it is the case for heterostructures consisting of different materials^{10,47}.

The direct calculation by (5) for H^{LK} ($n = 3$, $m = 6$) leads us to the 18×18 matrix M^{LK} with the following distinct eigenvalues:

$$\begin{aligned} \lambda_1 &= -E(\gamma_1 + 4\gamma_2 + 6\gamma_3), & \lambda_2 &= E(3\gamma_3 - \gamma_1 - 4\gamma_2), \\ \lambda_3 &= E(2\gamma_2 - \gamma_1 + 3\gamma_3), & \lambda_4 &= E(2\gamma_2 - \gamma_1 - 3\gamma_3), \end{aligned} \quad (7)$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ have the multiplicity 2, 4, 6 and 6, respectively, $E = \frac{\hbar^2}{2m_0}$ and $\gamma_1, \gamma_2, \gamma_3$ are the Luttinger material parameters mentioned above. By substituting (7) into (6), we receive the system of linear inequalities with respect to $\gamma_1, \gamma_2, \gamma_3$. They describe the feasibility region Λ_- in the space of ordered triplets $\gamma_1, \gamma_2, \gamma_3$. In this work we shall call a triplet of numbers a, b, c feasible if $(a, b, c) \in \Lambda_-$. More generally, we call a set of material parameters admissible if the Hamiltonian based on this set is an elliptic partial differential operator.

When $(\gamma_1, \gamma_2, \gamma_3) \in \Lambda_-$ the Hamiltonian H^{LK} is an elliptic PDO with the semi-bounded sequence of eigenvalues. One can use similar reasoning to obtain the corresponding inequalities for other common representations of H^{LK} like those through the parameters A, B, C^1 . Evidently, any solution of (6) for (7) would have a unique corresponding solution in the A, B, C notation¹⁹ (the aforementioned ellipticity analysis in full detail is presented in [41]). The region Λ_- comprises an unbounded pyramid in \mathbb{R}^3 (cf. Fig. 1) with the following rays as its edges:

$$\begin{aligned} l_1 &= (8t, t, -2t), & l_2 &= (2t, t, 0), \\ l_3 &= (3t, 0, t), & l_4 &= (4t, -t, 0), \end{aligned}$$

where $t \in [0, \infty)$, and the vertex is situated at the origin $\gamma_1 = \gamma_2 = \gamma_3 = 0$. The boundary of Λ_- and the edges l_1, l_2, l_3, l_4 are illustrated in FIG. 1.

To determine ellipticity of H^{LK} , we gathered in Table I the material parameters $\gamma_1, \gamma_2, \gamma_3$ for GaAs, AlAs, InAs,

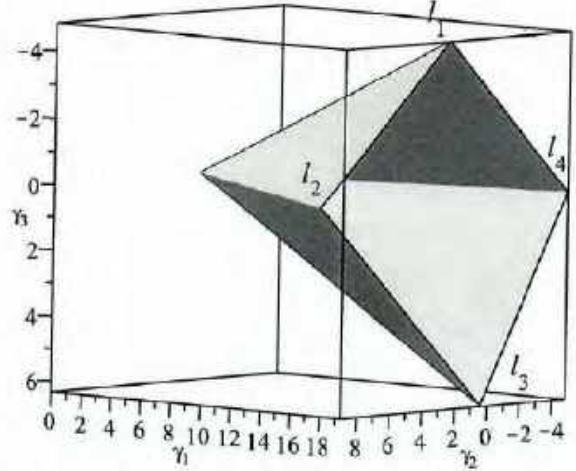


Figure 1. The part of the boundary of the feasibility region Λ_- along with the edges l_1, l_2, l_3, l_4 (color online).

GaP, AlP, InP, GaSb, AlSb, InSb, GaN, AlN, InN, C and evaluated $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ for the gathered triplets. As it turned out, the eigenvalues $\lambda_1, \lambda_2, \lambda_4$ are negative for all analysed parameter sets. In that case the ellipticity is determined by the value λ_3 . We provide the values of λ_3/E along with two other parameter dependent quantities which are important for the current work's ellipticity analysis. The first is the distance d from the parameter triplet $(\gamma_1, \gamma_2, \gamma_3)$ to Λ_- . Second is an absolute ratio ρ between positive and negative values of $\lambda_1, \lambda_2, \lambda_3, \lambda_4$.

From Table I one can observe that among all analysed materials only carbon has admissible sets of parameters (the last two sets from of Table I indicated by 0 in the ρ column). All other gathered parameters yield $\lambda_3 > 0$. That is why the Hamiltonian H^{LK} is not elliptic for the corresponding materials. It may even have no symmetric domain $D(H^{LK})$ as opposed to the original partial differential operator H_0 . Moreover, instead of the inclusion $D(H^{LK}) = D(H) \subset D(H_0) \subset \mathcal{H}^3(\Omega)$ we have only

$$D(H^{LK}) = D(H) \subset \mathcal{H}^1(\Omega). \quad (8)$$

It means that the discontinuous solutions of (1) are theoretically possible. They will occur in the models with jump discontinuous coefficients⁵², which is the case for heterostructure materials. Additionally, the double degeneracy of $\lambda_3 > 0$ from (7) means that for certain Ω there exists a two-dimensional manifold of $D(H^{LK})$ with non-physical in terms of (8) solutions to (1). Thus, the momentum operator from (1), will be ill-defined for such eigenstates of H^{LK} (by the embedding theorems, [p. 119, 49]). All the above arguments allow us to conclude that the H^{LK} does not provide a sufficiently good approximation to H_0 , preserving the type of the PDO, for the most of available data.

Let us return to the feasible parameters from Table I. For carbon the parameter values were analyzed earlier⁵³

Table I. The material parameters for ZB type crystals, d – distance from the point $(\gamma_1, \gamma_2, \gamma_3)$ to the feasibility region Λ_-

#	El	γ_1	γ_2	γ_3	λ_3/E	d	ρ	#	El	γ_1	γ_2	γ_3	λ_3/E	d	ρ
1	GaAs ³⁰	6.980	2.060	2.930	5.930	1.585	0.117	23	InP ^{ob}	5.040	1.560	1.730	3.270	0.874	0.094
2	GaAs ^a	7.100	2.020	2.910	5.670	1.515	0.111	24	InP ³¹	6.280	2.085	2.755	6.156	1.645	0.129
3	GaAs ^b	7.800	2.460	3.300	7.020	1.876	0.121	25	GaSb ³⁰	13.400	4.700	6	14	3.742	0.134
4	GaAs ^c	6.950	2.250	2.860	6.130	1.638	0.119	26	GaSb ^{ag}	11.800	4.030	5.260	12.040	3.218	0.132
5	GaAs ^d	6.850	2.100	2.900	6.050	1.617	0.120	27	GaSb ^b	13.100	4.500	6	13.900	3.715	0.136
6	GaAs ^e	6.800	2.400	1	1	0.267	0.025	28	GaSb ²⁹	13.300	4.400	5.700	12.600	3.367	0.125
7	GaAs ^g	7.200	2.500	1.100	1.100	0.294	0.025	29	GaSb ³¹	11	3	4.368	8.105	2.166	0.105
8	GaAs ³¹	7.150	2.030	2.959	5.788	1.547	0.113	30	AlSb ³⁰	5.180	1.190	1.970	3.110	0.831	0.090
9	AlAs ³⁰	3.760	0.820	1.420	2.140	0.572	0.087	31	AlSb ²⁹	4.150	1.010	1.750	3.120	0.834	0.108
10	AlAs ²⁹	3.760	0.900	1.420	2.300	0.615	0.091	32	AlSb ³¹	4.120	1.045	1.715	3.115	0.832	0.108
11	AlAs ³¹	4.030	1.045	1.697	3.150	0.842	0.110	33	InSb ³⁰	34.800	15.500	16.500	45.700	12.214	0.154
12	InAs ³⁰	20	8.500	9.200	24.600	6.575	0.148	34	InSb ^a	36.130	16.240	17.340	48.370	12.927	0.156
13	InAs ²⁹	20.400	8.300	9.100	23.500	6.281	0.142	35	InSb ^b	36.410	15.940	16.990	40.440	12.412	0.151
14	InAs ²⁹	19.670	8.370	9.290	24.940	6.665	0.151	36	InSb ^c	35.080	15.640	16.910	46.930	12.543	0.156
15	InAs ³¹	19.700	8.400	9.280	24.939	6.665	0.151	37	InSb ³¹	35	15.700	16.821	46.864	12.525	0.156
16	GaP ³⁰	4.050	0.490	1.250	0.680	0.182	0.030	38	GaN ³⁰	2.670	0.750	1.100	2.130	0.569	0.111
17	GaP ^{ag}	4.200	0.980	1.660	2.740	0.732	0.096	39	GaN ²⁹	3.080	0.860	1.260	2.420	0.647	0.110
18	AlP ³⁰	3.350	0.710	1.230	1.760	0.470	0.081	40	GaN ⁶¹	5.050	0.600	1.787	1.511	0.404	0.051
19	AlP ^{ag}	3.470	0.060	1.150	0.100	0.027	0.006	41	AlN ³⁰	1.920	0.470	0.850	1.570	0.420	0.115
20	InP ³⁰	5.080	1.600	2.100	4.420	1.181	0.118	42	InN ^k	3.720	1.260	1.630	3.690	0.986	0.129
21	InP ^a	5.150	0.940	1.620	1.590	0.425	0.052	43	C ^b	2.540	-0.100	0.606	-0.922	0	0
22	InP ^{bg}	6.280	2.080	2.780	6.220	1.662	0.130	44	C ^b	3.610	0.090	1.101	-0.127	0	0

^a Set 1 from 29 ^b Set 2 from 29 ^c Set 3 from 29 ^d Set 4 from 29 ^e Set 5 from 29^f Set 6 from 29 ^g Obtained by extrapolation from 5-level model ^h Measured at $T = 300K$ ⁱ Set from 31 ^k The sets from 30-29

and it was noted that they don't agree well with the Hall effect experimental measurements. In the earlier work¹⁹ we showed that experimentally consistent sets for C are not admissible in terms of ellipticity.

Concerning the rest of the materials from Table I we observe a clear correlation between the average distance to Λ_- per material and the size of the fundamental bandgap. Namely the sets for the large-bandgap materials: AlP, AlAs, GaP, GaN, InP are noticeably close ($d < 1$) to Λ_- . The closest in terms of the distance set number 19 for AlP can be made elliptic by the direct adjustment. Other materials have smaller gap and as a consequence are further away. The average distance to Λ_- for GaAs is around 1.7. For InAs the distance is more than 6. The indium antimonide is an extreme case here, having distance of more than 12. This material has the smallest bandgap and the high curvature of light-hole bands. It is known from the experiments^{2,54,55} that the valence-band-only Luttinger-Kohn model is insufficient for InSb like materials, and prescuted analysis support this fact theoretically. The ellipticity of the higher band $k \cdot p$ models are considered in the next sections.

IV. EIGHT-BAND HAMILTONIANS

This section is devoted to the analysis of Kane model^{33,36}. The basis set of 8×8 Kane Hamiltonian³³

contains two more elements $|S \uparrow\rangle$ and $|S \downarrow\rangle$ in addition to the basis set of H^{LK} . These new elements of basis represent the influence of the innermost conduction band. Recall that the influence of the out-of-basis states is again treated perturbatively up to the second order by using the Löwding perturbation theory. In this section we will follow the exposition of [33], because it presents the most general description of 8×8 Kane Hamiltonian for zinc blende crystals. Naturally, the results presented here remain valid⁵⁶ for other versions^{37,57,58} of the same Hamiltonian. Since our main focus is to check the ellipticity conditions we shall drop the spin-orbit interaction part, labelled as $H_{s.o.} + H'_{s.o.}$ in eq. (13) from [33]. This part of the Hamiltonian is linear in k and therefore won't affect the form of G . (as we have mentioned before, only second order terms in k are essential for ellipticity analysis) Then, following Kane³⁶, we rewrite the resulting operator in the block-diagonal form

$$H^K = \begin{pmatrix} H^K_\uparrow & 0 \\ 0 & H^K_\downarrow \end{pmatrix},$$

where H^K is the Kane 4×4 interaction matrix², given by (9) in the basis $|S \uparrow\rangle, |X \uparrow\rangle, |Y \uparrow\rangle, |Z \uparrow\rangle$ ³³. The matrix H^K_\downarrow that is also defined by (9), acts upon the spin-down part of the basis $|S \downarrow\rangle, |X \downarrow\rangle, |Y \downarrow\rangle, |Z \downarrow\rangle$.

$$H_K^K = \begin{pmatrix} E_c + (E + A')k^2 & iP_0k_x + Bk_yk_z & iP_0k_y + Bk_xk_z & iP_0k_z + Bk_xk_y \\ -iP_0k_x + Bk_yk_z & E_v + M'(k_y^2 + k_z^2) + L'k_x^2 + Ek^2 & N'k_xk_y & N'k_xk_z \\ -iP_0k_y + Bk_xk_z & N'k_xk_y & E_v + M'(k_y^2 + k_z^2) + L'k_x^2 + Ek^2 & N'k_yk_z \\ -iP_0k_z + Bk_xk_y & N'k_xk_z & N'k_yk_z & E_v + M'(k_y^2 + k_z^2) + L'k_x^2 + Ek^2 \end{pmatrix}. \quad (9)$$

Parameters A', B, P_0, M', N', L' are known as Kane parameters³⁶, their definitions are provided in Table 4.2 of 37. The quantities E_c and E_v are the conduction- and valence-band energies correspondingly, E is equal to $\frac{\hbar^2}{2m_0}$, as before. The parameter A' represents the influence of the higher bands on the conduction band included into the basis. The parameter P_0 accounts for a mixing of conduction and valence band states away from $\mathbf{k} = \mathbf{0}$. B is a so-called inversion asymmetry parameter. It is equal to zero in the materials with centrosymmetric crystal structure like diamond³³. By setting $B = 0$ in (9) we obtain a simplified version of (9) that is known as Bir-Pikus 4×4 Hamiltonian. The general case of H_K when $B \neq 0$ was studied by T. Bahder (Eq. (15) in [33]). In practice the mentioned parameters are fitted to experimental data; It is frequently assumed in the literature that the simplified version of H_K provides a sufficiently good description of the physical phenomena in ZB crystals with face-centered lattice too. As we will later demonstrate, the Hamiltonian of such simplified model is non-elliptic for all studied material parameter sets and therefore is prone to the appearance of spurious solutions. The parameter B can not be set to zero for the materials where $E + A' < 0$.

Similarly to the 6×6 case, it is common to rewrite Hamiltonian H_K in the basis where its spin-orbit interaction part becomes diagonal. Usually one additionally pre-multiplies the original basis functions to make interband matrix elements and possibly other physically relevant quantities real-valued.

Direct calculation of eigenvalues for the quadratic form associated with H_K^K , described in details for the Luttinger-Kohn case from the previous section, gives us five distinct eigenvalues

$$\begin{aligned} \lambda'_1 &= E + L' + N', \\ \lambda'_2 &= E + L' - \frac{1}{2}N', \\ \lambda'_3 &= E + M' - \frac{1}{2}N', \\ \lambda'_4 &= E + \frac{2A' + 2M' + N'}{4} - \sqrt{\frac{(2A' - 2M' - N')^2}{16} + \frac{B^2}{2}}, \\ \lambda'_5 &= E + \frac{2A' + 2M' + N'}{4} + \sqrt{\frac{(2A' - 2M' - N')^2}{16} + \frac{B^2}{2}}. \end{aligned} \quad (10)$$

The presence of the second order conduction-valence band mixing, characterized by the parameter B of Kane Hamiltonian (9), is reflected in (10) by the pair of eigenvalues λ'_1, λ'_5 , which are both determined by the whole set

of the principal Hamiltonian parameters N, M, L, A', B . Note that, if one removes the mixing by setting $B = 0$, this property disappears and the eigenvalues λ'_4, λ'_5 are turned into

$$\lambda'_{04} = E + M' + \frac{1}{2}N', \quad \lambda'_{05} = E + A'.$$

A. Ellipticity analysis in the absence of inversion asymmetry

We analyze the set $\lambda'_1, \lambda'_2, \lambda'_3, \lambda'_{04}, \lambda'_{05}$ associated with $B = 0$ in H_K^K first. The fifth eigenvalue λ'_{05} in (10) is related to the conduction band of (9) because its corresponding three-dimensional eigenspace (λ'_{05} is triple degenerate) has only 3 first coordinates not equal to zero. Hence this eigenspace is orthogonal to the space associated with the valence bands. Those are characterized by the eigenvalues $\lambda'_1, \lambda'_2, \lambda'_3, \lambda'_{04}$ with degeneracy 1, 2, 3, 3, respectively. The following system of inequalities ensures ellipticity of 8×8 ZB Hamiltonian³³ with zero B

$$\begin{cases} E + L' + N' < 0 \\ E + L' - \frac{1}{2}N' < 0 \\ E + M' - \frac{1}{2}N' < 0 \\ E + M' + \frac{1}{2}N' < 0 \\ E + A' > 0. \end{cases} \quad (11)$$

As we mentioned, the eigenvalues $\lambda'_1, \lambda'_2, \lambda'_3, \lambda'_{04}$ are related to the valence band, hence the sign of the first four inequalities from (11) is the same as in (6). The opposite sign of the fifth inequality reflects its correspondence to the conduction band. Due to the electron-hole duality, the conduction band eigen-energies need to be semi-bounded from below. The presence summand E in system (11) is connected with the differences in the definition of Dresselhaus parameters⁵⁹ and L', M' ³³.

To compare the result for the 8×8 ZB Hamiltonian with the previously obtained results for the 6×6 Hamiltonian we define the dimensionless parameters $\gamma'_1, \gamma'_2, \gamma'_3$

similar to the Luttinger triplet^{37,54,57}

$$\begin{aligned}\gamma'_1 &= -\frac{1}{3}(L' + 2M') \frac{2m_0}{\hbar^2} - 1 \\ \gamma'_2 &= -\frac{1}{6}(L' - M') \frac{2m_0}{\hbar^2} \\ \gamma'_3 &= -\frac{1}{6}N' \frac{2m_0}{\hbar^2}.\end{aligned}$$

Hereby, the system (11) is transformed to

$$\begin{cases} -\gamma'_1 - 4\gamma'_2 - 6\gamma'_3 < 0 \\ -\gamma'_1 - 4\gamma'_2 + 3\gamma'_3 < 0 \\ -\gamma'_1 + 2\gamma'_2 + 3\gamma'_3 < 0 \\ -\gamma'_1 + 2\gamma'_2 - 3\gamma'_3 < 0 \\ 1 + A > 0,\end{cases} \quad (12)$$

with $A = A'/E$.

The modified and the original Luttinger parameters $\gamma_1, \gamma_2, \gamma_3$ are connected by the formulas⁵⁴

$$\gamma'_1 = \gamma_1 - \frac{E_p}{3E_g}, \quad \gamma'_2 = \gamma_2 - \frac{E_p}{6E_g}, \quad \gamma'_3 = \gamma_3 - \frac{E_p}{6E_g}, \quad (13)$$

where $E_p = P_0^2/E$, $E_g = E_c - E_v$ is a fundamental bandgap energy, P_0 is the Kane parameter from (9).

As it was expected, four out of five obtained inequalities (12), which represent the ellipticity constraints for the valence band part of H^K , have the structure equivalent to that for the LK Hamiltonian (7). Hence, the feasibility region of the valence-band part of H^K in the space of parameters $(\gamma'_1, \gamma'_2, \gamma'_3)$ coincides with the feasibility region Λ_- of H^{LK} , depicted in FIG. 1. It means that if $(\gamma'_1, \gamma'_2, \gamma'_3) \in \Lambda_-$, the valence-band part of the Hamiltonian H^K in the position representation is an elliptic partial differential operator. Then, the transformation given by (13) can be geometrically interpreted as a shift in the space of parameters proportional to vector $\mathbf{v}' = (-2, -1, -1)$. This shift reduces the value of λ'_3 and, as we shall soon see, brings the majority of the non-elliptic parameter triplets $(\gamma_1, \gamma_2, \gamma_3)$ closer to the feasibility region.

The dimensionless parameter A from the fifth inequality, that complements a set of ellipticity constraints (12), is responsible for a coupling between the conduction band and other states. It is commonly assumed that the in-basis valence bands are the major contributors to A . The value of A is determined by matching its value to the effective mass of conduction band m_c , determined experimentally using the formula

$$A = \frac{m_0}{m_c} - 1 - E_p \frac{E_g + \frac{2}{3}\Delta}{E_g(E_g + \Delta)}. \quad (14)$$

The magnitude of this parameter is clearly affected by the size of band-gap E_g and spin-splitting Δ . The experimental nature of m_c does not factor out other possible contributions to A . For that reason we extended the collection of parameter sets from Table I by those stemming from the same sets of Luttinger parameters and the

different values of bandgap energy E_g (measured within different experimental setups). We also added a parameter set obtained by fitting the bandstructure of 8×8 Hamiltonian to the bandstructure calculated by ab-initio methods⁵⁸. All the data pertaining to the ellipticity analysis of 8×8 Hamiltonian is collected in Table II. In each case the modified Luttinger parameters $\gamma'_1, \gamma'_2, \gamma'_3$ were calculated by using (13) and the values of P_0^2, E_g provided in the dataset source. For those sources from the table that have P_0^2 unavailable we use the values collected by I. Vurgaftman, J. R. Meyer and L. R Ram-Mohan³⁰.

The ellipticity conditions of H^K are still violated for all materials presented in Table I. The situation is, however, more complex than for the 6×6 Hamiltonian. To illustrate that, we supplied in Table II the values of $\lambda'_1 - \lambda'_{05}$, the distance d to the feasibility region from $(\gamma'_1, \gamma'_2, \gamma'_3)$ and the measure of non-ellipticity ρ which is defined in the same way as for the 6×6 Hamiltonian case.

Overall, we can confirm the reduction of average distance to the feasibility region for all materials, especially for InAs and InSb. Furthermore, for several materials there exist parameter sets that are close to satisfy the full set of ellipticity constraints described by (12). Those are narrow gap semiconductor InSb (sets #37; #38 from Table II) and, perhaps more surprisingly, the materials with larger band-gap InP, AlAs and AlSb (sets #25, #10; #11, and #33). For these materials the corresponding parameter sets can be made elliptic by direct adjustment of $\gamma'_1, \gamma'_2, \gamma'_3, A$.

Certain parameter sets for AlP, AlSb and InAs satisfy ellipticity conditions for the valence band part of the Hamiltonian and do not satisfy the conduction-band constraint (inequality 5 from (12)). Among those, the sets #20, #33 for AlP, AlSb reported in [30] differ sharply in the size of γ_2 from two other sets for these materials collected in Table II. For AlP this can explained by the fact that in the absence of direct experimental data most of the material parameters were extrapolated from measurements for ternary alloys and ab-initio calculations which carries a lot of uncertainty. The authors of [30] performed readjustment of the Luttinger parameters to better match the experimental photoluminescence results on AlP/GaP heterostructures⁶¹. The set #33 for AlSb is based on the available theoretical calculations from various sources and the simultaneous fitting of $\gamma_1, \gamma_2, \gamma_3$ to the experimentally-determined hole effective masses along [001], [110] and [111] directions³⁰.

For InAs we can judge from the size of $\lambda'_1 - \lambda'_{04}$ that the triplets $(\gamma'_1, \gamma'_2, \gamma'_3)$ of its parameter sets are right near the side of Λ_- described by $\lambda'_2 = 0$. Two are inside (sets #12; #13) and two others are slightly off (sets #14; #15). The values of λ'_{05} for all four parameter sets are grouped near the value $\lambda'_{05} = -4.8$ and thus the conduction part of the Hamiltonian is, again, far from being elliptic.

All material parameter sets for GaAs violate two out of four ellipticity conditions for the valence-band part, although one parameter set #3 from Table II stays close to Λ_- ($d = 0.34$). However, it violates the ellipticity con-

Table II. The material data for 8×8 ZB Hamiltonian with $B = 0$, d – distance from the point $(\gamma_1, \gamma_2, \gamma_3)$ to the feasibility region Λ_- . The positive values of $\lambda'_1/E - \lambda'_{05}/E$ are printed in bold.

#	El	E_p	E_g	Δ_{SO}	A'	γ'_1	γ'_2	γ'_3	λ'_1/E	λ'_2/E	λ'_3/E	λ'_{04}/E	λ'_{05}/E	d	ρ	Δ_{03}^{\min}	Δ_{03}^{\max}
1	GaAs ³⁰	28.80	1.52	0.341	-3.88	0.66	-1.10	-0.23	5.12	3.05	-3.55	-2.17	-2.88	0.70	1.43	5.73	6.67
2	GaAs ^a	28.80	1.52	0.341	-3.88	0.78	-1.14	-0.25	5.28	3.03	-3.81	-2.31	-2.88	0.73	1.36	5.69	7.15
3	GaAs ^b	28.80	1.52	0.341	-3.88	1.48	-0.70	0.14	0.48	1.74	-2.46	-3.30	-2.88	0.34	0.39	3.27	4.62
4	GaAs ^c	28.80	1.52	0.341	-3.88	0.63	-0.91	-0.30	4.81	2.11	-3.35	-1.55	-2.88	0.66	1.41	3.96	6.29
5	GaAs ^d	28.80	1.52	0.341	-3.88	0.48	-0.76	-2.16	15.52	-3.92	-8.48	4.48	-2.88	2.13	1.61	8.41	15.92
6	GaAs ^e	28.80	1.52	0.341	-3.88	0.88	-0.66	-2.06	14.12	-4.42	-8.38	3.98	-2.88	1.94	1.41	7.47	15.74
7	GaAs ³¹	28.80	1.52	0.346	-3.86	0.83	-1.13	-0.20	4.89	3.09	-3.69	-2.49	-2.86	0.67	1.29	5.79	6.93
8	GaAs ⁵⁸	25.47	1.52	0.341	-3.34	1.28	-0.73	0.03	1.46	1.73	-2.65	-2.83	-2.34	0.34	0.58	3.25	4.98
9	AlAs ³⁰	21.10	3.10	0.280	-0.95	1.49	-0.31	0.29	-1.94	0.62	-1.26	-2.98	0.05	0.12	0.10	1.21	3.46
10	AlAs ²⁹	21.10	3.10	0.280	-0.95	1.49	-0.23	0.29	-2.26	0.30	-1.10	-2.82	0.05	0.06	0.05	0.59	2.15
11	AlAs ³¹	21.10	3.14	0.275	-0.87	1.79	-0.07	0.58	-4.95	0.24	-0.21	-3.67	0.13	0.05	0.03	0.47	0.41
12	InAs ³⁰	21.50	0.42	0.390	-5.79	2.81	-0.09	0.61	-6.08	-0.62	-1.18	4.82	-4.79	-0.12	0	4.79	1.98
13	InAs ²⁹	21.50	0.42	0.390	-5.79	3.21	-0.29	0.51	-5.08	-0.52	-2.28	-5.32	-4.79	-0.10	0	4.79	3.82
14	InAs ³⁴	21.50	0.42	0.390	-5.79	2.48	-0.22	0.70	-5.77	0.50	-0.84	-5.02	-4.79	0.10	0.04	4.79	1.41
15	InAs ³¹	21.50	0.42	0.380	-5.81	2.55	-0.17	0.71	-6.11	0.26	-0.78	-5.02	-4.81	0.05	0.02	4.81	1.31
16	GaP ³⁰	31.40	2.89	0.080	-4.09	0.42	-1.32	-0.56	8.25	3.18	-4.76	-1.38	-3.09	1.13	1.86	6.30	9.43
17	GaP ²⁶	22.20	2.88	0.080	-0.95	1.63	-0.30	0.38	-2.66	0.71	-1.11	-3.37	0.05	0.14	0.10	1.42	2.21
18	GaP ^b	22.20	2.88	0.080	-0.95	1.48	-0.79	-0.03	1.91	1.59	-3.17	-2.97	0.05	0.31	0.57	3.16	6.29
19	GaP ³¹	31.40	2.90	0.080	-4.06	0.58	-0.82	-0.15	3.63	2.24	-2.69	-1.77	-3.06	0.50	1.32	4.45	5.34
20	AlP ³⁰	17.70	3.63	0.070	-1.30	1.72	-0.10	0.42	-3.82	-0.06	-0.68	-3.18	-0.30	-0.01	0	0.30	1.35
21	AlP ²⁶	17.70	3.63	0.070	-1.30	1.84	-0.75	0.34	-0.86	2.18	-2.34	-4.36	-0.30	0.43	0.29	4.33	4.65
22	AlP ³¹	17.70	3.63	0.070	-1.30	1.84	-0.75	0.33	-0.85	2.14	-2.34	-4.34	-0.30	0.42	0.28	4.26	4.66
23	InP ³⁰	20.70	1.42	0.108	-2.62	0.23	-0.82	-0.32	5.00	2.09	-2.85	-0.91	-1.62	0.69	1.89	4.08	5.57
24	InP ^a	16.70	1.45	0.108	0.36	1.32	-0.97	-0.29	4.34	1.69	-4.15	-2.39	1.36	0.60	0.92	3.31	8.11
25	InP ^b	20.40	1.56	0.108	-1.22	1.92	-0.10	0.60	-5.11	0.28	-0.32	-3.92	-0.22	0.06	0.03	0.55	0.63
26	InP ^c	17.50	1.47	0.108	-0.09	1.06	-0.43	-0.26	2.23	-0.12	-2.70	-1.14	0.91	0.31	0.56	1.09	5.28
27	InP ³¹	20.70	1.34	0.108	-3.44	1.15	-0.48	0.19	-0.35	1.35	-1.54	-2.68	-2.44	0.26	0.29	2.63	3.01
28	GaSb ³⁰	27	0.81	0.760	-3.25	2.32	-0.84	0.46	-1.70	2.43	-2.63	-5.37	-2.25	0.48	0.25	4.07	4.40
29	GaSb ²⁹	22.40	0.81	0.725	1.39	2.60	-0.57	0.66	-4.31	1.65	-1.75	-5.73	2.39	0.32	0.14	2.79	2.95
30	GaSb ^b	26.10	0.81	0.725	-2.45	2.39	-0.86	0.64	-2.81	2.97	-2.17	-6.03	-1.45	0.58	0.27	5.01	3.66
31	GaSb ²⁹	25	0.81	0.725	-1.31	3.04	-0.73	0.57	-3.52	1.59	-2.79	-6.21	-0.31	0.31	0.13	2.69	4.71
32	GaSb ³¹	27	0.75	0.756	-5.34	-1	-3	-1.63	22.79	8.11	-9.89	-0.11	-4.34	3.13	3.09	13.50	16.48
33	AlSb ³⁰	18.70	2.39	0.676	-1.12	2.57	-0.12	0.66	-6.09	-0.11	-0.81	-4.79	-0.12	-0.02	0	0.12	1.50
34	AlSb ²⁹	18.70	2.39	0.680	-1.12	1.54	-0.30	0.44	-3.02	0.98	-0.80	-3.46	-0.12	0.19	0.13	1.81	1.48
35	AlSb ³¹	18.70	2.30	0.673	-1.37	1.41	-0.31	0.36	-2.33	0.91	-0.95	-3.11	-0.37	0.18	0.14	1.68	1.76
36	InSb ³⁰	23.30	0.24	0.810	-0.46	1.75	-1.02	-0.02	2.50	2.27	-3.87	-3.73	0.54	0.45	0.63	3.37	5.75
37	InSb ^b	23.20	0.24	0.803	-0.13	3.25	-0.20	0.90	-7.85	0.25	-0.95	-6.35	0.87	0.05	0.02	0.37	1.41
38	InSb ^c	23.42	0.24	0.803	-0.37	3.44	-0.54	0.51	-4.31	0.25	-3.01	-6.05	0.63	0.05	0.02	0.37	4.47
39	InSb ^d	23.10	0.24	0.803	0.12	2.31	-0.74	0.53	-2.50	2.24	-2.22	-5.38	1.12	0.44	0.22	3.32	3.29
40	InSb ³¹	23.30	0.18	0.810	-21.07	-8.15	-5.87	-4.75	60.16	17.39	-17.86	10.66	-20.07	8.26	4.94	25.29	25.98
41	GaN ³⁰	25	3.30	0.017	-1.90	0.14	-0.51	-0.16	2.89	1.42	-1.66	-0.68	-0.90	0.40	1.84	2.83	3.31
42	GaN ⁶⁰	25	3.30	0.017	-1.90	0.17	-0.50	-0.15	2.76	1.38	-1.64	-0.72	-0.90	0.38	1.75	2.75	3.27
43	GaN ²⁹	25	3.30	0.017	-1.90	0.55	-0.40	-0.00	1.08	1.05	-1.37	-1.35	-0.90	0.21	0.78	2.09	2.73
44	GaN ⁵¹	25	3.44	0.017	-1.59	2.63	-0.61	0.58	-3.64	1.54	-2.12	-5.58	-0.59	0.30	0.14	3.08	4.24
45	GaN ³²	16.86	3.07	0.017	-1.30	0.68	-0.28	0.06	0.07	0.63	-1.05	-1.42	-0.30	0.12	0.28	1.25	2.09
46	AlN ³⁰	27.10	6	0.019	-1.51	0.41	-0.28	0.10	0.13	1.01	-0.69	-1.27	-0.51	0.20	0.58	2.01	1.38
47	AlN ⁶⁰	27.10	5.40	0.019	-2.01	0.25	-0.37	0.01	1.14	1.26	-0.94	-1.02	-1.01	0.25	1.22	2.52	1.88
48	AlN ³²	23.84	5.63	0.019	-2.07	0.04	-0.36	-0.13	2.15	1.01	-1.13	-0.37	-1.07	0.30	2.10	2.01	2.26
49	InN ³⁰	25	1.94	0.006	-5.54	-0.58	-0.89	-0.52	7.23	2.57	-2.75	0.35	-4.54	0.99	3.69	5.14	5.50
50	InN ⁶⁰	17.20	0.78	0.005	-8.72	-3.63	-2.42	-2.05	25.56	7.16	-7.34	4.94	-7.72	3.51	5.13	14.28	14.64
51	InN ³²	11.37	0.53	0.005	-3.87	-0.34	-0.77	-0.46	6.13	2.04	-2.56	0.17	-2.87	0.84	3.25	4.06	5.11

^a Set 1 from 29 ^b Set 2 from 29 ^c Set 3 from 29^d Set 4 from 29^e Set 5 from 29^f Set 6 from 29^g Obtained by extrapolation from 5-level model^h Measured at $T = 300K$

dition for the conduction-band part by the same margin of approximately -2.9 as do other sets of GaAs parameters, let alone set #8 where the margin is slightly lower: $\lambda'_{05} \approx -2.34$. The indicated reduction of margin should be attributed to the optimization procedure⁵⁸ used to acquire set #8. As far as the ellipticity is concerned, this optimization procedure is no more effective than other acquisition methods.

The sets for GaN and AlN are failing first two valence-part constraints from (12) just like the most of the other material parameters. One exception is the set number 44 for GaN⁵¹, where the spherical symmetry of the heavy-hole and the light-hole bands is assumed. This assumption leads to the larger values of γ_1, γ_3 and smaller γ_2 ; and, as a consequence, more than five times smaller ratio ρ between positive and negative eigenvalues.

Even a more severe situation is observed for InN. The conduction-band eigenvalue λ'_{05} is noticeably below zero for all three available datasets ($\lambda'_{05} \approx -4.54, -7.72, -2.87$ for sets #49; #50; #51 accordingly). In addition, three out of four valence-band conditions are violated. It is important to highlight that the recently obtained set of parameters #51 features roughly two times larger values of γ_1, γ_2 and γ_3 and noticeably smaller E_p, E_g than two other sets #49, #50 reported earlier^{30,60}. As demonstrated in [32] set #51 recovers bandstructure better than two others sets for InN, discussed above. In terms of ellipticity, this set results in a lower, than others, distance to Λ_- ($d \approx 0.84$) and lower $\lambda'_{05} \approx 0.17$.

For GaP and GaSb the data seem inconclusive as the size and the sign of eigenvalues (10) are dependant on the choice the material parameter dataset. Sets #17, #29 from Landolt-Börnstein²⁹, based on the earlier data of P. Lawaetz²⁸, are most favourable in terms of ellipticity: $d \approx 0.14$, $\lambda'_{05} \approx 0.05$ for GaP; $d \approx 0.32$, $\lambda'_{05} \approx 2.39$ for GaSb. As a summary of the above analysis, we visualize in FIG. 2 the values of λ'_{05}, d for the selected parameter sets with the material-wise minimal distance to Λ_- . In this figure the ellipticity of Hamiltonian is depicted by the region (shaded in gray) where $\lambda'_{05} > 0$ and $d < 0$ simultaneously.

It is worth noticing that roughly 76% of parameter sets for analyzed materials fail the conduction-band constraint $\lambda'_{05} > 0$. This group includes all datasets for GaAs, InAs, AlP, AlSb, GaN, AlN, InN, quite important for applications. The positive (negative) sign of λ'_{05} is responsible for positive(negative) gain in the energy as we go from one conduction-band eigenvalue of Hamiltonian to the next in the position representation. In the momentum representation, the eigenvalue's sign and its magnitude is responsible for upward (downward) curvature of conduction band. In addition to the highlighted in section III issues caused by the non-ellipticity of valence-band part of the Hamiltonian H^K , the violation of condition $\lambda'_{05} > 0$ entails the existence of conduction-band related eigenstates of H^K with energies in the band-gap or the regions related to valence bands^{14,16,17,57}. This obviously poses a serious problem in applications.

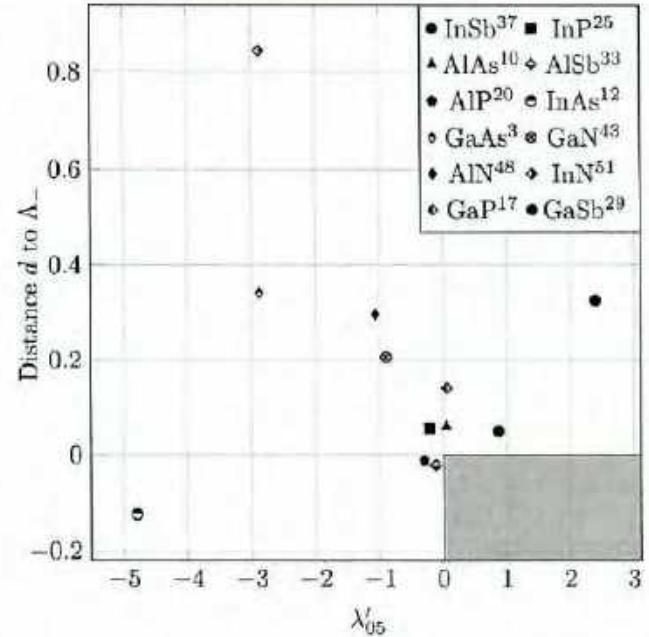


Figure 2. The values of quantities λ'_{05}, d for the list of selected material parameter sets from Table II (color online). The shaded region indicates pairs (λ'_{05}, d) based on the ellipticity constraints of 8×8 model³³

A rescaling procedure was introduced by B. Foreman in [20] (see also the work of S. Birner [57]) and has been adopted^{62,63} ever since as a way to make λ'_{05} positive and avoid the above-described type of spurious solutions. The idea of the procedure is to adjust the momentum matrix element E_p so that λ'_{05} is no longer negative. Let us assume that we target some value of $\lambda'_{05} = a$. Then, by using the definition of λ'_{05} and (14), we obtain⁵⁷

$$E_p = \left(\frac{m_0}{m_e} - a \right) \frac{E_g(E_g + \Delta_{SO})}{E_g + \frac{2}{3}\Delta_{SO}}. \quad (15)$$

Two values $a = 0$ and $a = 1$ are considered in the literature as a target for rescaling. The new value of E_p will necessarily affect the values of modified Luttinger parameters $\gamma'_1, \gamma'_2, \gamma'_3$ which are defined by (13). To figure out how this procedure would impact the ellipticity of the entire 8×8 ZB Hamiltonian H^K one needs to rewrite eigenvalues $\lambda'_1 - \lambda'_{04}$ as functions of E_p, E_g :

$$\lambda'_1 = \lambda_1 + 2E \frac{E_p}{E_g}, \quad \lambda'_{2/04} = \lambda_{2/4} + \frac{E}{2} \frac{E_p}{E_g}, \quad \lambda'_3 = \lambda_3 - \frac{E}{2} \frac{E_p}{E_g}.$$

By combining the above representations with (15) we obtain a new version of ellipticity constraints (12) for the valence-band part of H^K

$$\begin{aligned} \lambda_1 + 2E_m - \frac{2a}{E_r} &< 0, & \lambda_2 + \frac{1}{2}E_m - \frac{a}{2E_r} &< 0, \\ \lambda_3 - \frac{1}{2}E_m + \frac{a}{2E_r} &< 0, & \lambda_4 + \frac{1}{2}E_m - \frac{a}{2E_r} &< 0, \end{aligned} \quad (16)$$

where $E_m = \frac{m_0}{m_e E_r}$, $E_r = \frac{E_g + \frac{2}{3}\Delta_{SO}}{E(E_g + \Delta_{SO})}$ are two material-dependent constants. Now we substitute back $a = \lambda'_{05} + \Delta_{05}$ and solve system of inequalities (16) with respect to Δ_{05} . As a result, it will give us the range for the values of the rescaling parameter Δ_{05} that make the valence-band part of Kane Hamiltonian elliptic:

$$E_r \max \left\{ \frac{1}{2} \lambda'_1, 2\lambda'_2, 2\lambda'_{04} \right\} < \Delta_{05} < -2\lambda'_3 E_r. \quad (17)$$

The calculated values for the ranges from (17) are provided in the last two columns of Table II. If $-\lambda'_{05} < -2\lambda'_3 E_r$, then the Hamiltonian can be made elliptic by setting Δ_{05} to the arbitrary value within range (17) so that $\lambda'_{05} + \Delta_{05}$ is positive. In practice one would also like to make sure that the numerical inaccuracies introduced by the eigenvalue calculation procedure for H^K will not overturn any of the signs of $\lambda'_1 - \lambda'_{05}$. To minimize that possibility and to keep Δ_{05} reasonably small we suggest the following formula for the selection of Δ_{05} :

$$\Delta_{05} = \begin{cases} 2\Delta_m + 0.1, & \Delta_m + \lambda'_3 E_r < -0.1, \\ \Delta_m - \lambda'_3 E_r, & \text{otherwise.} \end{cases} \quad (18)$$

with $\Delta_m = \max \left\{ \frac{1}{4} \lambda'_1 E_r, \lambda'_2 E_r, \lambda'_{04} E_r, -\frac{\lambda'_{05}}{2E} \right\}$.

We carried out the rescaling procedure for the material parameters from Table II and selected the sets with minimal Δ_{05} for every given material. The resulting values of readjusted E_p , A' along with new values of modified Luttinger parameters are presented in Table III. It is also worth noting that the resulting value of $1 + A$ in our case is never equal to zero or one, as it was usually assumed by authors before^{20,57}. For many materials the value of adjusted parameter A is greater than 0.

To see the impact of rescaling on the band dispersion, in Table III we also supplied a maximum absolute difference (adjustment error) between the corresponding bands of bandstructure calculated over the 20% of three high symmetry paths ΓL , ΓK and ΓX pertaining to the first Brillouin zone (FBZ). Such a size of the domain for comparison is common³⁰ and motivated by the existing evidence⁵⁸ that the accurate fit of the $k \cdot p$ bandstructure to the state-of-the-art ab-initio calculations is possible over this part of FBZ. To ascertain the band that contributes most to the error we supplied in FIG. 3 a), and FIG. 4 the graphical comparison of band-structure diagrams for every set from Table III and the original sets of material parameters from Table II, on which they are based. For clarity, only bands with even numbers in the representation of 8×8 Hamiltonian³³ are plotted in these figures.

As one immediately notices from the last column of Table III, the chosen sets for AlN, AlP, AlSb, and AlAs materials are least susceptible to the performed rescaling procedure. The differences between the bandstructure for modified parameters sets (four topmost rows from

Table III) Selected material parameters, rescaled via (18) together with the difference between corresponding bands of 8×8 Hamiltonian³³ for original and rescaled parameters

El ^a	Δ_{05}^b	E_p	A'	γ'_1	γ'_2	γ'_3	λ_v^c	err^d
AlN ⁴⁸	2.11	11.92	0.049	0.744	-0.004	0.226	-0.05	2.78
AlP ²⁰	0.40	16.24	-0.900	1.859	-0.036	0.484	-0.26	4.29
AlSb ³³	0.22	18.14	-0.900	2.646	-0.077	0.703	-0.23	5.84
AlAs ¹⁰	0.69	18.90	-0.262	1.728	-0.116	0.404	-0.05	9.35
GaN ³³	2.19	17.75	0.296	1.287	-0.037	0.363	-0.05	11.58
InP ²⁵	0.59	19.46	-0.632	2.120	0	0.700	-0.02	11.69
Gap ¹⁷	1.52	17.80	0.569	2.140	-0.050	0.630	-0.05	23.34
InSb ³⁷	0.47	23.05	0.336	3.462	-0.094	1.006	-0.07	32.66
InN ⁵¹	4.16	9.16	0.285	1.054	-0.071	0.240	-0.05	35.03
GaAs ³	3.37	23.35	-0.509	2.676	-0.102	0.738	-0.05	50.62
GaSb ²⁹	2.87	19.63	4.263	3.740	0	1.230	-0.05	171.40

^a Refer to the original dataset number from Table II

^b The quantity Δ_{05} describes the size of adjustment to A'

^c The values of λ_v are calculated via $\lambda_v = \max\{\lambda'_1, \lambda'_2, \lambda'_3, \lambda'_{04}\}$ ^d Maximum difference in meV between the bandstructure for the original parameters from Table II and the rescaled parameters calculated by using (18) over the 20 % of the paths ΓL , ΓK and ΓX

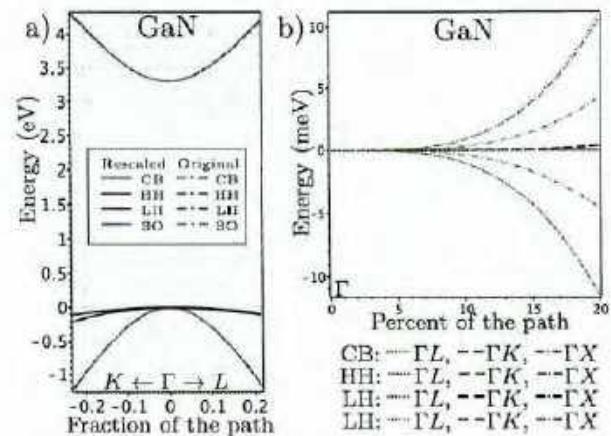


Figure 3. Comparison of original and rescaled parameter sets for GaN (color online): Conduction band (CB), heavy-hole (HH), light-hole (LH) and the split-off band (SO). a) Bandstructure along the fraction of symmetry path $K - \Gamma - L$: original set 43 (solid), rescaled set from Table III (dashed); b) bandstructure adjustment error along the paths ΓL , ΓK and ΓX

Table III) and the original sets for materials from this group are less than 10 meV. We will call these differences bandstructure adjustment errors or simply errors, when it is unambiguous. For GaN and InP the errors of approximately 11 meV are also visually indistinguishable in FIG. 3 a) and FIG. 4. Thus we supplied in FIG. 3 b) their plot for GaN, that has an appropriate vertical scaling. This plot, typical for all analyzed materials except of InN, shows the behaviour of bandstructure adjustment error along three main paths ΓL , ΓK and ΓX . For the GaN, the errors along directions ΓL , ΓK are about 11

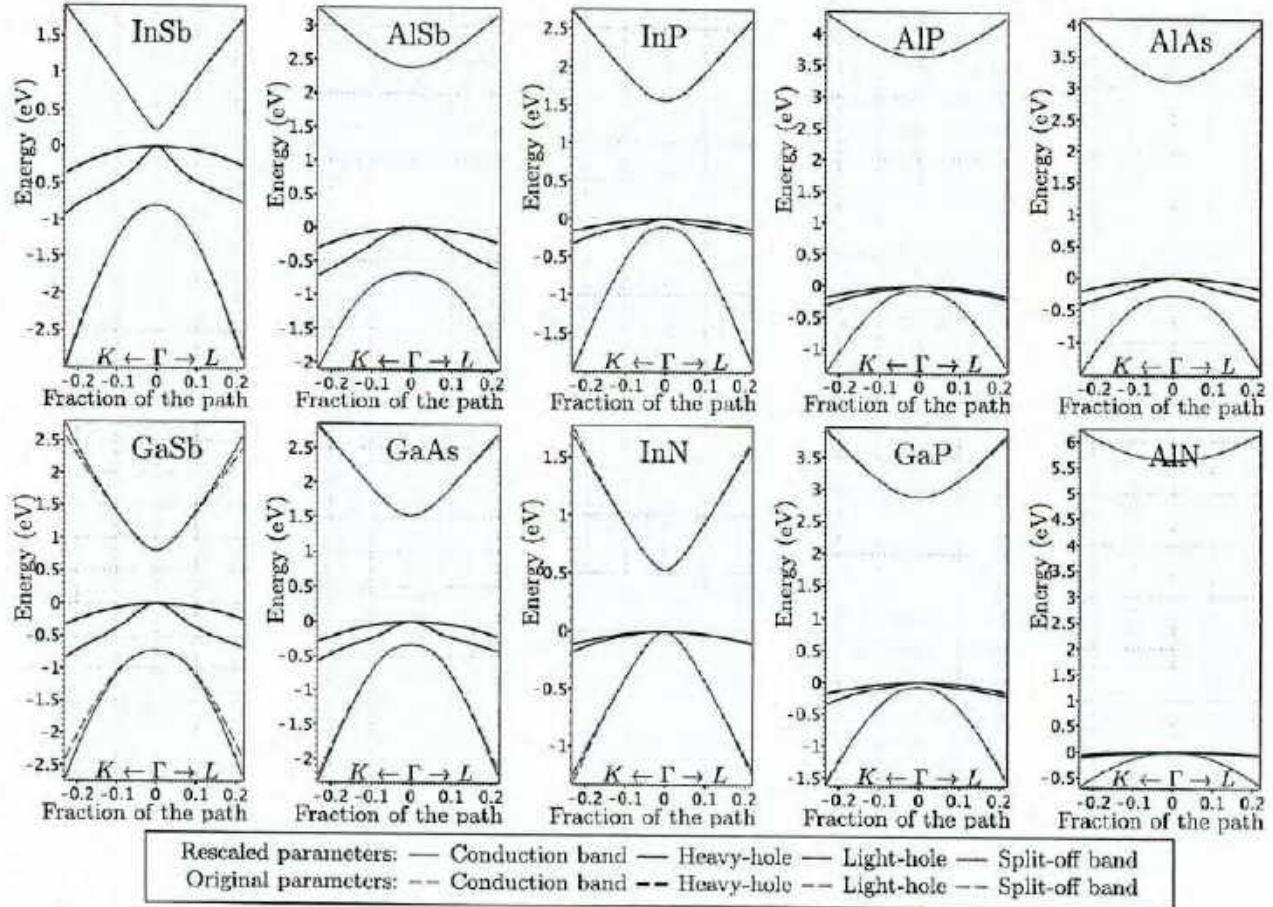


Figure 4. Comparison of bandstructure for the selected original parameter sets from Table II and the rescaled sets from Table III (color online). The band dispersion is plotted from even eigenenergies of 8×8 Hamiltonian³³ along the fraction of symmetry path $K -\Gamma -L$ in the vicinity of Γ : original sets (solid line), rescaled sets (dashed line).

meV, while the band errors in the direction ΓX is around two times lower. What makes this material unique is the fact that such tiny errors are the result of a significant change in the values of material parameters during the rescaling. Namely, the difference in E_p is around 29% and above 100% for γ'_1, γ'_2 .

The situation is more close to the anticipated for another group of materials: GaP, InN, GaAs. The relative differences in $E_p \approx 19\%$ for all three materials, but the error is higher for GaAs than for GaP and InP: 50.62 meV vs 23.34 meV and 35.02 meV, respectively. This can be explained by a closer proximity of p -like conduction bands in GaAs, that are treated perturbatively in the current model. For indium antimonide the band adjustment error of 32.66 meV (barely visible as a slightly higher curvature of conduction and SO bands in the first plot of FIG. 4) lays within the same range as for GaP, InN, GaAs. What is unusual is that these differences in band dispersion were produced by the smallest (among all analyzed materials) adjustment of $E_p - 0.6\%$, which resulted in only approximately 10% increase of γ'_1, γ'_3 . Such error sensitivity might be attributed to the very

small bandgap (see FIG. 4). The conduction band (CB) adjustment error for InSb is equal to 10.87 meV. That is about 3 times smaller than the valence band adjustment error and therefore invisible in the plot. The same is true for heavy hole and light hole bands.

Similar tendencies are valid for other materials from Table III. The performed adjustment of A leads to a slightly noticeable change in the conduction band dispersion. Heavy hole (HH) and light hole (LH) bands remain visually unaffected even though the differences are non-zero. The rescaling also causes an increase in the curvature of the split-off (SO) band, making it the main source of total valence-band adjustment error.

The maximum adjustment error of 171.4 meV was observed in gallium antimonide. We postpone a detailed discussion of GaSb till the next subsection and focus now on the following question: How the dispersion of CB and SO band can be corrected without braking ellipticity of the H^K Hamiltonian?

B. Ellipticity analysis for 8×8 ZB Hamiltonian with inversion-asymmetry present

To answer the question posed at the end of previous subsection, we will consider here the ellipticity conditions for the case of non-zero B in (9). In this case the ellipticity region in the parameter space $A', B, \gamma'_1, \gamma'_2, \gamma'_3$ is described by the system of inequalities

$$\left\{ \begin{array}{l} \max \{-4\gamma'_2 - 6\gamma'_3, 3\gamma'_3 - 4\gamma'_2, 3\gamma'_3 + 2\gamma'_2\} < \gamma'_1 \\ E + A' + \lambda'_{04} - \sqrt{(E + A' - \lambda'_{04})^2 + 2B^2} < 0 \\ E + A' + \lambda'_{04} + \sqrt{(E + A' - \lambda'_{04})^2 + 2B^2} > 0. \end{array} \right. \quad (19)$$

The first inequality is just a compact form of inequalities 1-3 from (12), the value of λ'_{04} is equal to the one defined above, but written in a new parameter notation $\lambda'_{04} = -\gamma'_1 + 2\gamma'_2 - 3\gamma'_3$.

Despite a more complicated structure than in the situation with zero B , discussed earlier, one out of two B -dependent constraints in (19) is always fulfilled. To be more specific: if $E + A' \geq -\lambda'_{04}$ the third inequality from (19) is redundant, else, the second one is redundant. In each case, the remaining non-redundant inequality leads to the following constraint on B^2

$$B^2 - 2E^2(1+A)(-\gamma'_1 + 2\gamma'_2 - 3\gamma'_3) > 0. \quad (20)$$

The combination of (20) with the first inequality from (19) yields a system of ellipticity constraints for 8×8 ZB Hamiltonian³³ with non-zero B

$$\left\{ \begin{array}{l} \max \{-4\gamma'_2 - 6\gamma'_3, 3\gamma'_3 - 4\gamma'_2, 3\gamma'_3 + 2\gamma'_2\} < \gamma'_1 \\ 2E^2(1+A)(-\gamma'_1 + 2\gamma'_2 - 3\gamma'_3) < B^2. \end{array} \right. \quad (21)$$

Inequality (20) is fulfilled for any B , when the parameter set $A, \gamma'_1, \gamma'_2, \gamma'_3$ satisfies conditions 4-5 from (12). Consequently, the admissible, in terms of (12), material parameters with zero B remain admissible even after the value of B is set to some nonzero number. In other words, B can be treated as an additional fitting parameter to be used in the subsequent adjustment step after the rescaling procedure is performed, but the additional ellipticity preserving readjustment of bandstructure is needed. We are especially interested in correcting the improper dispersion of the CB and SO bands, since it is a major source of errors (see Table III) for most of the rescaled material parameter sets.

Conducted numerical experiments⁵⁶ with different values of B indicate that the increase in $|B|$ leads to the increase in the curvature of conduction and SO bands along ΓK , ΓL and ΓX directions. For InN, that makes the CB adjustment error along those directions smaller at the expense of larger difference in SO band dispersion between the original set and the rescaled parameter set with non-zero B . The indicated behaviour of error and the fact that the error's dominating contribution comes

from CB and SO bands (see FIG. 3 b) mean that there exists an optimal value B that minimizes the error for chosen energy bands. For small errors and $B > 0$ the indicated behaviour is also influenced by the spin-slitting of bands away from Γ . Error can not be minimized for GaSb material, because the rescaled parameters with $B = 0$ already yield visibly higher curvature of CB than the original parameters (see GaSb plot in FIG. 4). We performed the error minimization by adjusting B for each selected parameter set from Table III and confirm that for all materials, except of InN, increase in $|B|$ makes error larger.

The first larger value of $|B| = 15.006$ for InN is a result of the error minimization over two conduction bands only. The conduction band error $\text{err}_{\text{CB}} = 8.352$ meV signifies that the correct dispersion of CB can be recovered almost perfectly by selecting the appropriate $|B|$. These stated differences between the obtained dispersion and the one for original parameter sets are caused by the non-zero spin-splitting of the CB states for $B \neq 0$, which is also witnessed experimentally^{64,65}. To illustrate the effects of spin-splitting and visualize the behaviour of adjustment error, we provide in FIG. 5 bandstructure plots for InN along selected directions together with the plot of direction-wise maximal absolute error between the bandstructure of original set #51 from Table II with $B = 0$ and rescaled set from Table III with $B = 15.006$.

Notice from FIG. 5 b),c) that the spin-splitting is even more evident for LH and SO valence bands than for the conduction band depicted in FIG. 5 a). Direction-wise the magnitude of CB spin-orbit splitting depends on the ratio of the individual momentum components k_x/k_y , k_x/k_z , k_y/k_z . It is non-zero if all these ratios are not zero, infinity or one.

The overall eight-bands' adjustment error $\text{err} = 9.839$ meV is still more three times smaller for the calculated B than for $B = 0$ (green dotted and dash-dotted lines vs red solid line in FIG. 5 d)). Now this error is dominated by the error of SO band that comes from dispersion along ΓK direction (FIG. 5 b). This kind of dominance is typical for the consider materials.

Starting from the same sets of parameters in Table III, we performed another optimization procedure with the aim of verifying at what extent the overall eight-bands' adjustment error can be minimized with help of B . The resulting value of $B = 14.805$ and error $\text{err} = 9.057$ meV are non-significantly differ from the results of the previous optimization procedure. The corresponding band dispersion for InN is visualized in FIG. 6 by using the layout of the previous figure.

The obtained minimal error is around 10 % smaller than the eight-bands' error of the previous optimization procedure and almost 6 times smaller than the bandstructure error of rescaled parameters with zero B (see the last column of Table III).

For InN, the same conclusion can be made from FIG. 6 d), where the band-wise error dispersion (dotted and dash-dotted lines) is plotted along with the error dispersion of the elliptic parameter set with $B = 0$. The overall

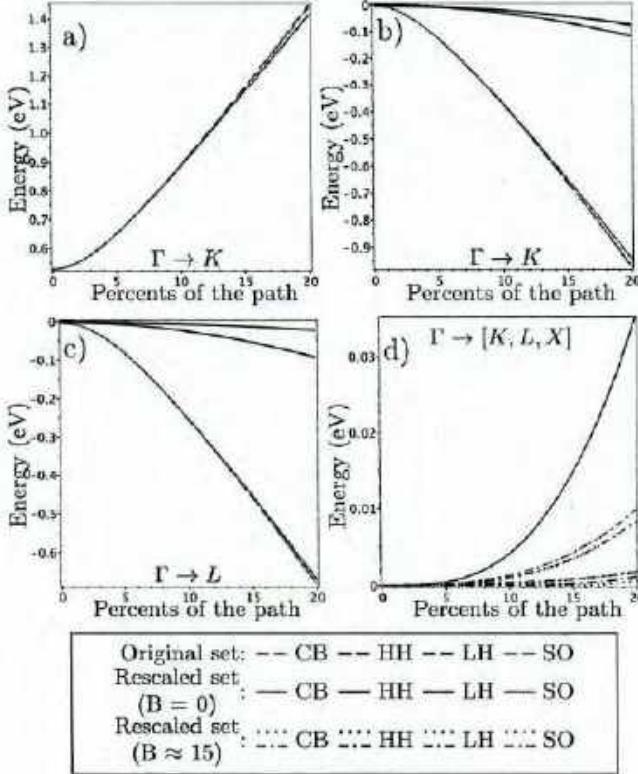


Figure 5. Comparison of original and rescaled parameter sets for InN (color online): original set #43 (solid); rescaled set from Table III with $B = 0$ (dashed) and with optimal $B = 15.006$ (dotted and dash-dotted). Bandstructure along the fraction of symmetry paths: a) Conduction band (CB) along ΓK ; b) Valence bands along ΓK ; c) Valence bands along ΓL ; d) Maximum absolute difference (direction-wise) between original and rescaled sets

error is apparently dominated by the adjustment error of CB and SO bands (FIG. 6 a,b). The errors of other bands are below 1 meV.

The error plots of FIG. 5 and FIG. 6 are also useful to quantify the magnitude of spin-splitting in CB, LH and SO bands for $B = 15.006$ and $B = 14.805$. The comparison of calculated spin-splitting parameters for GaAs and AlAs from Table III and the experimental values for CB along $(1, 1, 0)$ direction suggests that the value of B around 70–80 eV is needed for the 8×8 model to reach the reported experimental values^{66–68}. For such B we observed the deviation in band dispersion of around 0.35–0.5 eV from the dispersion for zero B . Thus, spin-splitting errors are more dominant than the adjustment errors for selected material sets and possibly others (for the experimental values of B see Table 5.5 from [21] and the references wherein). In order to achieve better accuracy with $B \neq 0$ one should use the bandstructure diagram with realistic spin-splitting of bands as an optimization target. Having that in hands, one can possibly get better results by applying the implemented two-step adjustment

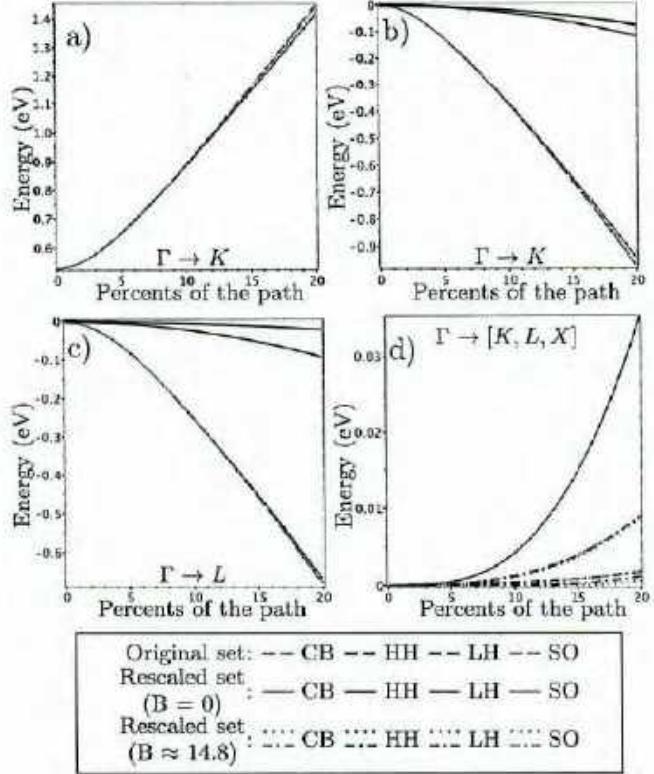


Figure 6. Comparison of original and rescaled parameter sets for InN (color online): original set #43 (solid); rescaled set from Table III with $B = 0$ (dashed) and with optimal $B = 14.805$ (dotted and dash-dotted). Bandstructure along the fraction of symmetry paths: a) Conduction band (CB) along ΓK ; b) Valence bands along ΓK ; c) Valence bands along ΓL ; d) Maximum absolute difference (direction-wise) between original and rescaled sets

procedure to other materials data entries from Table II where the range $(\Delta_{05}^{\min}, \Delta_{05}^{\max})$ for the adjustment parameter Δ_{05} is non-empty.

To clarify the use of $|B|$ in the above calculations we note, that similarly to ellipticity constraints (21), the sign of B^{55} has no effect on the eigenvalues of the Hamiltonian in momentum or position representation. It, however, affects the eigenstates of the Hamiltonian in both representations and therefore must be taken into consideration for experiments^{69,70} that make use of the eigenstates.

One can further increase the accuracy of admissible parameter set of 8×8 Hamiltonian by fitting⁵⁸ the full set of $A, \gamma'_1, \gamma'_2, \gamma'_3, B$ and using inequalities (21) as constraints for the fitting method. Our initial results in that direction show that this is possible for a wide range of materials. Besides, the adjustment with B alone is not a universal substitute for the optimally fitted parameter set $A, \gamma'_1, \gamma'_2, \gamma'_3, B$ because the effect of B on the bandstructure disappears if two out of three momentum components are zero.

Now, let us get back to ellipticity conditions (21). So

far we used B as a bandstructure fitting parameter, after the ellipticity of parameter set was established by rescaling procedure (13), (15), (18) with zero B . The procedure with adjustment of B can be used directly for the materials with the elliptic valence band part (e.g. sets #12, #13, #20, #33). For such materials we can overturn the negative sign of λ'_5 and make the Hamiltonian fully elliptic by setting B to the appropriate (in terms of (20)) nonzero value. This allows to bypass the rescaling procedure altogether, which might be favourable in the light of its phenomenological nature. Another, more physically convenient, way to increase the accuracy of $k \cdot p$ Hamiltonian is to extend its basis set by adding new energy band states. The resulting Hamiltonians are analyzed in the next section.

To conclude the discussion on the ellipticity of 8×8 ZB Hamiltonians we apply the direct adjustment of B to several material parameter sets from Table II; namely sets #12, #13 for InAs, set #20 for AIP and set #33 for AlSb. Only those listed parameter sets yield elliptic valence band part of the Hamiltonian, i.e. the corresponding distance $d = 0$ in Table II. Notably, for each of the three materials the first listed set was reported in [30] – the work that is highly regarded as a source of overall physically consistent material parameters. For InAs, the direct adjustment of B is the only option to obtain the admissible parameter sets based on the data from Table II, because the admissible range $(\Delta_{05}^{\min}, \Delta_{05}^{\max}) \ni \Delta_{05}$ is empty for all its four table entries. For each above mentioned parameter sets we performed bandstructure error minimization procedure over the interval $|B| \in (|B|_{\min}, \infty)$ and reported the resulting value $|B|$, along with the errors in Table IV. Here $|B|_{\min}$ is the minimal solution of (20), with 0.1 substituted in place of 0 in the right-hand side to accommodate for possible numerical errors.

Table IV. Results of direct bandstructure error minimization procedures based on the adjustment of B .

El ^a	$ B _{\min}$	$ B ^b$	err ^c	err _{CB} ^d	err _{HH} ^d	err _{LH} ^d	err _{SO} ^d
InAs ¹²	25.90	25.90	0.073	0.073	0.013	0.061	0.073
InAs ¹³	27.21	27.21	0.081	0.077	0.013	0.064	0.081
AIP ²⁰	5.27	5.27	0.003	0.002	0.001	0.002	0.003
AlSb ³³	4.06	4.06	0.012	0.003	0.002	0.012	0.008

^a Refer to the original dataset number from Table II

^b B that minimizes the bandstructure error

^c The minimal value of the error in eV calculated for the eight bands with a given B over 20 % of the paths ΓL , ΓK and ΓX

^d The errors err_{CB}, err_{HH}, err_{LH}, err_{SO} (all in eV) of conduction bands, heavy holes, light holes and split-off bands accordingly

For brevity we do not provide bandstructure or error plots based on the data from Table IV. Alternatively, we supplied the errors of CB, HH, LH, SO bands as separate entries in the table. For InAs and AIP the introduced bandstructure error is dominated by the differences in CB, SO and LH bands. The situation is different for AlSb for which the main contribution to the error comes from LH bands.

The results reported in Table IV clearly indicate that set #20 for AIP along with $B = 5.272$ lead to the smaller bandstructure adjustment error than the same-material set reported in Table III with $B = 0$. We recommend this for simulations based on the 8×8 ZB Hamiltonian³³ in the position representation. For InAs we suggest using the set 12 with $B = 25.898$. The set obtained as a result of two step optimization, reported above, is the most optimal for InN. For the rest of materials analyzed in this work we recommend the sets from Table III.

V. ELLIPTICITY OF 14×14 BAND MODELS

In this section we focus our attention on the ellipticity of two 14×14 ZB Hamiltonians that are frequently used in the literature^{34,42,71–79}. These two models are based on the extended basis set: six p -like valence band states and two s -like conduction states comprising the basis of 8×8 Hamiltonian studied in the previous section, plus six additional p -like conduction band states. They are introduced to better describe anisotropy of conduction band in the materials like GaAs, InP, InSb, where it is evidenced experimentally^{34,71,73,74}.

The first Hamiltonian proposed by W. Zawadzki, P. Pfeiffer and H. Sigg in [72] and then extended^{34,74} to account for the influence of the out-of-basis bands perturbatively. We base our analysis on this later extended version described by equation (5) from [34]. The calculated⁵⁶ eigenvalues $\lambda''_1 - \lambda''_5$ of the quadratic form associated with this 14×14 ZB Hamiltonian are as follows

$$\begin{aligned}\lambda''_1 &= E(-\gamma''_1 - 4\gamma''_2 - 6\gamma''_3) \\ \lambda''_2 &= E(-\gamma''_1 - 4\gamma''_2 + 3\gamma''_3) \\ \lambda''_3 &= E(-\gamma''_1 + 2\gamma''_2 + 3\gamma''_3) \\ \lambda''_4 &= E(-\gamma''_1 + 2\gamma''_2 - 3\gamma''_3) \\ \lambda''_5 &= E.\end{aligned}\quad (22)$$

The CB part of the Hamiltonian is elliptic by design, since $\lambda''_5 > 0$ independently of materials parameters. The ellipticity of the valence-band part is guaranteed when $\lambda''_1 - \lambda''_4$ are all negative simultaneously. So in the end, we are getting exactly the same ellipticity conditions as for the 6×6 ZB Hamiltonian, albeit with the different Luttinger-like parameters (compare the above $\lambda''_1 - \lambda''_4$ with $\lambda'_1 - \lambda'_4$ from (7)). These new Luttinger-like parameters $\gamma''_1, \gamma''_2, \gamma''_3$ can be obtained from the conventional Luttinger parameters by subtracting from $\gamma_1, \gamma_2, \gamma_3$ the contributions of p -like CB bands, that are no longer treated perturbatively. More precisely,

$$\begin{aligned}\gamma''_1 &= \gamma'_1 - \frac{Q^2}{3EE'_0} - \frac{Q^2}{3E(E'_0 + \Delta'_0)}, \\ \gamma''_2 &= \gamma'_2 + \frac{Q^2}{6EE'_0}, \quad \gamma''_3 = \gamma'_3 - \frac{Q^2}{6EE'_0}.\end{aligned}\quad (23)$$

Here E_0 is a fundamental bandgap, E'_0 is a gap between first two bottommost conduction bands, Δ_0, Δ'_0 are the

Table V. The material data for 14×14 ZB Hamiltonian³⁴, d – distance from the point $(\gamma_1'', \gamma_2'', \gamma_3'')$ to the feasibility region Λ_- . The positive values of $\lambda_1''/E, \lambda_2''/E, \lambda_3''/E, \lambda_4''/E$ are printed in bold.

# El ^a	γ_1''	γ_2''	γ_3''	λ_1''/E	λ_2''/E	λ_3''/E	λ_4''/E	d
1 GaAs ^b	0.18	0.42	0.11	-2.49	-1.55	0.98	0.35	0.26
2 GaAs ^b	-0.59	-0.02	-0.34	2.69	-0.34	-0.46	1.55	0.41
3 GaAs ^c	-1.48	-0.03	-0.61	5.23	-0.24	-0.39	3.26	0.87
4 AlAs ^c	-0.91	0.34	-0.31	1.38	-1.39	0.67	2.52	0.67
5 InAs ^c	0.76	0.59	0.12	-3.84	-2.73	0.78	0.05	0.21
6 GaP ^c	-1.55	-0.16	-0.84	7.25	-0.33	-1.31	3.75	1.00
7 AlP ^c	-1.22	0.22	-0.46	3.10	-1.04	0.28	3.04	0.81
8 InP ^d	0.44	0.46	-0.13	-1.49	-2.67	0.08	0.87	0.23
9 InP ^e	-0.50	-0.15	-0.68	5.19	-0.94	-1.85	2.24	0.71
10 InP ^c	-1.54	-0.03	-0.66	5.63	-0.34	-0.50	3.48	0.93
11 GaSb ^c	-0.39	0.59	-0.03	-1.79	-2.04	1.48	1.65	0.44
12 AlSb ^c	-0.99	0.58	-0.32	0.61	-2.31	1.18	3.13	0.84
13 InSb ^c	-2.89	-0.92	-1.61	16.20	1.75	-3.77	5.87	2.23

^a Set 1 from [34] ($\alpha = 0.065$) ^b Set 2 from [34] ($\alpha = 0.085$)

^c Parameters obtained via (23) from the data in [80] and [30]

^d Set 1 from [34] ($\alpha = 0.12$) ^e Set 2 from [34] ($\alpha = 0.2$)

spin-splitting parameters of the valence and conduction bands correspondingly, Q is the interband momentum matrix element (see Fig 1 in [34]). For a complete description of the Hamiltonian one additionally needs to define other material dependent parameters $P'_0, \bar{\Delta}, \kappa, C_k$. Those are determined by fitting the bandstructure to experimental data³⁴. For that reason, we are focused only on the sources where the full sets of fitted Hamiltonian parameters have been reported in the context of the considered 14×14 model.

Beside the sets $\gamma_1'', \gamma_2'', \gamma_3''$ from the original paper³⁴, that provides them for GaAs and InP explicitly, we used the parameters sets from J.-M. Jancu et al. [80] and recommended there Luttinger parameters from [30] to calculate the respective values of $\gamma_1'', \gamma_2'', \gamma_3''$ for GaAs, AlAs, InAs, GaP, AlP, InP, GaSb, AlSb and InSb. All calculated parameters along with the results of ellipticity analysis are collected in Table V.

Each of the considered in Table V sets fails two out of four ellipticity constraints except the sets for AlAs, AlP and InSb, for which three ellipticity constraints are violated. For GaAs set 1 from Table V we can confirm a reduction of distance d to the feasibility region Λ_- in the space $\gamma_1'', \gamma_2'', \gamma_3''$ compared to the best of GaAs sets for 6×6 and 8×8 Hamiltonians. This set and set #8 are taken from the original work. Both sets were calculated using cyclotron resonance experiments³⁴. The second pair of sets #2, #9, which are deemed more consistent experimentally³¹, is slightly off the region Λ_- ; but the corresponding values of d are within the range of the same-material values of d from Table II. Parameter sets for other materials are even further away from Λ_- than the un-rescaled same-materials sets for 8×8 Hamiltonian.

The observed increase of the distance to Λ_- seems to be theoretically unfounded, especially in the view of for-

mula (3). Recall, that the relative norm⁴⁹ of the perturbative term from (3) should decrease after eigenstates are moved from perturbative class (class B) in to the basis (class A). It can be explained as follows.

In the 8×8 model the influence of valence bands on the CB states was represented directly by the parameters P_0, B and, we suppose, indirectly by the perturbative CB parameter A' . The absence of A' in the CB eigenvalue from (22) suggest that the 14×14 model was derived under assumption that A' depends on the upper CB states only, now included in the basis. In such a situation, all cross-influence between valence and conduction bands are incorporated into P and Q by using fitting to experimental data. Then it is propagated to $\gamma_1'', \gamma_2'', \gamma_3''$ with help of formula (23). But the terms $\gamma_1', \gamma_2', \gamma_3'$ in the right-hand side of (23) were fitted to experiments under assumption of the non-zero valence band contribution to A' . That explains why the parameter triplets $\gamma_1'', \gamma_2'', \gamma_3''$ for materials with smaller fundamental bandgap E_0 (InAs, GaSb, InSb) end up having larger d .

On the other hand, the conduction band states in the materials with larger E_0 (AlAs, AlP, AlSb) may in reality be influenced by the higher bands not included in the basis. That influence is assumed to be zero in the model, because CB eigenvalues are equal to E even for the newly included in the basis p -like bands. If non-negligible, the influence is accounted by P, Q and then propagated to $\gamma_1'', \gamma_2'', \gamma_3''$ by the mechanism described above. That explain the increase in d for large-bandgap materials from Table V (AlAs, AlP, AlSb).

For some parameter sets from Table V the ellipticity might be corrected by rescaling of P, Q in a way similar to the rescaling procedure from Section IV. This will, of course, affect the accuracy of bandstructure and therefore must involve the optimization procedure with respect to the parameters P, Q and possibly $\gamma_1'', \gamma_2'', \gamma_3''$, if our hypothesis holds true.

Now we proceed to the second implementation of 14×14 ZB Hamiltonian model. The initial version of this model was derived by U. Rössler using the theory of invariants⁸² and then extended in the work of H. Mayer and U. Rössler⁷³ by adding first-order perturbative corrections to the lowest conduction and upper valence bands. The most recent version of the Hamiltonian was provided by R. Winkler³⁵. It additionally includes the second order conduction-valence band mixing parameters similar to B from Kane Hamiltonian (9).

All three mentioned versions of 14×14 ZB Hamiltonian are connected by the common assumption that six second-order diagonal terms related to the newly added p -like CB states are neutralized by the counter-influence of other bands, e.g. the representation of H_{scbe}, H_{8sc} from Table C.5 of [35]. In terms of ellipticity such an assumption results in the presence of zero eigenvalue among the set of eigenvalues of the quadratic form associated with this implementation. Our calculations⁵⁶ confirm that. Therefore, this Hamiltonian is not elliptic by design.

It is worthwhile pointing that, unlike first, the second implementation of 14×14 Hamiltonian^{35,73,82} can be regarded as an extension of Kane model studied in section IV. The Hamiltonian contains the perturbative correction to s -like CB states and the conduction-valence band mixing parameters. So, all inter-band interaction effects embodied in the 8×8 representation³³ can be properly accounted for. In our opinion, two analyzed implementations of the 14×14 Hamiltonian model are less universal material-wise than 8×8 Hamiltonians, despite being more accurate at describing CB related phenomena^{78,79}. For such higher band models, the assumptions regarding the interactions of in-basis conduction band states require a revision.

CONCLUSIONS

We performed a systematic study of ellipticity conditions for 6×6 , 8×8 , 14×14 $k \cdot p$ Hamiltonians in the bulk zinc blende crystals. The conditions take roots in the fundamental axioms of quantum mechanics concerning the description of observable states and properties of Hamiltonian as a differential operator. They appear in the form of constraints on the values of material parameters pertaining to the second-order-in- k terms from the Hamiltonian in the momentum representation.

For 6×6 and 8×8 models we examined an extensive number of parameter sets for GaAs, AlAs, InAs, GaP, AlP, InP, GaSb, AlSb, InSb, GaN, AlN, InN and C that are gathered from the widely accepted sources of reference literature on material parameters^{29–31}. The results of the performed analysis reaffirm earlier conclusions on the violation of Hamiltonian ellipticity^{16,27} and its cause^{19,41}. Furthermore, we demonstrated that this violation is a much more common problem material-wise. Among all analyzed materials only carbon has parameter sets that make 6×6 Hamiltonians elliptic and therefore admissible from a theoretical point of view. Other sets of material parameters incur violation of one out of four ellipticity constraints: $2\gamma_2 - \gamma_1 + 3\gamma_3 < 0$. This can be traced to a non-negligible influence of conduction bands on the heavy-hole and light-hole, accounted perturbatively in $\gamma_1, \gamma_2, \gamma_3$. We conclude that this model is not accurate enough to describe all considered bands reliably and to remain elliptic at the same time.

The situation becomes more complex for 8×8 Hamiltonians, where the bottom-most conduction band is included into the basis. None of the analyzed parameter sets are admissible, because the conduction-band ellipticity constraint is violated for all sets, when the absence of inversion asymmetry ($B = 0$) is assumed. However, the degree of non-ellipticity in the valence-band part of the Hamiltonian, which we characterize in terms of distance to the feasibility region Λ_- in the space of Luttinger-like parameters, decreases. Several parameter sets for InAs, AlP and AlSb satisfy the ellipticity constraints for the valence-band part. It corroborates the evidence on the

perturbative source of non-ellipticity. We note, that in the case of $B = 0$, these constraints have the same structure as the ellipticity constraints for 6×6 Hamiltonians.

As one possible way to remedy the situation with the lack of ellipticity in the 8×8 model we propose a parameter rescaling procedure. It is based on the idea of adjusting the first-order conduction-valence mixing parameter P_0 to change $A, \gamma'_1, \gamma'_2, \gamma'_3$ and make the Hamiltonian elliptic. The proposed here rescaling procedure accounts for a full set of ellipticity constraints and thus improve the previous approaches^{20,57}, targeted solely at imposing the conduction band constraint $1 + A > 0$.

The results of the rescaling procedure for all materials except InAs, are presented in Table III. Each of the admissible sets is made via (18) from one of the original sets that lead to a minimal absolute difference in the parameter A per material. Notwithstanding the attempt to minimize the effects of rescaling on the band-structure, these effects are negligible (≤ 11 meV) only for AlP, AlSb, AlAs, InP and GaN (see FIG. 3). They may be considered small (≤ 50 meV) for GaP, InSb, InN and can not be ignored for the rest of materials from the table (see FIG. 4 for visual comparisons). For them, the rescaling leads to a noticeable change in the conduction band dispersion. Heavy hole (HH) and light hole (LH) bands remain visually unaffected even though the differences are non-zero. The rescaling also causes an increase in the curvature of the split-off (SO) band which makes it the main source of total valence-band adjustment error. For all mentioned materials excluding GaP and AlN, the magnitude of this error is proportional to the relative change in E_p . Therefore, in most cases the Hamiltonian based on the new parameters is elliptic, yet incapable of reliably describing the conduction-valence band transition phenomena, except those occurring at the band edge.

In attempt to counter for the observed bandstructure discrepancies and to derive the admissible parameter set for InAs we consider the use of B as an additional adjustment parameter. This requires a generalization of the ellipticity conditions for the 8×8 Hamiltonian to the case of non-zero B . Recall, that, due to the inversion asymmetry, this case is theoretically more relevant to the majority of zinc blende materials. The form of the generalized ellipticity conditions allows us to draw two important conclusions.

First, setting B to some nonzero value will not break ellipticity of 8×8 Hamiltonians if the parameter set – the Hamiltonian is based upon – is admissible with zero B . We use this property to calculate two distinct values of B for the materials. The larger value of B minimizes the error of conduction band and makes the errors in the dispersion of other (most notably SO) bands larger. We left aside a discussion on physical relevance of the calculated values of B and presented the bandstructure plots with $B \neq 0$ for InN only. Our intent here has been to show that B -adjustment can be used to partially correct the bandstructure distorted by rescaling.

Second, the parameter B could not be set to zero for the materials where $1 + A' < 0$ without sacrificing ellipticity of the 8×8 Hamiltonian. At the same time, the adjustment of B can be used to correct the ellipticity of CB part of the Hamiltonian provided that the valence-band ellipticity constraints are fulfilled by $\gamma'_1, \gamma'_2, \gamma'_3$. We discovered four parameter sets for InAs, AlP, AlSb where this is true. These parameter sets are highly regarded as overall physically consistent³⁰. Four minimally admissible values of B that complement each of the mentioned sets to make the Hamiltonian elliptic, are collected in Table IV.

Besides being the source of admissible B , the data from Table IV illustrates that the admissible sets obtained by the rescaling procedure are not the best option in terms of the bandstructure fit, at least for AlP. We postulate that there exist admissible parameters of 8×8 Hamiltonian³³ better in terms of bandstructure error for other materials too. To find them one should fit the Hamiltonian bands' dispersions to the spin-resolved bandstructure by adjusting the entire set of the Hamiltonians parameters simultaneously and by using the ellipticity conditions as constraints for the fit. This idea is supplemented by the fact, that the ellipticity region in the space of parameters that satisfy the constraints is convex and connected. The results of such fitting procedure will be also useful to quantify the limits of this and other $k \cdot p$ models, assuming that the bandstructure used as a fitting target is reliable⁸³. This would constitute an important for the future studies.

Finally, we analyzed two popular implementations of the 14×14 ZB Hamiltonian model^{34,35}. The ellipticity of the first implementation is described by precisely the same set of constraints as for the 6×6 model, but written in terms of the reduced Luttinger-like parameters $\gamma''_1, \gamma''_2, \gamma''_3$. Unfortunately, the parameter analysis shows that none of the available sets for a studied list of materials is admissible in terms of the Hamiltonian ellipticity. We conclude that an overly-strict set of assumptions regarding the perturbative influence of outer bands on the model's conduction bands is to be responsible for the lack of ellipticity. The second analysed 14×14 Hamiltonian is more general in that regard. It is however non-elliptic by design owing to the fact that the second-order-in- k terms are zero for three upper p -like conduction bands.

Based on the supplied evidence we surmise that both 14×14 implementations are less universal than the previously studied 8×8 Hamiltonian. The revision of indicated

assumptions and, perhaps, some unifications are necessary to bring these extended models to a strict theoretical ground.

The analysis conducted in this paper covers possible extensions of the considered models, such as the inclusion strain-stress, electromagnetic or other phenomena, as long as such extensions do not change the structure of second-order-in- k terms of the Hamiltonian. The analysis can be easily transferred to the cases when momentum quantization is applied in one- or two-dimensions (quantum wells, wires, etc). It can also be applied to the materials, that are intrinsically non-three dimensional (non-3D), like graphene, silicene or others; especially given that many high-accuracy bandstructure diagrams for those kind of materials are readily available⁸⁴⁻⁸⁶.

We note that ellipticity conditions stated here are not valid for anything other than the considered three dimensional $k \cdot p$ Hamiltonians for ZB crystals. The whole analysis will have to be repeated in each specific non-3D case.

The applications to materials with other-than-zinc-blende crystal structures are also possible. The Hamiltonian parameters for ternary alloys, for instance, are typically calculated by using a linear combination of the parameters for the constituents. Therefore the alloys' parameters will be elliptic if the parameters of constituents are, because the ellipticity region is connected and convex in the space of parameters. Similar reasoning can be applied to the calculation of time-dependent Hamiltonian parameters with help of the Varshni formulas. There might be some complications with analytic calculation of quadratic form's eigenvalues, for more complicated Hamiltonians with different symmetry-structure. This is not a major issue, because for any specific material the ellipticity of $k \cdot p$ Hamiltonian can also be verified numerically.

ACKNOWLEDGEMENTS

The first author acknowledges the partial financial support from The Royal Society of Canada via 2017 RSC-Ukraine exchange program. Both authors are grateful to Dr. Sunil Patil for our earlier discussions on the topic. The support of NSERC and the CRC Program is also acknowledged.

* sytnikd@gmail.com

[†] rmelnik@wlu.ca

¹ J. M. Luttinger and W. Kohn, Phys. Rev. **97**, 869 (1955).

² E. O. Kane, Journal of Physics and Chemistry of Solids **1**, 249 (1955).

³ S. Prabhakar, R. V. Melnik, P. Neittaanmäki, and T. Tiisanen, Journal of Computational and Theoretical Nanoscience **10**, 534 (2013).

⁴ S. Prabhakar, R. V. Melnik, P. Neittaanmäki, and T. Tiisanen, Physica E: Low-dimensional Systems and Nanostructures **46**, 97 (2012).

⁵ J. M. Xia, Phys. Rev. B **43**, 9856 (1991).

⁶ S. Prabhakar, R. Melnik, and L. L. Bonilla, Journal of Applied Physics **113**, 244306 (2013).

- ⁷ M. Alvaro, L. Bonilla, M. Carretero, R. Melnik, and S. Prabhakar, *Journal of Physics: Condensed Matter* **25**, 335301 (2013).
- ⁸ S. R. Patil and R. V. N. Melnik, *Nanotechnology* **20**, 125402 (13pp) (2009).
- ⁹ F. J. G. de Abajo, *Reviews of Modern Physics* **79**, 1267 (2007).
- ¹⁰ S. L. Chuang, *Wiley series in pure and applied optics* (Wiley, New York, 1995) p. 736.
- ¹¹ S. L. Chuang, *Physics of photonic devices* (John Wiley & Sons, Hoboken, N.J., 2009) p. 821.
- ¹² D. L. Smith and C. Mailhot, *Phys. Rev. B* **33**, 8345 (1986).
- ¹³ F. Szmulowicz, *Phys. Rev. B* **54**, 11539 (1996).
- ¹⁴ B. A. Foreman, *Phys. Rev. B* **75**, 235331 (2007).
- ¹⁵ W. Yang and K. Chang, *Phys. Rev. B* **72**, 233309 (2005).
- ¹⁶ R. G. Veprek, S. Steiger, and B. Witzigmann, *Phys. Rev. B* **76**, 165320 (2007).
- ¹⁷ R. G. Veprek, S. Steiger, and B. Witzigmann, *Journal of Computational Electronics* **7**, 521 (2008).
- ¹⁸ B. Lassen, R. Melnik, and M. Willatzen, *Commun. Comput. Phys.* **6**, 699 (2009).
- ¹⁹ D. Sytnyk, S. Patil, and R. Melnik, ArXiv e-prints (2010), arXiv:1004.4152 [cond-mat.mtrl-sci].
- ²⁰ B. A. Foreman, *Phys. Rev. B* **56**, R12748 (1997).
- ²¹ X. Cartoixa Soler, *Theoretical methods for spintronics in semiconductors with applications*. Ph.D. thesis, California Institute of Technology (2003).
- ²² K. I. Kolokolov, J. Li, and C. Z. Ning, *Phys. Rev. B* **68**, 161308 (2003).
- ²³ R. Eppenga, M. F. H. Schuurmans, and S. Colak, *Phys. Rev. B* **36**, 1554 (1987).
- ²⁴ M. V. Kisin, B. L. Gelmont, and S. Luryi, *Phys. Rev. B* **58**, 4605 (1998).
- ²⁵ T. Eissfelder and P. Vogl, *Phys. Rev. B* **84**, 195122 (2011).
- ²⁶ V. Berestetskii, E. Lifshitz, and L. Pitaevskii, *Quantum Electrodynamics*, 2nd ed., Course of Theoretical Physics, Vol. 4 (Pergamon Press, 1982).
- ²⁷ R. G. Veprek, S. Steiger, and B. Witzigmann, *Opt. Quantum Electron.* (2009), 10.1007/s11082-008-9259-9.
- ²⁸ P. Lawaetz, *Phys. Rev. B* **4**, 3460 (1971).
- ²⁹ O. Madelung, U. Rössler, and M. Schulz, *Group IV Elements, IV-IV and III-V Compounds. Part b - Landolt-Börnstein - Group III Condensed Matter Numerical Data and Functional Relationships in Science and Technology*, Vol. 41A1b (Springer-Verlag, 2002).
- ³⁰ I. Vurgaftman, J. R. Meyer, and L. R. Ram-Mohan, *Journal of Applied Physics* **89**, 5815 (2001), <https://doi.org/10.1063/1.1368156>.
- ³¹ O. Madelung, *Semiconductors : data handbook* (Springer, Berlin : New York, 2004) p. 691.
- ³² P. Rinke, M. Winkelkemper, A. Qteish, D. Bimberg, J. Neugebauer, and M. Scheffler, *Phys. Rev. B* **77**, 075202 (2008).
- ³³ T. B. Bahder, *Phys. Rev. B* **41**, 11992 (1990).
- ³⁴ P. Pfeffer and W. Zawadzki, *Phys. Rev. B* **53**, 12813 (1996).
- ³⁵ R. Winkler, *Spin-orbit coupling effects in two-dimensional electron systems* (Springer, Berlin; New York, 2003) p. 238.
- ³⁶ E. O. Kane, "Energy band theory," (North-Holland, 1982) Chap. 4A, pp. 193–216.
- ³⁷ L. Voon and M. Willatzen, *The k p Method: Electronic Properties of Semiconductors* (Springer Science+Business Media, 2009).
- ³⁸ G. Bir and G. Pikus, *Symmetry and Strain-Induced Effects in Semiconductors* (Wiley, New York, 1974) p. 484, translated from Russian by P. Shehitz.
- ³⁹ J. M. Luttinger, *Phys. Rev.* **102**, 1030 (1956).
- ⁴⁰ T. Kato, *Perturbation theory for linear operators*. (Grundlehren der mathematischen Wissenschaften. Band 132. Berlin-Heidelberg-New York: Springer-Verlag. xx, 592 p. with 3 figures DM 79.20, 1966).
- ⁴¹ D. Sytnyk, *Mathematical modeling of quantum dots with generalized envelope functions approximations and coupled partial differential equations*, Master's thesis, Wilfrid Laurier University (Canada) (2010).
- ⁴² P. Pfeffer and W. Zawadzki, *Phys. Rev. B* **41**, 1561 (1990).
- ⁴³ S. I. Dorozhkin, *JETP Lett.* **88**, 819 (2008).
- ⁴⁴ L. E. Ballentine, *Quantum mechanics: a modern development* (World Scientific, Singapore, 2006) p. 672.
- ⁴⁵ P. A. M. Dirac, *The principles of quantum mechanics* (Clarendon Press, Oxford, UK, 1981) p. 340.
- ⁴⁶ G. Teschl, *Mathematical methods in quantum mechanics*, Graduate Studies in Mathematics, Vol. 99 (American Mathematical Society, Providence, RI, 2009) pp. xiv+305, with applications to Schrödinger operators.
- ⁴⁷ M. G. Burt, *J. Phys.: Condens. Matter* **11**, 53 (1999), 53.
- ⁴⁸ L. Hörmander, *The analysis of linear partial differential operators. II: Differential operators with constant coefficients*. (Springer-Verlag: Berlin Heidelberg-NewYork - Tokyo, 1983).
- ⁴⁹ Y. Egorov and M. Shubin, *Foundations of the classical theory of partial differential equations. Transl. from the Russian by R. Cooke. 2nd printing of the 1st ed. 1992.* (Berlin: Springer, 1998) p. 259.
- ⁵⁰ L. Hörmander, *The analysis of linear partial differential operators. I: Distribution theory and Fourier analysis*. (Springer-Verlag: Berlin Heidelberg-NewYork - Tokyo, 1983).
- ⁵¹ P. Y. Yu and M. Cardona, *Fundamentals of semiconductors: physics and materials properties*, 3rd ed. (Springer, 2005) p. 639.
- ⁵² R. Courant and P. D. Lax, *Proc. Natl. Acad. Sci. U. S. A.* **42**, 872 (1956).
- ⁵³ Electronic Transport, D Opt Waechter, Other Authors Zukotynski, *Phys. Rev. B* **28**, 3550 (1983).
- ⁵⁴ C. R. Pidgeon and R. N. Brown, *Phys. Rev.* **146**, 575 (1966).
- ⁵⁵ M. Cardona, N. Christensen, M. Dobrowolska, J. Furyna, and S. Rodriguez, *Solid State Communications* **60**, 17 (1986).
- ⁵⁶ All computations were performed in Maple. Codes are available at www.imath.kiev.ua/~sytnik/projects/kp.
- ⁵⁷ S. Birner, in *Multi-Band Effective Mass Approximations* (Springer, 2014) pp. 193–244.
- ⁵⁸ C. M. O. Bastos, F. P. Sabino, P. E. F. Junior, T. Campos, J. L. F. D. Silva, and G. M. Sipahi, *Semiconductor Science and Technology* **31**, 105002 (2016).
- ⁵⁹ G. Dresselhaus, A. F. Kip, and C. Kittel, *Phys. Rev.* **98**, e368 (1955).
- ⁶⁰ I. Vurgaftman and J. R. Meyer, *Journal of Applied Physics* **94**, 3675 (2003), <https://doi.org/10.1063/1.1600519>.
- ⁶¹ F. Issiki, S. Fukatsu, and Y. Shiraki, *Applied Physics Letters* **67**, 1048 (1995), <https://doi.org/10.1063/1.114460>.

- ⁶² F. Viñas, H. Q. Xu, and M. Leijnse, Phys. Rev. B **95**, 115420 (2017).
- ⁶³ S. Birner, T. Zibold, T. Andlauer, T. Kubis, M. Sabathil, A. Trellakis, and P. Vogl, IEEE Transactions on Electron Devices **54**, 2137 (2007).
- ⁶⁴ Z. Zhang, R. Zhang, B. Liu, Z. Xie, X. Xiu, P. Han, H. Lu, Y. Zheng, Y. Chen, C. Tang, et al., Solid State Communications **145**, 159 (2008).
- ⁶⁵ F. Mei, N. Tang, X. Wang, J. Duan, S. Zhang, Y. Chen, W. Ge, and B. Shen, Applied Physics Letters **101**, 132404 (2012).
- ⁶⁶ R. Eppenga and M. F. H. Schuurmans, Phys. Rev. B **37**, 10923 (1988).
- ⁶⁷ B. Jusserand, D. Richards, G. Allan, C. Priester, and B. Etienne, Phys. Rev. B **51**, 4707 (1995).
- ⁶⁸ D. Richards, B. Jusserand, G. Allan, C. Priester, and B. Etienne, Solid-State Electronics **40**, 127 (1996), proceedings of the Seventh International Conference on Modulated Semiconductor Structures.
- ⁶⁹ A. N. Chantis, M. Cardona, N. E. Christensen, D. L. Smith, M. van Schilfgaarde, T. Kotani, A. Svane, and R. C. Albers, Phys. Rev. B **78**, 075208 (2008).
- ⁷⁰ A. N. Chantis, N. E. Christensen, A. Svane, and M. Cardona, Phys. Rev. B **81**, 205205 (2010).
- ⁷¹ M. Braun and U. Rossler, Journal of Physics C: Solid State Physics **18**, 3365 (1985).
- ⁷² W. Zawadzki, P. Pfeffer, and H. Sigg, Solid State Communications **53**, 777 (1985).
- ⁷³ H. Mayer and U. Rössler, Phys. Rev. B **44**, 9048 (1991).
- ⁷⁴ W. Zawadzki, I. T. Yoon, C. L. Little, X. N. Song, and P. Pfeffer, Phys. Rev. B **46**, 9469 (1992).
- ⁷⁵ P. Pfeffer and W. Zawadzki, Phys. Rev. B **74**, 233303 (2006).
- ⁷⁶ N. Cavassilas, F. Aniel, K. Boujdaria, and G. Fishman, Phys. Rev. B **64**, 115207 (2001).
- ⁷⁷ R. D. R. Bhat, P. Nemec, Y. Kerachian, H. M. van Driel, J. E. Sipe, and A. L. Smirl, Phys. Rev. B **71**, 035209 (2005).
- ⁷⁸ M. El kurdi, G. Fishman, S. Sauvage, and P. Boucaud, Phys. Rev. B **68**, 165333 (2003).
- ⁷⁹ M. Gladysiewicz, R. Kudrawiec, and M. S. Wartak, Journal of Applied Physics **118**, 055702 (2015), <https://doi.org/10.1063/1.4927922>.
- ⁸⁰ J.-M. Jancu, R. Scholz, E. A. de Andrada e Silva, and G. C. La Rocca, Phys. Rev. B **72**, 193201 (2005).
- ⁸¹ I. Gorczyca, P. Pfeffer, and W. Zawadzki, Semiconductor Science and Technology **6**, 963 (1991).
- ⁸² U. Rössler, Solid State Communications **49**, 943 (1984).
- ⁸³ T. Okuda, Journal of Physics: Condensed Matter **29**, 483001 (2017).
- ⁸⁴ D. Massatt, S. Carr, M. Luskin, and C. Ortner, Multiscale Modeling & Simulation **16**, 429 (2018).
- ⁸⁵ H. Pan, Z. Li, C.-C. Liu, G. Zhu, Z. Qiao, and Y. Yao, Physical review letters **112**, 106802 (2014).
- ⁸⁶ W. Lin, J. Li, W. Wang, S.-D. Liang, and D.-X. Yao, Scientific reports **8**, 1674 (2018).



Proceedings of the First Fields-MITACS Industrial Problems Workshop

The Fields Institute for Research in Mathematical Science
Toronto, Ontario

August 14-18, 2006

Editors:

Dhavide A. Aruliah
(University of Ontario Institute of Technology)
and

Gregory M. Lewis
(University of Ontario Institute of Technology)

Sponsored by:

The Fields Institute for Research in Mathematical Science
and
Mathematics of Information Technology and Complex Systems
(MITACS)

Nonlinear Dimension Reduction for Microarray Data (Small n and Large p)

Problem Presenter: Christopher Bowman (National Research Council of Canada)

Academic Participants: Dhavide Aruliah (University of Ontario Institute of Technology), Guangzhe Fan (University of Waterloo), Roderick Melnik (Wilfred Laurier University), Suzanne Shontz (Pennsylvania State University), Steven Wang (York University), Jiaping Zhu (University of Waterloo)

Report prepared by: Suzanne Shontz¹

1 Introduction

Over the last decade or so, researchers have developed techniques for measuring the expression level of many genes in an organism simultaneously. One such technique is the cDNA microarray [13, 11]. Such techniques generate a torrent of data that can be used to then learn more about gene functions, response to stimuli, and interactions.

A cDNA microarray is a glass slide on which many (usually thousands) segments of DNA (often genes, but not always) are attached in distinct spots. Messenger RNA is then extracted from two different populations of cells (for example, cancer and normal tissue) and reverse transcribed to complementary DNA (cDNA). Each of the two sets of cDNA is tagged with a molecule of fluorescent dye; usually they are red and green, respectively. The cDNA solutions are then washed over the glass slide and hybridized with the genetic material spotted onto the slide. When a molecule of cDNA matches the DNA spotted onto the slide, it reacts and binds to it, bringing along the fluorescent dye molecule. The greater the number of copies of the appropriate piece of cDNA present in the sample, the greater the number of dye molecules which will bind to that particular spot, creating a stronger signal. If red and green dyes are used, the spots will appear to fluoresce with varying intensities of red, green, and yellow (when cDNA from both samples bind to the spot). These intensities can be measured by a scanner to determine the relative expression level of each gene in each of the two cell populations.

Microarray experiments thus typically have thousands of variables explaining each individual sample in the experiment and typically only a handful (a few hundred at most, often

¹shontz@cse.psu.edu

many fewer) distinct samples. Furthermore, the thousands of genes in an organism are not independent entities, and each reacts to the activation level of other genes in a complicated and not well-understood way. This raises a mathematical challenge. Each experimental sample can be viewed as a point in a space of dimension equal to the number of genes being measured. The question is, can one find a lower-dimensional space in which to work? Or, stated more precisely, given a set of n points in an p -dimensional linear space, drawn from some unknown distribution V , find a d -dimensional (possibly nonlinear, $d \ll p$) manifold that approximates the points well?

There are many measures for determining whether or not a lower-dimensional manifold approximates the points well. The simplest is that the orthogonal distance from each point to the manifold should be minimised, but this is by no means the only choice, and other options will be discussed below. It is important to note, however, that any notion of goodness-of-fit should apply not just to the data that has already been collected, but also to future data points drawn from the distribution V . To ensure this, a cross-validated estimate of error must be used, for example, the leave-one-out error described below.

2 Filtering the data

2.1 Motivation for filtering. Typical microarray data have quite high dimensionality due to the number of genes involved. For example, the simplest biological model, yeast, has more than 6000 genes. In 2003, estimates from gene-prediction programs suggested there might be as many as 24,500 protein-coding genes [12]. The Ensemble genome-annotation system estimates their number at 23,299. Therefore, the dimensionality is very high. On the other hand, the number of observations that is available is usually very low due to fact the microarray experiments are too expensive to produce many replications. This is known as the problem of “*large p and small n*”.

Although there are many human genes, often medical researchers are only concerned with a dozen or fewer genes if they are interested in one particular disease. Therefore it is not necessary to consider all the genes in the analysis of microarray data. Furthermore, the information or “signal” for the genes of interest could be overwhelmed by the genes that are not relevant to the current analysis. Dimensionality reduction is necessary given the fact that there are not many observations that are scattered in very high-dimensional space.

Furthermore, the filtering is crucial for any data mining technique to work. For example, the clustering procedure is often applied to microarray data to divide the large-dimensional space into subspaces such that the subspaces are much more manageable than the whole space. However, most clustering algorithms would rely on a proper choice of distance function. There are many distance functions proposed in the literature. We will demonstrate our argument by using the most commonly used distance function, *i.e.*, Euclidean distance. To make our argument more transparent, we suppose that there are p random variables that are independent and identically distributed with a standard normal distribution, *i.e.*

$$X_1, X_2, \dots, X_p \sim N(0, 1).$$

Let us further assume that only Y_1 and Y_2 are important or relevant to us. Furthermore, we assume that

$$Y_1 \sim N(10, 1) \quad \text{and} \quad Y_2 \sim N(20, 1)$$

for the observations from the treatment group and $Y_1, Y_2 \sim N(0, 1)$ for the control group.

However, the vector $(Y_1, Y_2, X_1, X_2, \dots, X_p)$ will not be informative if the Euclidean distance is used. Let $O_k = (Y_{k1}, Y_{k2}, X_{k1}, X_{k2}, \dots, X_{kp})$ and $O_j = (Y_{j1}, Y_{j2}, X_{j1}, X_{j2}, \dots, X_{jp})$

represent two observations in a microarray data set. Note that $Y_{k1} \sim N(10, 1)$ and $Y_{k2} \sim N(20, 1)$.

It can be verified that

$$\text{dist}(O_j - O_k)^2 = \sqrt{(Y_{k1} - Y_{j1})^2 + (Y_{k2} - Y_{j2})^2 + \chi^2(p)}$$

where $\chi^2(p)$ denotes the central χ^2 distribution with mean p and variance $2p$.

It can be verified that

$$\frac{(Y_{k1} - Y_{j1})^2 + (Y_{k2} - Y_{j2})^2 + \chi^2(p)}{\chi^2(p)} \xrightarrow{P} 1, \quad \text{as } p \rightarrow \infty. \quad (2.1)$$

This implies that the Euclidean distance function will be dominated by those variables that are pure noise in general. Although another distance function might be a better choice, they all suffer from the same problem but to a lesser degree.

Therefore, filtering is very important to quickly reduce the genes of interest and avoid the aforementioned problems introduced by those noisy variables which contain no information at all.

2.2 Example of preprocessing. Various methods of preprocessing have been proposed in the literature. These methods mainly deal with problems having class labels. For example, Golub et al. [3] proposed a univariate ranking criterion for each gene in a two-class situation. The criterion is can be defined as the ratio of the absolute mean difference of gene expression levels of the two classes with respect to the sum of standard deviations of the two classes for each gene considered. Higher ratio indicate higher ranking of the gene. Some other author used the ratio of between-class sum of squares to within-class sum of squares of each gene for multi-class problems. Later, Tibshirani et al. [14] used the shrunken centroids method for gene classification. Shrinkage and gene selection are integrated into a naive Bayes classifier. The univariate gene selection is based on t-statistics for each gene of each class.

Another method is the random forest procedure [1]. Random forest is an ensemble tree approach in data mining. It has been used as a popular approach in gene selection. Basically, random forest build tree classifiers. A tree classifier recursively partition the data to classify. The tree is built by greedily searching locally optimal split rules recursively. Instead of using all variables (genes) to build the trees, random forest uses a random sample of variables (genes) during the tree construction. The randomness causes different trees built even following the same procedure each time. In this sense, we can get many (usually more than 50) different trees using the same data set. These trees are used as an ensemble classifier via voting. Random forest can have high accuracy in classification. It also effectively estimates the performance of these randomly selected variables (genes) and provides an overall measure of variable (gene) importance. Different ways of evaluating the variables (genes) can be found in the random forest manual. These importance measures will consider interactions among the variables (genes) due to the nature of the tree classifiers.

As an example, let us look at the famous Leukemia data. This data set has 38 training samples and 34 test samples of two types of acute leukemias, acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) [3]. Each sample is related to 7129 genes.

Due to the large number of genes, we propose a combination of univariate ranking and random forest. First, univariate ranking is performed to select the best 200 genes. Then the random forest procedure is performed on these 200 genes to obtain their measures of importance. Below is a figure showing the importance measure of the 200 pre-selected genes.

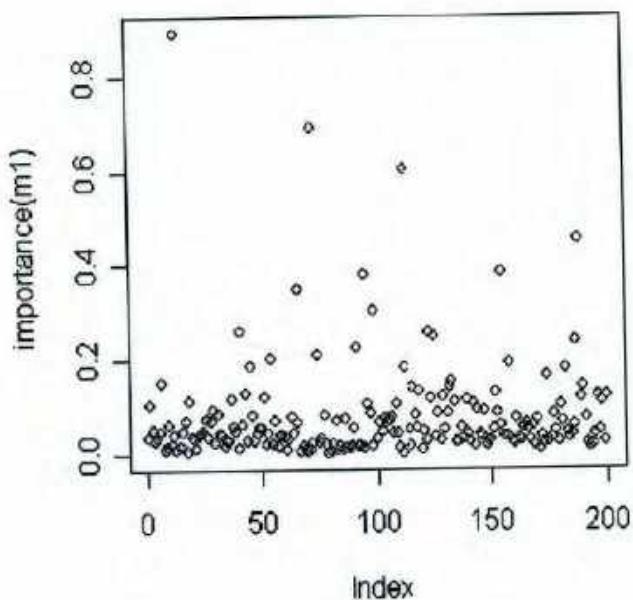


Figure 1 Variable (Gene) Importance using Random Forest for 200 Pre-selected Genes

We see that the variable importance using random forest is not the same as that using univariate ranking when interactions are considered. We can select a number of genes based on the variable importance for our needs.

2.3 Future avenues. As seen in the previous example, univariate ranking methods implicitly assume weak or no interactions or correlations among genes, which may not be true in practice of microarray analysis. The ideal situation would be to have a method which considers all variables together for selection. We now introduce the support vector machine classifier. The support vector machine basically searches for the optimal hyperplane that separates the data. Here an optimal hyperplane is the one that maximizes the geometric margin (the closest distance of the observations to the hyperplane). For a nonlinear pattern, the original space can be mapped to a high-dimensional space using kernel functions and the so-called Reproducing Kernel Hilbert Space (RKHS). So support vector machines can learn very well for a general problem.

In particular, for gene expression data, Guyon et al. [4] proposed a recursive feature eliminating method for gene selection. First, all the variables (genes) are used in model fitting. Then a large number of variables (genes) which are not significant in the model are removed. The model is then rebuilt on the rest of the variables (genes), and we can recursively repeat the procedure until only a small number of genes is left in the model. Zhu and Hastie used penalized logistic regression with similar ideas of gene selection [16].

For example, in the Golub 1999 data set just described in the previous section, Zhu and Hastie report 26 genes selected with two cross-validated errors on the training set and one error on the test set. The support vector machine selects 31 genes with similar performances.

In the future, we could try to combine random forest with support vector machine to perform gene selection.

3 Nonlinear dimension reduction

While filtering the data requires the methodologies to be able to identify in some sense redundant information in the already existing data, the ultimate purpose of dimensionality reduction is to reduce the data to a low-dimensional manifold in such a way that unsupervised learning is possible on new data. In other words, we have to discover the main trend in the existing data in order to be able to deal with newly acquired data in a similar way. Along with the (probabilistic) density estimation techniques, dimensionality reduction is a key methodology in developing algorithms for unsupervised learning [10]. Although these two methodologies can be applied simultaneously, nonlinear reduction can also be developed in a non-probabilistic framework.

Historically, first developed techniques for dimensionality reduction were based on linear versions of principal component analysis (PCA) (described below) and other eigenvalue-based methods such as variants of centre manifold reduction techniques. Due to difficulties with the mapping of the higher-dimensional data (even if represented by clusters) into a single coordinate system of lower dimensionality in applications of such techniques, most recent developments were centered around nonlinear dimensionality reduction methods. These techniques, similarly to traditional linear methods, are essentially based on *spectral embeddings*, but with a key new feature now of being able to generate nonlinear embeddings.

3.1 Spectral embedding methodologies.

3.1.1 *Generic setting for the spectral embedding.* Nonlinear procedures of interest can be cast in the following generic setting:

1. Given the input space, compute neighbourhoods;
2. Construct the cost function to determine the weight matrix as a result of an optimisation procedure (e.g., minimising the generalized error function);
3. Based on the eigenvectors of the above matrix (can be shown with the Rayleigh-Ritz ansatz), calculate the spectral embedding.

3.1.2 *The nearest neighbour parameter as a key to success.* The starting point of spectral embedding methodologies is computing neighbourhoods. In all the algorithms currently available to the National Research Council (NRC) of Canada, our industrial partner, we have one of the two situations:

- either the number of nearest neighbours is predefined by a certain value, denoted further by K (as it is the case, for example, in the Locally Linear Embedding (LLE)),
- or the input information requires the neighbourhood radius, denoted further by ϵ (as is the case, for example, in Isomap).

This may bring difficulties in some of the practical situations where a nonlinear dimensionality reduction algorithm is applied to a set of new data. Indeed, if we choose an estimate for K in the LLE such that K is very small compared to the real situation, a neighbourhood can falsely divide the underlying manifold. On the other hand, if we choose an estimate for K in the LLE such that K is large, the resulting manifold will be excessively smoothed and important small-scale features will be completely missing. Similar difficulties arise in the choice of ϵ .

3.1.3 *Description of some algorithms.* **Locally Linear Embedding (LLE):** The Local Linear Embedding (or LLE) Algorithm is a spectral embedding method for nonlinear dimension reduction developed by Roweis and Saul [8]. This algorithm takes as input X , a $p \times n$ matrix (whose columns contain the n data points in \mathbf{R}^p) and outputs Y , a $d \times n$

matrix, where $d < p$ is the dimensionality of the embedding of the input. The idea behind this algorithm is to characterize the local geometry of patches by linear coefficients that reconstruct each data point from its nearest neighbours when determining the underlying lower-dimensional manifold.

There are three major steps in the LLE algorithm. The first is to determine the neighbours in X -space. The second step is to solve for reconstruction weights W_{ij} which allow each point \mathbf{X}_i to be reconstructed from its neighbours \mathbf{X}_j . The final step is to compute the embedding coordinates Y using the reconstruction weights W . We now describe these three steps in more detail.

The first step is to determine the neighbours for each data point. This can be done in various ways described above. For our purposes, we compute the K -nearest neighbours for each data point.

The second step is to determine the reconstruction weights. To that end, we let W_{ij} denote the contribution of the j^{th} data point to the i^{th} reconstruction. Then, the weights W_{ij} are computed that minimise the following cost function:

$$E(W) = \sum_i \|\mathbf{X}_i - \sum_j W_{ij} \mathbf{X}_j\|^2. \quad (3.1)$$

which is known as the reconstruction error. The minimisation is performed subject to two constraints. The first constraint is that \mathbf{X}_i is reconstructed only from its K nearest neighbours. Thus, we set W_{ij} to 0 if \mathbf{X}_j is not a neighbour of \mathbf{X}_i . The second constraint is that each set of local weights must sum to 1. Determining the optimal weights is a least squares optimisation problem that is described in further detail in Appendix A of [9].

The final step is to compute the embedding coordinates Y using the weights W . This is done by choosing the Y_i that minimise:

$$\Phi(Y) = \sum_i \|\mathbf{Y}_i - \sum_j W_{ij} \mathbf{Y}_j\|^2. \quad (3.2)$$

This specifies a quadratic form in Y which can be minimised by solving a sparse $N \times N$ eigenvector problem. See Appendix B in [9] for more details.

Isomap: The Isomap algorithm consists of three primary steps:

1. Construct the neighbourhood graph G .

As with the LLE method, the first step of the Isomap method is to determine which points are neighbours based on Euclidean distances between points in the input data in \mathbb{R}^p . The neighbours are obtained using one of the two basic approaches for finding nearest neighbours outlined above. The information about the neighbourhoods is collected into a weighted neighbourhood graph G that has a node for each data point in the original input. The nodes of G are connected iff they are neighbours and the weights on the edges connecting nodes are the corresponding Euclidean distances between neighbouring data points in \mathbb{R}^p . The graph G is particularly easy to construct when the number of data points n is smaller than the dimension p of the space in which the data is embedded.

2. Construct the matrix D containing the shortest paths between all pairs of points in the graph G .

Isomap estimates the geodesic distances between all pairs of points on the manifold. This is achieved by computing the shortest path distances between vertices in the

weighted graph G constructed in the first step. Dijkstra's algorithm is known to be a good algorithm to find a shortest path in a weighted graph.

3. Apply Multi-Dimensional Scaling to the matrix D to determine the d -dimensional embedding.

The final step of Isomap applies classical Multi-Dimensional Scaling (MDS) to the matrix of D of graph distances as computed in the second step. The MDS algorithm constructs an embedding of the data in a d -dimensional Euclidean space that best preserves the intrinsic geometry of the manifold as determined by the relative distances of the points. The particular embedding found results from minimising a particular measure of error; the solution of this optimisation problem reduces to an eigenvalue problem.

Some more comments are in order concerning the final step of the Isomap algorithm. Given a set of n input vectors $\{\mathbf{X}_1, \dots, \mathbf{X}_n\} \subset \mathbf{R}^p$, the Isomap algorithm returns a set of n vectors $\{\mathbf{Y}_1, \dots, \mathbf{Y}_n\} \subset \mathbf{R}^d$ where $d < p$ is prescribed. Let

$$X = [\mathbf{X}_1, \dots, \mathbf{X}_n] \in \mathbf{R}^{p \times n} \text{ and } Y = [\mathbf{Y}_1, \dots, \mathbf{Y}_n] \in \mathbf{R}^{d \times n}$$

be rectangular matrices with the input and output column vectors stacked in sequence. The pre-images Y of the input data X are found as the minimisers of a cost function

$$E(Y) = \|\tau(D) - \tau(D_Y)\|_F, \quad (3.3)$$

where $\|A\|_F = \left[\sum_{i,j} |A_{i,j}|^2 \right]^{1/2}$ is the usual Frobenius matrix norm. Further, in the definition of the cost function in (3.3), D_Y denotes the matrix of Euclidean distances between all the columns of Y taken pairwise, i.e.,

$$(D_Y)_{i,j} = \|\mathbf{Y}_i - \mathbf{Y}_j\| \quad (i, j = 1, \dots, n).$$

The operator τ in (3.3) is defined as

$$\tau(A) := -\frac{1}{2}H(A \cdot A)H \quad (3.4a)$$

where $A \cdot A$ is the Hadamard (entrywise) product of A with itself and H is a centering matrix; explicitly,

$$(A \cdot A)_{i,j} := A_{i,j}^2 \quad (3.4b)$$

$$H_{i,j} := \delta_{i,j} - \frac{1}{n} \quad (3.4c)$$

where n is the number of data points and $\delta_{i,j}$ is the usual Kronecker delta. The operator τ expresses the (Frobenius) distance between matrices using matrix products and thus makes the minimisation of $E(Y)$ easier.

3.1.4 Optimal number of nearest neighbours. In what follows, we focus on the problem of optimal choice of nearest neighbour numbers.

As proposed by Kouropeteva et al. (with modifications recently suggested by Samko et al.) [6], we choose a set of values of K from $[K_{min}, K_{max}]$. The simplest choice of K_{min} in our case is 1.

Next, for each $K \in [K_{min}, K_{max}]$, we calculate the cost function:

$$E = \|\tau(\bar{D}) - \tau(D_Y)\|, \quad (3.5)$$

where

$$D_Y = \{d_Y(i, j) = \|\mathbf{Y}_i - \mathbf{Y}_j\| = \sqrt{\sum_{k=1}^d (\mathbf{Y}_i^k - \mathbf{Y}_j^k)^2}\} \quad (3.6)$$

is the matrix of Euclidean distances in the output space, while \bar{D} is different for different spectral embedding algorithms. We focus on a generalization of LLE where we have

$$\bar{D} \equiv D_X = \{d_X(i, j) = \|\mathbf{X}_i - \mathbf{X}_j\| = \sqrt{\sum_{k=1}^p (\mathbf{X}_i^k - \mathbf{X}_j^k)^2}\} \quad (3.7)$$

As usual, the operator τ converts distances to inner products (to simplify the optimisation procedure).

Then all K values where minima of $E(K)$ are achieved will form the set S_K of initial candidates for the optimum value of K .

Finally, the nonlinear reduction algorithm should be run for each $K \in S_K$. And K_{opt} is determined by the following formula based on minimising the residual variance:

$$K_{\text{opt}} = \arg \min_K (1 - \rho_{D_X D_Y}^2), \quad (3.8)$$

where ρ is the linear correlation coefficient taken over all entries of the matrices D_X and D_Y which contain Euclidean distances between pairs of points in the input (dimension p) and output (dimension d) spaces, respectively.

3.1.5 Choice of the cost function and how to avoid ill-posedness. The above choice of the cost function in the form of (3.5) is not the only possible one. In the generalized LLE we worked on, the cost function is taken as a measure of the reconstruction error:

$$E(W) = \sum_i \left\| \mathbf{X}_i - \sum_j W_{ij} \mathbf{X}_j \right\|^2. \quad (3.9)$$

To avoid ill-posedness it is essential to add constraints to this optimisation problem. The most natural are related to the weights, and probably the simplest one is

$$\sum_j W_{ij} = 1. \quad (3.10)$$

Further, as we observed in our experiments (described below), it might be essential to precondition the Gram matrix by a regularization procedure.

3.1.6 Further improvements. The search for the neighbours can be improved further if it is carried out with respect to the geodesic distance, rather than the Euclidean distance as it is commonly done. Only a slight modification of the LLE algorithm is required in this case [15]. In order to eliminate the necessity to estimate geodesic distances between faraway inputs on the manifold, and hence to improve the efficiency, we can apply a semidefinite embedding as recently proposed by Weinberger et al.

Finally, further modifications of the spectral embedding algorithms, described here, can be introduced with the stochastic neighbour embedding which could be useful for relatively noisy data.

Potential gaps in existing literature: There is a lot of work and experience to draw upon from the machine learning community to help with the problem of dimensionality reduction for microarray data. However, there are some features of algorithms for nonlinear dimensionality reduction described in the existing literature that complicate the present

study rather than making matters more clear. We describe some of these issues with the hope that they can be resolved in later studies.

A number of nonlinear dimensionality reduction algorithms are based on the assumption that there exists a nonlinear manifold embedded in \mathbf{R}^p that underlies the given set of data. The input of the algorithms usually consists of a set of n data points in \mathbf{R}^p and possibly the dimension d of the manifold sought. The output of the algorithms typically consists of a set of n vectors in \mathbf{R}^d with $d < p$ that would be the pre-images of the input data vectors in a presumably lower-dimensional space. However, manifolds consist of uncountably many coordinate charts (smooth mappings from \mathbf{R}^d into \mathbf{R}^p whose ranges are contained in the points on the manifold) in an atlas that cover the whole manifold. It is not typical for a single coordinate chart to cover the whole manifold. Moreover, when the output from a dimensionality reduction algorithm consists of the pre-images of the data under a particular coordinate chart, it is not obvious which coordinate chart is being used. Some charts have greater utility than others in different regions of the manifold.

To clarify the preceding discussion, assume that the manifold from which the input data is sampled is the unit sphere in \mathbf{R}^3 (i.e., $p = 3$ and $d = 2$). Consider the mappings $\psi_1 : (0, \pi) \times (\pi, \pi) \rightarrow \mathbf{R}^3$ and $\psi_2 : \{(x, y) \in \mathbf{R}^2 : x^2 + y^2 < 1\} \rightarrow \mathbf{R}^3$ that are defined by

$$\begin{aligned}\psi_1(\phi, \theta) &:= (\cos \theta \sin \phi, \sin \theta \sin \phi, \cos \phi) & (\phi, \theta) &\in (0, \pi) \times (\pi, \pi) \\ \psi_2(x, y) &:= (x, y, \sqrt{1 - x^2 - y^2}) & (x, y) &\in \{(x, y) \in \mathbf{R}^2 : x^2 + y^2 < 1\}.\end{aligned}$$

Both of these charts cover portions of the unit sphere in \mathbf{R}^3 . However, while ψ_1 covers the region near the point $(0, 0, 1)$ relatively poorly due to a coordinate singularity in ψ_1 near $\phi = 0$, the chart ψ_2 can be used near that region without difficulty. Similarly, the mapping ψ_2 encounters difficulty near the boundary of the unit disk in \mathbf{R}^2 for exactly the same reason whereas ψ_1 has quite well-behaved derivatives near the region where $\phi = \pi/2$. As such, the output of algorithms such as LLE or Isomap consist of vectors in \mathbf{R}^d , but it is in no way obvious which chart has been selected and whether it is one that is appropriate in the region of the manifold being sampled.

Another problem shared by many algorithms is the number of heuristic parameters inherent even in deterministic algorithms. For instance, in LLE, the dimension d of the lower-dimensional manifold on which the sampled data lies is an input parameter of the algorithm. (Admittedly, the Isomap algorithm does not share this particular shortcoming in that it starts from $d = 1$ and increments d until a suitable value of d is determined.) Other heuristic choices in the development of the algorithms include the number of nearest neighbours to choose, the method by which nearest neighbours are measured, and the choice of metric in the objective function to minimise in finding the reconstruction weights.

The most perplexing difficulty arises when trying to compare the performance of distinct algorithms. If the output of algorithm A is a set of pre-images of the data under one coordinate chart and the output of algorithm B is a similar set of pre-images, does it follow that the outputs can be compared? This is a vexing issue for assessing the numerical accuracy and the asymptotic complexity of dimensionality reduction algorithms. Convergence properties of, say, numerical approximation schemes for partial differential equations, can be estimated by numerical experiments where the exact solution is known even when convergence proofs are unattainable. Such numerical experiments are invaluable when new schemes are suggested for comparison to existing frameworks. It does not seem that the literature on nonlinear dimension reduction algorithms has analogous criteria for comparison of algorithms.

3.2 Kernel Principle Component Analysis (KPCA).

3.2.1 *Description of the PCA method.* Principal component analysis (PCA) is one of the statistical methods to extract the patterns in the data and to represent the original data in another way based on their similarity and dissimilarity. PCA is not only widely-used for pattern extraction but also for dimensionality reduction and data visualization. Once the patterns hidden in the data are identified, we can project the data into lower dimension by selecting several most important patterns and without losing too much information. As one might expect, it is nontrivial to identify these patterns in the high-dimensional data.

PCA is essentially a basis transformation. Suppose the data points are given by $\{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n\}$, where $\mathbf{X}_i \in R^p$ and are centered, i.e., such that $\sum_{i=1}^n \mathbf{X}_i = 0$. The covariance matrix is then defined by

$$C = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^T.$$

Since the orthonormal eigenvectors of the covariance matrix form a basis in space, we can express the data in terms of the eigenvector basis, instead of Cartesian coordinates. Actually, the eigenvectors show the directions of variance in the data. In addition, the corresponding eigenvalues indicate the proportions of the variances. The eigenvector corresponding to the largest eigenvalue is called the principal component.

It will be easier to analyse the data if their dimension is much smaller. Therefore, we can project the data onto a lower-dimensional space. Then the data are approximated by the linear combination of the selected d eigenvectors ($d < p$) and truncating the tails of the vectors creating vectors of length d . The value of d may be decided by the specific need, such as a value of two or three for visualization, or by minimising the difference between the original data and its approximation.

3.2.2 *Description of the Kernel PCA method.* The principle component analysis has a very long history and is known to be very powerful for the linear case. However, the sample space that many research problems are facing, especially the sample space of microarray data, are considered nonlinear in nature. One reason might be that the interaction of the genes are not completely understood. Many biological pathways are still beyond human comprehension. It is then quite naive to assume that the genes should be connected in a linear fashion.

To handle nonlinear spaces, a natural idea is to make a suitable transformation that tries to make the transformed space linear. Although this idea has been mentioned in the literature many times, the breakthrough did not come until the last 20 years during which time the computational issue has been resolved.

In order to capture nonlinear patterns, it is often useful to consider a nonlinear transformation of the original variables. For example, given two random variables, we might consider only the linear combination, i.e., $a_1x_1 + a_2x_2$. To capture any nonlinear relationship, we might want to consider the ensemble of

$$\mathcal{E} = \{X_1, X_2, X_1^2, X_2^2, X_1X_2, X_1^3, X_2^3, X_1X_2^2, X_1^2, X_2, \dots\}.$$

Although this can be done, the computational burden associated with the expansion into a higher-dimensional space is very costly. Given the fact the microarray data already has a very high dimensionality to begin with, this does not appear to be feasible.

To be more specific, we consider a mapping:

$$\mathcal{K} : \mathcal{X} \longrightarrow \mathcal{F} \tag{3.11}$$

where \mathcal{X} and \mathcal{F} are the sample and output spaces, respectively, and K is the kernel function.

However, the Kernel PCA method does exactly this seemingly impossible task. The key element of the Kernel PCA is that the original space is transformed into an output space through kernel functions. Kernel functions are designed to capture the nonlinear nature of the original space by expanding the basis functions into a much higher-dimensional space. However, any computation after the transformation can be done using the kernel function and the inner product of the original space. In other words, no actual transformation is necessary, and the results are obtained without significant computational cost.

Given $f \in \mathcal{F}$ and $g \in \mathcal{F}$, we then have

$$\langle f, g \rangle = \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j K(x_i, x_j). \quad (3.12)$$

Detailed discussions of the Kernel PCA can be found in [12].

3.2.3 Choice of kernel functions and future research. There are many kernel functions that have been proposed in the literature. Gaussian functions are commonly used. However, this is no guarantee that a Gaussian function would be applicable all the time. One possible approach is to use the idea of model averaging. To be more specific, one could use an array of kernel functions and evaluate each kernel function for its effectiveness using some loss function, for example, the mean squared error (MSE).

4 Numerical experiments and results

4.1 Available datasets. We now describe the two types of datasets of interest to us: a microarray dataset from the NRC and some synthetic datasets which we have designed.

4.1.1 Microarray dataset. The NRC provided us with an AML microarray dataset of genetic data which sampled 7129 genes in 72 patients; this corresponds to 72 vectors of data in \mathbf{R}^{7129} . The patterns in the microarray data are nonlinear and are thus quite complicated. In addition, the data are noisy due to the nature of the microarray experiment.

4.1.2 Synthetic datasets. The papers of Roweis and Tannenbaum make extensive use of test data sets for their algorithms. These include an S-shaped ribbon (a two-dimensional manifold embedded in \mathbf{R}^3 , the standard unit sphere $S^2 \subset \mathbf{R}^3$, and a number of variants involving translations of a fixed image. These examples support the strength of these algorithms in the event that the number n of samples available is larger than the dimension p of the space from which the data are sampled. Unfortunately, this is not the case with cell microarray data.

We advocate generating synthetic test data known to lie on a manifold with known structure of arbitrary dimensions as a means of testing prospective algorithms. The procedure mentioned here was developed using random sampling on the unit sphere $S^d \subset \mathbf{R}^{d+1}$. This does in fact require some care; randomly sampling points on the unit hypersphere needs to be done in a way to ensure that samples are not clustered near poles. Fortunately, a very simple framework for doing so is provided by Knuth.

1. Generate random vectors $\mathbf{Y}_1, \dots, \mathbf{Y}_n \in \mathbf{R}^{d+1}$, each component being observations of a random variable with Gaussian distribution with mean at 0.
2. $\mathbf{Y}_k \mapsto \mathbf{Y}_k / \|\mathbf{Y}_k\|_2$ ($k = 1, \dots, n$) generates set of n random vectors uniformly-distributed on the unit sphere in \mathbf{R}^{d+1} .
3. Embed vectors \mathbf{Y}_k into column vectors $\mathbf{X}_k \in \mathbf{R}^p$ by padding with zeros.

The vectors generated thusly lie on the unit sphere $S^d \subset \mathbf{R}^p$. This procedure can be adapted to make data points from the manifolds $S^{d_1} \times S^{d_2}$ embedded in \mathbf{R}^p or any similar Cartesian product manifold. To obscure the obvious manifold structure of a set of vectors in \mathbf{R}^p with zeros in most of the components, a number of strategies can be used.

1. Make the substitutions $\mathbf{X}_k \mapsto P\mathbf{X}_k$ where P is a random $p \times p$ permutation matrix.
2. Make the substitutions $\mathbf{X}_k \mapsto Q\mathbf{X}_k$ where $Q = I - 2\mathbf{u}\mathbf{u}^T$ is a random orthogonal Householder reflection ($\|\mathbf{u}\|_2 = 1$).
3. Make the substitutions $\mathbf{X}_k \mapsto \mathbf{X}_k + \mathbf{a}$ where $\mathbf{a} \in \mathbf{R}^p$ is a random translation.
4. Add Gaussian noise to all of the components of each vector.

4.2 Numerical experiments. For our numerical experiments, we tested the LLE algorithm on the NRC microarray dataset described above. In particular, we performed a leave-one-out cross-validation experiment and measured the reconstruction error for several combinations of d and K .

Leave-one-out is a cross-validation technique in which the data is divided into n subsets each corresponding to one data point. Training on the data is performed h times, each time using only the omitted subset to compute the error criterion of interest. The following pseudo-code demonstrates the use of the leave-one-out technique for the LLE algorithm:

```

for i = 1:n
    1. Remove X_i from X, i.e., set Xhat_i = X\{X_i}.
    2. Compute the manifold on Xhat_i.
    3. Project X_i onto Xhat_i.
    4. Compute the reconstruction error for X_i.
end

```

Then, the reconstruction error is the sum of the reconstruction errors for each of the n data points.

We repeated this experiment on the filtered dataset (using the result from the Random Forest algorithm described above). Before discussing our numerical results, we describe three main metrics for analyzing the error in the nonlinear dimension reduction process.

4.3 Measures for error analysis. There are three main measures for the error in the nonlinear dimension reduction algorithms: the distortion of the distances, the residual variance, and the reconstruction error.

To measure the distortion of the distances, we compute $E(W) = \sum_{i < j} W_{ij}(D_{ij} - \Delta_{ij})^2$, where D_{ij} is the distance between the points in \mathbf{R}^p , and Δ_{ij} is the corresponding distance in \mathbf{R}^d . Although an interesting error metric, we have not studied this metric in favor of pursuing others.

The measures for characterizing the quality of the nonlinear dimension reduction procedure described here are based on the following two choices:

- Minimisation of the generalized reconstruction error:

$$E_Y = \frac{1}{n} \sum_i \|\mathbf{X}_i - P\mathbf{X}_i\|^2, \quad (4.1)$$

where P is the projection operator, $P^2 = P$ such that $\mathbf{Y}_i = P\mathbf{X}_i$. Note that we go from a space of dimensionality p to a space of dimensionality d where for the linear pieces we have

$$\text{var}(\mathbf{Y}) = \text{Tr}(P\mathbf{C}\mathbf{C}^T), \quad P = \sum_{\alpha=1}^d \mathbf{e}_\alpha \mathbf{e}_\alpha^T, \quad (4.2)$$

$$\mathbf{C} = \frac{1}{n} \sum_i \mathbf{X}_i \mathbf{X}_i^T = \sum_{\alpha=1}^p \lambda_\alpha \mathbf{e}_\alpha \mathbf{e}_\alpha^T, \quad (4.3)$$

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p. \quad (4.4)$$

- Minimisation of the residual variance:

$$1 - \rho_{D_X D_Y}^2 \quad (4.5)$$

is as discussed above (see (3.8)).

Determining the error of the reduction with (4.5) was proposed for the original LLE algorithm [8]. Note, however, that since all the available algorithms require computing spectral characteristics of the underlying data in one form or another, the computational dichotomy of spectra may represent a non-trivial problem in practice [7]. Nevertheless, recent applications of spectral embedding non-linear reduction techniques, such as LLE and Isomap, to high-density microarray data sets have demonstrated their robustness [5, 2]. Finally, note that in the linear case, the approach based on (4.5) is the standard maximisation of variance subspace:

$$\text{var}(\mathbf{Y}) = \frac{1}{n} \sum_i \|P\mathbf{X}_i\|^2, \quad (4.6)$$

and is equivalent to the procedure (4.1) based on the minimum reconstruction error.

4.4 Numerical results. Now that we have described three possible error metrics, we return to a description of the results from obtained from the LLE algorithm by running the leave-one-out cross-validation experiment on the NRC microarray dataset. We will also describe the results from running leave-one-out on the corresponding filtered dataset.

The following two figures show the results from the leave-one-out cross-validation experiment using LLE on the microarray and filtered microarray datasets. We see from both figures that the best results are obtained for roughly $K = 12$ nearest neighbours. This result is independent of the choice of d . When a greater number of nearest neighbours is used, the LLE algorithm is more expensive, and there is little to no additional benefit, i.e., the reconstruction error decreases little. The figures also demonstrate that the reconstruction error is minimised for low-dimensional manifolds of smaller dimension. This result is also independent of whether or not filtering was applied. As was expected, the amount of reconstruction error decreased when filtering was applied to the microarray dataset before the leave-one-out cross validation experiment was performed; this demonstrates the success of the filtering process.

5 Discussion and recommendations

As discussed above, support vector machines may be useful for gene selection when combined with the random forest algorithm. There are other issues to explore within the filtering context such as filtering time trends.

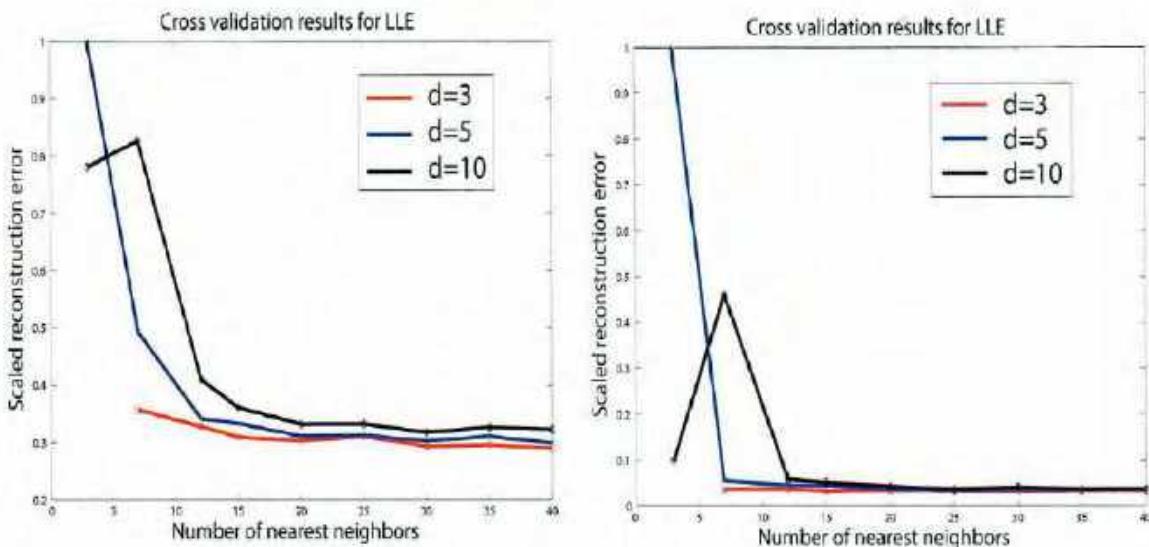


Figure 2 Leave-one-out cross validation results for the LLE algorithm on the NRC microarray (left) and filtered microarray (right) datasets. These figures demonstrate that 12 is a good choice for the number of nearest neighbours in this algorithm. In addition, the reconstruction error is smaller for lower-dimensional manifolds and for the filtered microarray dataset.

Within Kernel PCA, the choice of kernel function should be investigated to find the most useful type of kernel for the microarray and synthetic datasets.

There are several avenues to pursue within the spectral embedding framework for nonlinear dimensionality reduction. First, we would like to experiment with the Isomap algorithm and compare the results of that algorithm with the LLE experiments. Comparisons should also be made with the other spectral embedding algorithms. Thus far, our results indicate that the Random Forest and LLE algorithms were useful for filtering the genes and nonlinear dimensional reduction as tested on the NRC microarray dataset.

A second avenue to explore is the choice of the nearest neighbours in LLE and other spectral-embedding algorithms. There are many options for choosing the neighbours within LLE such as the using the K -nearest neighbours, the points within a ball of a specified radius, or using an adaptive local distance metric to choose the neighbours flexibly within various regions. It is expected that an adaptive choice for the neighbours will produce improved results. Above it was discussed how to pursue an optimal number of neighbours.

A third avenue that needs to be investigated is the choice of cost function. Our experiments measured the reconstruction error because this metric was of interest to the NRC, our industrial partner. It is not clear which error metric would be the most useful for the general case, as we have not run any experiments using the distortion of the distances or the residual variance as our error metric. We have not run any experiments on the synthetic datasets which we have designed; tests will need to be run on this dataset for us to be able to understand how these algorithms perform on additional datasets.

The final issue we have identified for investigation is the choice of cross-validation technique. Our experiments used the leave-one-out cross-validation technique. This can be

generalized to the leave- v -out technique, which is a more complicated version of leave-one-out in which all possible subsets of v data points are left out of the training set. As the choice of cost function changes, the most successful cross-validation technique may change as well.

Numerous experiments need to be performed on the microarray and synthetic datasets in order for us to better understand the performance of the algorithms described in this report on filtering of genes and the nonlinear dimensionality reduction problem.

References

- [1] L. Breiman. (2001), Random forests. *Machine Learning* 45:5–32.
- [2] K. Dawson, R. L. Rodriguez, and W. Malyj (2005), Sample phenotype clusters in high-density oligonucleotide microarray data sets are revealed using Isomap, a nonlinear algorithm. *BMC Bioinformatics* 6(185):1–17.
- [3] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, and M. Caligiuri (1999), Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* 286:531536–531536.
- [4] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik (2002), Gene selection for cancer classification using support vector machines. *Machine Learning* 46:389–422.
- [5] B. W. Higgs, J. Weller, and J. L. Solka (2006), Spectral embedding finds meaningful (relevant) structure in image and microarray data. *BMC Bioinformatics* 7(74):1–13.
- [6] O. Kouropeteva, O. Okun, and M. Pietikainen (2002), Selection of the optimal parameter value for the linear embedding algorithm. In *Proc. of the 1st International Conference on Fuzzy Systems and Knowledge Discovery* pages 359–363.
- [7] R. V. N. Melnik (2000), Topological analysis of eigenvalues in engineering computations. *Engineering Computations* 17(4):386–416.
- [8] S. T. Roweis and L. K. Saul (2000), Nonlinear dimensionality reduction by locally linear embedding. *Science* 290:2323–2326.
- [9] L. Saul and S. Roweis (2001), An introduction to locally linear embedding. Paper located at <http://www.cs.toronto.edu/~roweis/lle/publications.html>
- [10] L. K. Saul and S. T. Roweis (2003), Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research* 4:119–155.
- [11] M. Schena, D. Shalon, R. Heller, A. Chai, P. O. Brown, and R. W. Davis 1996), Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. In *Proc. Nat. Acad. Sci.* Volume 93, pages 10615–10619. National Academy of Sciences.
- [12] J. Shawe-Taylor and N. Cristianini (2006), *Kernel Methods for Pattern Analysis*. Cambridge University Press.
- [13] M. Shena, D. Shalon, R. W. Davis, and P. O. Brown (1995), Quantitative monitoring of gene expression patterns with complimentary DNA microarray. *Science* 270:467–470.
- [14] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu (2002), Diagnosis of multiple cancer types by shrunken centroids of gene expression. In *Proc. Nat. Acad. Sci.* volume 99, pages 6567–6572. National Academy of Sciences.
- [15] C. Varini, A. Degenhardt, and T. W. Nattkemper (2006), ISOLLE: LLE with geodesic distance. *Neurocomputing* 69:1768–1771.
- [16] J. Zhu and T. Hastie (2004), Classification of gene microarrays by penalized logistic regression. *Information Processing Systems* 14:427–443.

CSIRO Mathematical and Information Sciences

Modelling, Simulation, and Control of Polymer and Composite Systems.

Part 1: Distance Geometry Algorithms in Molecular Modelling

By R. V.N. Melnik and A. Uhlherr

May 5, 2000

Report Number: 2000/79

CSIRO Mathematical and Information Sciences,
Building E6B, Macquarie University Campus,
North Ryde NSW 2113
E-mail: Roderick.Melnik@cmis.csiro.au



Modelling, Simulation, and Control of Polymer and Composite Systems.

Part 1: Distance Geometry Algorithms in Molecular Modelling

R. V. N. Melnik* and A. Uhlherr†

Abstract

In this report microscopic, atomistically detailed models of polymer and composite systems are studied in the context of distance geometry algorithms. The development of such algorithms in the general framework of molecular modelling is closely associated with problems in computational optimisation, and is pursued here as a key to a better understanding of properties of polymer and composite systems at the macroscopic level. Applications of two distance geometry algorithms to topological optimisation of alkane chains and bulk decane structures are discussed with results of numerical simulations.

Key words: distance geometry algorithms, polymer and composite systems, computational optimisation.

AMS Subject Classification: 82D60, 65C20, 73S10

1 Introduction

Development in polymers and composites is at the heart of many new industrial materials technologies. Recent advances in this field include applications of environmentally stable high temperature thermoplastics, new thermoset materials, photoresists for fabrication of VLSI, flexible electronic conductors, highly polarised materials such as ferroelectrics, piezoelectrics, nonlinear optical materials. In such areas as aerospace and automotive industries in particular, further progress in this field is often impeded by inadequate understanding of how atomistically detailed molecular structures lead to macroscopic properties. Because the polymers and composites (in fact, many commercial polymers are composites, e.g. thermosetting resins containing fillers) applied in these areas of industry are generally disordered and heterogeneous (at least at the microscopic level), the current experimental methods can generally provide only hazy guidelines concerning local structures. Various properties of these materials are interrelated at the molecular level, and changing a single *variable*, say, the extent of crosslinking, the plasticizer content, or the polymer molecular weight, will have

*CSIRO Mathematical and Information Sciences, Macquarie University Campus, North Ryde, NSW 2113, Australia, E-mail: Roderick.Melnik@cmis.csiro.au

†CSIRO Molecular Science, Bag 10, Clayton South, VIC 3169, Australia

mutual effects on the use response [34]. In many cases it is important to know for any given material property the effect caused by the whole set of fabrication-controlled material variables or molecular characteristics because the advantages that composite or polymer materials have to offer have to be balanced against their undesirable properties (difficult fabrication techniques, complex rheological behaviour, etc). Therefore, modelling polymer and composite structures at the molecular level constitutes a natural approach to the solution of complex industrial engineering problems in the field where applications of mathematical and computational tools are both beneficial for industry and mathematically challenging. In this work we deal with such tools applied to molecular modelling, i.e. to the generation, manipulation, and representation of realistic 3D molecular structures with the purpose of understanding the physicochemical properties and macroscopic phenomena at the molecular level.

Molecular modelling has become a powerful approach in science and engineering in general, and in computational (bio)chemistry in particular. It is widely used for searching conformational space in order to find stable structures, for examining the effect of thermal motions on the structure of a molecule, for the analysis of the dynamics to obtain macroscopic materials properties as a function of temperature and pressure, etc. Since any molecule can be in an infinite number of spatial states, i.e. configurations, in practical applications we have to limit ourselves by subsets of such configurations, known as conformations, which share common (physico)chemical properties [23].

In principle, using methods of quantum mechanics we can compute first principles force fields and then, using the latter, we can apply molecular dynamics techniques to calculate trajectories which will allow to extract the information on the macroscopic properties of the material. However, if we take into account that to describe even partially crystalline polymers one needs $\sim 10^6$ atoms per unit cell, it becomes clear that the above classical *ab initio* procedure (and other methods of quantum chemistry such as pseudospectral or cell multipole methods) can be hardly applied to realistic polymers and composites. Typically, we have not only to average electronic degrees of freedom into force fields and charges, but also to impose "hard" constraints on bonds and angles to eliminate those degrees of freedom which are only lightly excited and hence have negligible effect on conformational (i.e. long-scale) changes in polymer or composite molecules. In particular, it is customary to assume that all bonds and angles are constrained and only the torsional degrees of freedom in a chain are relevant in determining its overall conformations [22]. Due to such simplifications, molecular modelling methods cannot be perfectly accurate, but they can (a) reasonably indicate whether it might be worthwhile to synthesise the material, (b) suggest alternative structures to consider [17]. Hence, with the spotlight fixed on these strengths of molecular modelling methods together with essential computational cost savings compared to the quantum mechanics methods, we conclude that the molecular modelling approach is a natural choice in simulating complex polymer and composite structures.

To get started with most of the molecular modelling procedures, including Molecular Dynamics (MD) and Monte-Carlo (MC), one has to determine an initial topological structure of the material under consideration. For complex polymer and composite systems this task is far from trivial and, what's more, the solution to this task keeps a key to success in the whole molecular modelling exercise. Indeed, since polymers and composites are characterised by a wide-range hierarchy of different length and time scales, in molecular simulations of complex materials such as long-chain polymers the large-scale conformational characteristics of the

system are given too little time to evolve. As a result, if the initial configuration is poorly chosen, the system often remains trapped within its neighbourhood which leads to insufficient samplings of the configuration space and unreliable estimations of the dynamic and structural properties of the material. Therefore, a major challenge in the field of molecular modelling of polymer and composite systems is to build realistic and computationally inexpensive initial configurations which can then be relaxed effectively by the available molecular modelling methodologies [28].

The rest of the report is organised as follows.

- In Section 2 we describe a general framework for modelling and simulation of polymers and composites and the place of distance geometry in this framework.
- In Section 3 we provide the reader with the mathematical foundations of the distance geometry methodology considered here as a starting point in molecular modelling.
- In Section 4 we focus on principal criteria for the design of distance geometry algorithms and discuss the effect of topological constraints on macroscopic properties of material structures. Other methods for finding starting geometries are also discussed in this section.
- In Section 5 the general problem of distance geometry is split into a series of problems in computational optimisation.
- In Section 6 we apply potential-function-smoothing procedures to simulation of short-chain polyethylene isolated in space and discuss options for improvements of such procedures.
- Section 7 is devoted to the description of a distance geometry algorithm capable of dealing with bulk material structures. Results of numerical simulations of long alkane chains and bulk decane structures are also presented in this section.
- Conclusions and future directions are discussed in Section 8.

2 Distance geometry methodology in a hierarchy of models for simulation of polymers and composites

Given (a) nuclear coordinates, (b) the number of electrons the system contains, and (c) a set of atomic orbital forms assigned to each atom, the most rigorous *ab initio* methodologies to atomistic simulation of structural, mechanical, electrical and optical properties of materials are those that are based on first principles quantum mechanics methods in which the electronic states and structures are calculated directly from the Schrödinger equation, which governs the motions of electrons and nuclei in atoms and molecules [18]. Such methodologies, known for almost a century¹, remain very restricted in size and time scales and are usually computationally impractical in modelling realistic polymer and composite systems. Typically, quantum mechanical calculations can provide information on potential surfaces for small molecules, but moving towards more complex polymer and composite systems, characterised by a wide-range hierarchy of different length and time scales, empirical energy functions of the molecular mechanics type become the only practical source of such information [32, 25]. In principle, however, using coarse graining and averaging procedures we can devise a hierarchy of models for simulation of polymers and composites, briefly summarised

¹In 1929 Paul Dirac noted that “The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble”.

below.

Model type	Methodologies	Space scales	Time scales
Quantum level	Quantum chemistry, Quantum Monte Carlo, Band theory	1 \AA - 10 \AA	$10^{-15} - 10^{-12}$ s
Atomic/Molecular Level	Classical dynamics, Statistical Mechanics	10 \AA - 100 \AA	$10^{-12} - 10^{-6}$ s
Phenomenological	Nanotechnological, Microstructural	100 \AA - 1 μm	$10^{-6} - 10^{-3}$ s
Continuum/Discrete	Microstructural, Semi-macrostructural, Macrostructural	1 $\mu\text{m} \rightarrow \text{cms}$	10^{-3} s \rightarrow mins
Physical/Engineering	Engineering design	Expanding limits	Expanding limits

Some entries in the above table are overlapping and could be attributed to the same group under other classification schemes. For example, in the context of this report it's worthwhile noting that models and methodologies related to quantum chemistry and molecular dynamics are referred generically to as atomistic simulation. Our interest in atomistic simulation of polymer and composite systems is limited to the assumption that the force field² has been derived empirically using the molecular mechanics approach [9]. In other words, by assuming that a model and a force field have been already chosen, we put ourselves in the framework of molecular modelling where the ultimate goal is to find the geometry with the minimum strain energy by rearrangement of the nuclei. Provided electronic effects are negligible, this approach is considered as a predictive tool leading to the possibility of evaluating macroscopic properties of the material from its microscopic structure.

Much progress has been recently achieved along this path with MC procedures applied to polymer and composite systems (in particular, with modifications of configuration-bias MC, concerted rotation and bridging moves methods [29]), where the focus has been placed on the algorithms for which the time of simulation depends *weakly* on the length of the chain. Although this feature gives a substantial advantage to MC procedure compared to the classical MD procedures, it would be proper to say that the former could be considered as a real alternative to the latter only for *static* thermodynamic properties. For the purpose of this work it is important to emphasise that most of the available molecular modelling procedures, including both MD and MC, exhibit a high sensitivity to parameterisation processes related to the choice of model and force field in a specific engineering problem, as well as the choice of the initial geometry which then is relaxed with such procedures.

Therefore, at the initial stage of molecular modelling we need an efficient method that requires neither force field parameters, nor a starting conformation, but at the same time is capable of predicting correctly the overall topological structure of the molecular system. To be viable for the analysis of complex polymer and composite systems, this method has to be computationally efficient and relatively simple. Distance geometry algorithms (DGAs) are

²i.e. the set of functions that defines an approximation to the strain energy together with the collection of terms that parametrises this approximation

the most obvious candidates for forming the basis of such a method, because (a) distances are coming naturally from chemical bonds, bond angles, and torsional angles, and (b) the complete conformational space of the system is contained within the set of minimum and maximum interatomic distances between all pairs of atoms in the molecule system. Since their first appearance in computational biochemistry in the late 1970s, DGAs have been applied not only in nuclear magnetic resonance structure determination and in conformational analysis of small molecules, where they performed as well as other sampling procedures, but also in such technologically important fields as constructing genetic maps [33]. These algorithms have recently attracted attention of researchers as a potentially powerful method for building approximate models of complex polymer and composite structures in conformational analysis [39]. By developing efficient methods for generating and minimising coordinates of each atom within the structure against the distance error function, we open a way to a substantial increase in the effectiveness of molecular dynamics codes and molecular modelling procedures that allow to extract physico-mechanical properties of technologically important polymers and composites [24, 18].

3 Mathematical foundations of molecular modelling: distance geometry as a starting point of simulations

The foundations of the distance geometry methodology in computational (bio)chemistry can be traced back to the Born-Oppenheimer approximation of the Schrödinger equation (for mathematical history of distance geometry the reader can consult [7, 19, 10]), where the nuclei are viewed as executing small oscillations about an equilibrium conformation which is a result of time-averaged or steady-state electronic configuration. If we assume that the (dynamical) behaviour of a molecular system is well predictable using classical mechanics, then this approximation allows us, at least in principle, to calculate time-dependent movement of each atom in a molecule by solving Newton's equations of motion for all degrees of freedom

$$(a) \frac{d^2x_i}{dt^2} = \frac{\mathbf{F}_i}{m_i}, \quad (b) \mathbf{F}_i = -\frac{\partial \mathcal{E}}{\partial x_i}, \quad i = 1, \dots, 3N. \quad (3.1)$$

given the energy surface \mathcal{E} and the derivative of energy in terms of nuclear coordinates x_i which determines the instantaneous force on any atom \mathbf{F}_i [17]. Assuming that from (3.1)(b) we can derive good approximations of forces between the atoms, we expect to have a good approximation of the geometry of the molecule by solving $3N$ equations (3.1)(a). Note that deriving such approximations is not a trivial task, because the force law in molecular dynamics requires accounting not only for the two-body force, but also for three-body force (by the bond angles) and four-body force (by the torsion angles). If the molecular system under consideration is Hamiltonian, we identify the total energy of the system with the Hamiltonian function

$$H(\mathbf{p}, \mathbf{q}, t) \equiv \mathcal{E} = K + P, \quad (3.2)$$

where kinetic energy, K , and the potential (strain) energy, P , are represented by

$$(a) K = \frac{1}{2} \langle M\mathbf{p}, \mathbf{p} \rangle, \quad (b) P(\mathbf{x}, t) = \sum_{\text{molecules}} (\mathcal{E}_b + \mathcal{E}_\theta + \mathcal{E}_\phi + \mathcal{E}_{nb}), \quad (3.3)$$

with $\mathbf{x} = (x_1, \dots, x_{3N})$ and $\mathbf{p} = (p_1, \dots, p_{3N})$ being the corresponding positions and momenta of the atoms, $M \in \mathbb{R}^{3N \times 3N}$ the diagonal mass matrix $\text{diag}(m_1, \dots, m_{3N})$, and $\langle \cdot, \cdot \rangle$ the Euclidean inner product in \mathbb{R}^{3N} , so that

$$K \equiv \frac{1}{2} \mathbf{p}^T M^{-1} \mathbf{p} = \sum_{i=1}^{3N} \frac{p_i^2}{2m_i}. \quad (3.4)$$

The summations with respect to each term in the potential energy function (3.3)(b) denote the total bond deformation (stretching) energy, the total valence angle deformation (bending) energy, the total torsional (dihedral) angle deformation energy, and the total nonbonded interaction (separated by at least 2 atoms, i.e. van der Waals) energy, respectively.

Molecular dynamics simulations are initiated by the kinetic energy of the atoms (3.4) chosen from a random velocity distribution, and we assume that

$$\mathbf{p}(t_0) = \mathbf{p}_0. \quad (3.5)$$

The resulting distortions from the equilibrium configuration are opposed by a force commensurate with the gradient of H . Since H can be only an approximation [30], obtained, for example, from the Consistent Valence Force Field Theory (see, for example, [35] and references therein), it is often convenient to split its “potential” part (3.3)(b) into the sum of weak and strong interactions [8]

$$P = P_{\text{weak}} + P_{\text{strong}}, \quad P_{\text{strong}} = \epsilon^{-2} U, \quad (3.6)$$

with appropriate scaling factor (or a singular perturbation parameter) ϵ and function U . In what follows, we assume only that all energy terms participating in the definition of H can be effectively expressed as a function of the coordinates of the atoms constituting the molecule system and a set of parameters computed from experimental data. Having obtained an approximation to the system Hamiltonian, we can reduce numerical integration of the equations of motions (3.1) to the solution of the following Hamilton’s equations:

$$\frac{\partial H}{\partial p_i} = \frac{dx_i}{dt}, \quad \frac{\partial H}{\partial x_i} = -\frac{dp_i}{dt}. \quad (3.7)$$

In the stationary case at constant temperature T and no volume change, this system can be simplified by introducing the (stationary) separable canonical density associated with the Hamiltonian approximation H [23]

$$f(\mathbf{x}, \mathbf{p}) = \frac{1}{Z} \exp(-\beta H(\mathbf{x}, \mathbf{p})) = X(\mathbf{x})P(\mathbf{p}), \quad \beta = 1/k_B T, \quad (3.8)$$

where k_B is the Boltzmann constant, and Z is the partition sum. This simplification becomes often a basis for MC computation where we typically keep those conformations where the Boltzmann factor $\exp(-\Delta E/k_B T)$ are larger than the random number from the interval $[0, 1]$. Our interest lies with the general dynamic case where we seek such positions of the atoms that satisfy the least action principle

$$\delta S = 0, \quad S = \int_{t_0}^{t_1} L dt, \quad L = K - P \quad (3.9)$$

where L is the Lagrangian of the system, δS is the first variation of S , and $[t_0, t_1]$ is the interval of observation. This leads to the system (3.1), or under appropriate assumptions, (3.7). In addition to an approximation of the Hamiltonian/Lagrangian of the molecular system, in order to start computation with a molecular modelling procedure, we have to compute good approximations to the initial conditions of the system, defined by the initial impulse (3.5) and the initial geometry of the system

$$\mathbf{x}(t_0) = \mathbf{x}_0. \quad (3.10)$$

The choice of initial geometries is an important issue for all energy-based molecular modelling procedures, and the role of this choice increases even further with the concept of conformation meta-stability being put forward and efficient dynamics-based clustering methods being developed [38, 23]. For successful simulations with MD procedures, which typically cover time-intervals of the order of tens of picoseconds up to a few nanoseconds [9], the quality of (3.10) is critical.

4 Accounting for the effect of topological constraints and other criteria in the design of distance geometry algorithms

When designing new materials and dealing with complex polymer and composite systems, little information is available about the high dimensional energy surface prior to minimisation by a molecular modelling code. However, it is known that the quality of this minimisation depends critically on the choice of the initial topology which, in turn, can be found using computational optimisation procedures based on

- deterministic (or grid search) methods that cover all areas of the potential energy surface systematically;
- stochastic methods (Monte-Carlo, DGAs, genetic algorithms, etc);
- molecular dynamics.

The most powerful (although still computationally very demanding) approach to the solution of many practical problems in materials science is the MD approach. While the MD methods are highly competitive in searching local conformational space, they often fail in generating starting geometries due to their inefficiency in overcoming energy barriers. The first two groups above are not influenced by high energy barriers and can be used for scanning large areas of the potential energy function. However, deterministic methods are becoming less feasible computationally and their efficiency decreases with the technology moving towards designing new complex material structures. Hence, only the stochastic group of methods is left as the most practical one for our purpose of building realistic initial configurations. Such initial configurations not only can then be relaxed effectively with MD, but efficient methodologies have already emerged to calculate such characteristics of the material as specific heat, thermal expansion tensor, compliance tensor and others directly from thermodynamic response relations, using MD simulations. Therefore, a good initial approximation becomes a key to success of the whole mathematical modelling exercise in this field. In choosing an algorithm for its construction we have to take into account several important issues.

Since the equilibration time for dense polymers is in many orders of magnitude larger than currently available MD simulation times, we have to be able to build an initial topological structure that resemble the equilibration one with little computational efforts. A closely connected to the **computing time** issue is the **algorithm efficiency**. Recall that the mean square end-to-end distance, an important relaxation characteristic of polymers, is approximated according to the Flory hypothesis by

$$\langle r^2 \rangle = NCl^2, \quad (4.1)$$

where C is the characteristic ratio of the polymer with the number N of skeletal bond lengths l . In order to achieve (4.1) with standard approaches we need a simulation time proportional to $O(N^3)$.

Finally, the algorithm for the construction of initial geometries has to be capable of dealing with **crosslink structures** such as polymer networks (both randomly- and end-crosslinked [13, 14]), composite materials such as interpenetrating polymer networks with organic and nonorganic fillers [40], etc. A good benchmark example for this capability of the algorithms is provided by high deformability networks with essentially complete recovery. Such networks were traditionally attributed to the domain of “rubber-like elasticity” where the crucial assumption, which usually justify the transition from the microscopic to the macroscopic level, is the *affine deformation* assumption [36]. This assumption states that points of crosslinkage move in such a way that components of the end-to-end length of each chain change in the same ratio as the corresponding dimensions of the bulk rubber during deformation. A network is seen then as a *phantom chain* which may pass freely through their neighbours and themselves and leads to the Gaussian distribution for the end-to-end vector of flexible polymers. More precisely, the number of possible confirmations for walks is assumed to be a function of the distance between the two endpoints taken as the Gaussian distribution for random walks. The resulting change in entropy under deformation (the entropic elasticity), leads to the elastic shear modulus (the plateau modulus measured by shear experiment)

$$G^0 \propto \frac{k_B T}{N_s} \quad (4.2)$$

with N_s being the characteristic of the chain length (the strand length between crosslinks). The point here is that for longer chains where $N_s > N_e$ (N_e is a characteristic entanglement length [27]), the approximation (4.2) can deviate substantially from the real situation. Shear modulus should be corrected to account not only for the entropic contributions, but also for the trapping contributions due to non-crossability of the chain [13, 14] since real networks are not phantom chains

$$G^0 = [((\nu - h\mu)/V)k_B T + T_e G_N^0], \quad (4.3)$$

where V is the total volume of the molecular system, ν is the number of elastically active strands, μ is the number of elastically active cross-linked, h is an empirical parameter ($0 \leq h \leq 1$), G_N^0 is the melt plateau modulus for the un-crosslinked system, and T_e is the trapping factor. There is evidence [26] that (4.3) converges to a limiting value around $k_B T/N_e$.

More generally, whenever $N_s > N_e$ the classical Rouse-type models [12] where the dynamics of short chains can be well understood with models based on the Langevin dynamics of individual random walks need correction due to the fact that the effect of topological constraints (entanglement effects) become markedly pronounced. For longer relaxation times (exceeding the Rouse relaxation time $\tau_e \propto N_e^2$) the constraints are supposed to become dominant and the chain moves along its own coarse grained contour. The transition from Rouse to reptation can be characterised by the diffusion coefficient $D(N_s)$ which is drastically decreasing during such a transition

$$D(N_s) \propto N_s^{-1} \implies D(N_s) \propto N_s^{-2}. \quad (4.4)$$

Dramatic qualitative changes are observed in other properties, including viscous, during such a transition.

Since the classical theories disregard topological constraints, and in real networks the strands are impenetrable and highly entangled, numerous attempts have been made to approximate topological constraints by local geometrical constraints [15]. However, only the reptation concept proposed by Edwards and developed by de Gennes takes the noncrossing of the chains explicitly into account. Incidentally, the idea of this concept originated from the investigation of the effect of topological constraints in polymer networks, and the observation that the topological constraints of each chain, as imposed by the surrounding, eventually cause a motion along the polymers own coarse-grained contour. "Tube" models are amongst simplest examples of approximating topological constraints by local geometrical constraints [16]. In particular, taking into account that the free ends of a melt would not alter significantly the behaviour for very long chains on scales much smaller than the chain diameter and for intermediate times, it is often assumed that they become confined to a tube with the diameter which coincides with the diameter of a subchain of length N_e , i.e.

$$\text{diam}_T \propto N_e^{1/2}. \quad (4.5)$$

Topologically trapped entanglements play an important role, particularly when the average strand length is much longer than the entanglement length of the un-crosslinked melt [13, 14]. Having accounted for these entanglements via topological constraints, many physico-chemical properties of polymer and composite materials, including elastic constants (Young's modulus, Poisson ratio, compressibility), thermodynamic properties and surface energies, stress-strain curves for finite deformations [18], can be defined more exactly. This can be done effectively using tools of molecular modelling, provided a good approximation to the initial topology has been obtained. Indeed, in designing new materials and modelling complex polymer and composite systems, a key to success is kept by topological interactions in networks, and it is advantageous to start simulations of such interactions with the distance geometry.

5 Distance geometry optimisation as a natural approach to modelling complex polymer systems and composites

In recent years classical distance geometry procedures, including metric matrix methods, embedding algorithms, torsional space methods, have provided important tools in molecular

modelling applications ranging from conformations of small molecules, protein and peptides to the time-average conformations of biological macromolecules (helping in the solution of the central problem of molecular biophysics), pharmacophore³ modelling and drug-receptor docking [6, 10, 21, 39]. In most of these applications the conformational space is search by generating a large number of independent solutions within the constraints of the model, and each structure is then evaluated for the inclusion in the final ensemble of low-energy conformations, using a force field or other energetic evaluation. Using such classical algorithms, the answer to the question on the existence of a model that satisfy our experimental constraints comes at a cost of dealing with (a) $\frac{1}{2}(N-1)N$ nonbonded terms (for structures of N atoms) [17], (b) $\sim 2N^2$ units of memory, (c) computationally costly smoothing techniques (i.e. triangle inequalities, see further details in Section 6).

In order to be successful for complex polymer and composite systems computational efficiency of distance geometry procedures used at the initial stage of molecular modelling should be improved. Since it becomes clear (see, for example, [39] and references therein) that such procedures have potential in the computer-aided molecular design where polymers and composites worthy of actual synthesis can be effectively predicted, in the last few years there has been an increasing interest to the development of such improved procedures. In the process of computer-aided molecular design distance geometry and molecular dynamics can be considered as complimentary methods [6] in a sense that efficient DG procedures can generate complex models rapidly at relatively small computational cost and produce structures that are appropriate starting point for molecular mechanics and dynamics calculations [41]. Although results on MD computations applied to polymers and composites are extensive, applications and analysis of improved DG procedures are still lacking in the literature.

Mathematically speaking, the problem of distance geometry applied to atomistic simulation of polymer and composite systems can be formulated as an optimisation problem.

1. Since measurements (obtained from NMR data or otherwise) may introduce errors into the symmetric matrix (of measured distances) D with nonnegative entries $d_{ij} \geq 0$, $i, j = 1, \dots, n$, $i \neq j$ and such that $d_{ii} = 0$, $i = 1, \dots, n$, in the general case this matrix, called a dissimilarity or pre-distance matrix, will deviate from the Euclidean matrix [1]. Recall ([31], e.g.) that if for a pre-distance matrix $D_{\text{Eucl.}} = (\delta_{ij})_{i,j=1,\dots,n}$ there exist n points $\mathbf{x}^l \in \mathbb{R}^r$, $l = 1, \dots, n$ such that

$$\|\mathbf{x}^i - \mathbf{x}^j\|_2 = \delta_{ij}, \quad (i, j) \in \mathcal{S}, \quad (5.1)$$

is satisfied, we call such a matrix the Euclidean distance matrix. In (5.1) δ_{ij} is the given distance between atoms i and j , \mathcal{S} is a subset (say, $i, j = 1, \dots, n$, $n \leq N$) of all atom pairs ($i, j = 1, \dots, N$), $\mathbf{x}^k \in \mathbb{R}^r$, $k = 1, \dots, N$ are the sought-for positions of the atoms in the molecular system, and r is the embedding dimension. Therefore, the first problem of distance geometry can be formulated as follows. Having a pre-distance matrix D , we have to "complete" it to a Euclidean distance matrix $D_{\text{Eucl.}}$. This problem is equivalent to the problem of finding such an Euclidean distance matrix $D_{\text{Eucl.}}$ from a set of all admissible Euclidean matrices that solves the weighted Euclidean matrix optimisation problem

$$f_1(D) := \|\Omega \circ (D - D_{\text{Eucl.}})\|_F \rightarrow \min, \quad (5.2)$$

³set of atoms/groups that are required for bio-activity of a molecule

where $\Omega = \{\omega_{ij}\}_{(i,j) \in \mathcal{S}}$ is the weight matrix (see [2] for a special case $H = I$), $\|\cdot\|$ denotes the Frobenius norm ($\|A\|_F = \sqrt{\text{tr} A^T A}$) and \circ denotes the Hadamard-Schur product (by definition, $A \circ B = (a_{ij} b_{ij})_{i,j=1,\dots,n}$, see, for example, [5, 4]). In [1, 2] (see also references therein) the problem (5.2) was considered for the case $n = N$.

2. In the general case we have to include the possibility of $n < N$ and this leads us to the problem of finding positions of atoms $\mathbf{x}^l \in \mathbb{R}^r$, $l = 1, \dots, N$ such that

$$f_2(\mathbf{x}) = \sum_{(i,j) \in \mathcal{S}} \omega_{ij} H(d_{ij}) \rightarrow \min, \quad (5.3)$$

where

$$d_{ij} \equiv \|\mathbf{x}^i - \mathbf{x}^j\|_2, \quad \text{and} \quad H(d_{ij}) = (d_{ij}^2 - \delta_{ij}^2)^2 \quad (5.4)$$

with known values of δ_{ij} . The problem (5.3)-(5.4) is solved by such $\mathbf{x} \in \mathbb{R}^{r \times N}$ that

$$f_2(\mathbf{x}) = 0. \quad (5.5)$$

We assume that the embedding dimension is 3 and $N \geq 4$ (since $r \leq N - 1$).

In reality, however, the distance matrix is given only approximately, and some elements are defined by lower and upper bounds that are not equal, as they are in (5.1). As a result, the process of converting the distance bonds matrix into Cartesian coordinates becomes much more tedious. Apart from the problem of whether the structure can be determined uniquely from incomplete but exact data given by inequalities ([20], e.g.), this leads to the problem of finding positions $\mathbf{x}^1, \dots, \mathbf{x}^N \in \mathbb{R}^3$ such that inequalities with the given lower, l_{ij} , and upper, u_{ij} , bounds on a set of distance constraints (5.1) with $\delta_{ij} \in [l_{ij}, u_{ij}]$ are satisfied, i.e.

$$l_{ij} \leq \|\mathbf{x}^i - \mathbf{x}^j\| \leq u_{ij}, \quad (i, j) \in \mathcal{S}. \quad (5.6)$$

Assuming (5.6) and following the approach originally proposed in [10], we are looking for the solution(s) of the following problem

$$\tilde{f}_2(\mathbf{x}) \rightarrow \min, \quad \mathbf{x} \in \mathbb{R}^{3N}, \quad (5.7)$$

where $\tilde{f}_2(\mathbf{x})$ can be thought as a “potential” function, although, strictly speaking, we do not invoke here any energetic considerations directly. The choice of the potential function $\tilde{f}_2(\mathbf{x})$ is not unique and can influence the computational effectiveness of the DG algorithm associated with such a choice. Here we aim only at a *representative sample* of local minimisers of the chosen objective function, and in the next two sections we concentrate of efficient algorithms for the construction of such samples.

6 Distance geometry algorithms with smoothing objective functions

To ensure stability, most existing energy optimisation codes require a starting model that is adequately close to the real topological structure. Of course, unstable energy-minimisation

processes can be controlled by “damping” the refinement (i.e. by decreasing the atomic shift). However, if the energy minimisation is unstable, even with damping, then it is necessary to go back to the starting model and to improve its topological representation. This process is trivial, at least from the theoretical point of view, for structures that can be determined by crystallographic or spectroscopic means (i.e. metals). For complex polymer and composite systems, where the structure may be unknown a priori with sufficient accuracy the construction of starting model⁴ becomes a non-trivial task of critical importance.

Although distance geometry algorithms have been extensively developed to provide better initial trial coordinates as well as better error refinement, it would be fair to say that such algorithms have been applied predominantly to small molecules. In application to large macromolecules and complex structures such as polymeric systems and composites, the choice of the objective function in the solution of problem (5.6)–(5.7) and the procedure for tracing local minimisers of this function deserve much more attention than they were accounted in the past.

Typical choices of the objective function try to make each term responsible for a specific constraint zero if the constraint is satisfied or monotonically increasingly positive as the violation of the constraint increases. This can be achieved for functions taken in the form

$$\tilde{f}_2(\mathbf{x}) = \sum_{(i,j) \in \mathcal{S}} \epsilon_{ij}^k, \quad \text{or} \quad \tilde{f}_2(\mathbf{x}) = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \epsilon_{ij}^k, \quad \text{where } \epsilon_{ij}^k : \mathbb{R}^n \rightarrow \mathbb{R}. \quad (6.1)$$

Particular examples include

$$\epsilon_{ij}^1 = \max^2[0, (d_{ij}^2 - u_{ij}^2)] + \max^2[0, (l_{ij}^2 - d_{ij}^2)], \quad k = 1, \quad (6.2)$$

$$\epsilon_{ij}^2 = \max^2 \left[0, \left(\frac{d_{ij}^2}{u_{ij}^2} - 1 \right) \right] + \max^2 \left[0, \left(\frac{l_{ij}^2}{d_{ij}^2} - 1 \right) \right], \quad k = 2, \quad (6.3)$$

$$\epsilon_{ij}^3 = \max^2 \left[0, \left(\frac{d_{ij}^2}{u_{ij}^2 - 1} \right) \right] + \max^2 \left[0, \left(\frac{2l_{ij}^2}{l_{ij}^2 + d_{ij}^2} - 1 \right) \right], \quad k = 3. \quad (6.4)$$

Such functions are, as a rule, non-differentiable, and lead to too many minimisers for any realistic structures. It is natural, therefore, to try to transform the original objective function to a smoother function that have smaller number of local minimisers, and then to trace back those minimisers to the original objective function. In the Wu-Moré method [31] the objective function is also taken in the standard form (6.1) with

$$\epsilon_{ik}^4 = \min^2 \left[\left(\frac{d_{ij}^2}{l_{ij}^2} - 1 \right), 0 \right] + \max^2 \left[\left(\frac{d_{ij}^2}{u_{ij}^2} - 1 \right), 0 \right], \quad k = 4. \quad (6.5)$$

This is reduced to (5.3) for the case where $l_{ij} = u_{ij} = \delta_{ij}$ and $\omega_{ij} = 1$. In this case the coordinates $\mathbf{x}^1, \dots, \mathbf{x}^N$ from the representative sample we mentioned above should satisfy the following inequalities

$$||\mathbf{x}^i - \mathbf{x}^j|| - \delta_{ij} \leq \epsilon, \quad (i, j) \in \mathcal{S}. \quad (6.6)$$

⁴that is a set of coordinates that defines the approximate geometry of the conformation and configuration of the molecule system of interest, and is input to the minimisation program

For fixed $\epsilon > 0$ this leads to the problem of finding ϵ -optimal solutions to our distance geometry problem which is known to be a NP-hard problem [20, 31].

With the Gaussian transform $g \in \mathbb{R}^n \rightarrow \mathcal{G}^\lambda(g) \in \mathbb{R}$

$$\mathcal{G}^\lambda(g) = \frac{1}{\pi^{n/2}\lambda^n} \int_{\mathbb{R}^n} g(\mathbf{y}) \exp\left(-\frac{\|\mathbf{y} - \mathbf{x}\|^2}{\lambda^2}\right) d\mathbf{y} \quad (6.7)$$

we transform our objective function \tilde{f}_2 defined by (6.1), (6.5) to a smoother function (with controlled degree of smoothing) with fewer local minimisers, then apply an optimisation algorithm to the transformed function, and finally trace the minimisers back to the original function [31]. For functions representable in the form $\tilde{f}_2(\mathbf{x}) = \sum_{(i,j) \in \mathcal{S}} h_{ij}(d_{ij})$ (in our case $h_{ij}(d_{ij}) \equiv \epsilon_{ij}^4$) it was shown in [31] that

$$\mathcal{G}^\lambda(\tilde{f}_2) = \sum_{(i,j) \in \mathcal{S}} \frac{1}{\sqrt{2\pi}d_{ij}} \int_{-\infty}^{+\infty} (d_{ij} + \lambda s) h_{ij}(d_{ij} + \lambda s) \exp(-s^2/2) ds. \quad (6.8)$$

This integral can be approximated with the Gauss-Hermite approximation

$$\mathcal{G}^\lambda(\tilde{f}_2) \approx \mathcal{G}^{\lambda,q}(\tilde{f}_2) = \sum_{(i,j) \in \mathcal{S}} \frac{1}{d_{ij}} \sum_{k=1}^q \omega_k(d_{ij} + \lambda s_k) h_{ij}(d_{ij} + \lambda s_k), \quad (6.9)$$

where ω_k and s_k , $k = 1, \dots, q$ are weights and nodes, respectively, for the Gaussian quadrature for the integral

$$I(g) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} g(s) \exp(-s^2/2) ds. \quad (6.10)$$

Then, in order to determine local minimisers of the function (6.9) for different values of λ , $\tilde{f} = \mathcal{G}^{\lambda,k}(\tilde{f}_2)$, $k = 0, 1, \dots, p$ ($\lambda_0 > \lambda_1 > \dots > \lambda_p$ and the original function is recovered in the limit $\lambda_p \rightarrow 0^+$), we follow [31] in using the quasi-Newton-type algorithm

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k H_k \nabla \tilde{f}(\mathbf{x}_k) \quad (6.11)$$

with $\alpha_k > 0$ and the approximation H_k of the inverse Hessian (k is the number of the current iteration). The quality of the result is judged upon the error functions

$$\epsilon_l = \min\left(\frac{d_{ij}}{l_{ij} - 1}, 0\right), \quad \epsilon_u = \max\left(\frac{d_{ij}}{u_{ij} - 1}, 0\right), \quad \epsilon_l^u = \max(\epsilon_l, \epsilon_u), \quad (6.12)$$

and the corresponding value of the objective function

$$\tilde{f}_2(\mathbf{x}^*) = \sum_{(i,j) \in \mathcal{S}} \left\{ \min^2 \left[\left(\frac{(d_{ij}^*)^2}{l_{ij}^2} - 1 \right), 0 \right] + \max^2 \left[\left(\frac{(d_{ij}^*)^2}{u_{ij}^2} - 1 \right), 0 \right] \right\}. \quad (6.13)$$

A typical result of computation with this algorithm, obtained for a short-chain polyethylene structure consisting of 250 backbone carbon atoms (polymer $-\text{CH}_2-\text{CH}_2-$ with 3-atom structural unit $-\text{CH}_2-$ [44]), is presented in Fig. 1. Such structures correspond to alkane liquids and important for petrochemicals and lubricants.

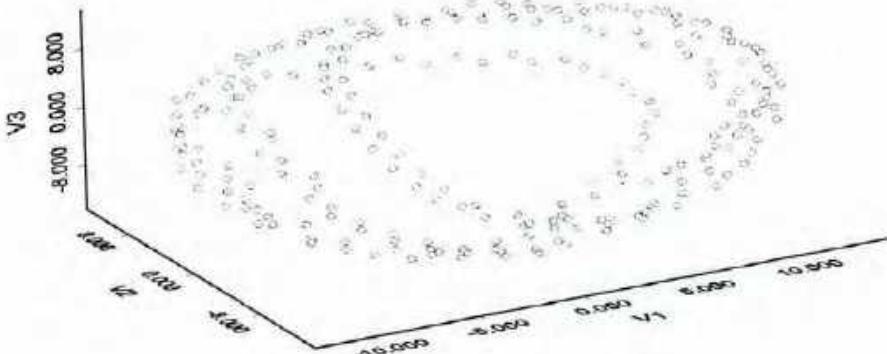


Figure 1: An alkane chain of 250 atoms isolated inspace (no periodic boundary conditions; all constraints)

In experimenting with this distance geometry algorithm no bound-smoothing procedures have been implemented. Such smoothing procedures, used in classical distance geometry algorithms, are usually implemented to lower some (unrealistic) default upper bounds in the constraints with the triangle inequality and to raise some (unrealistic) default lower bounds with the inverse triangle inequality. This step is aimed at achieving a better sampling of the conformational space at a high computational cost due to the increased number of effective constraints and the decrease in the regularity properties of the potential function. Despite the elimination of the bound-smoothing step, our experiments with larger chains and increased number of atoms showed that the algorithm described in this section is still computationally expensive. Although very reliable in calculating individual chains isolated in space, the algorithm needs further improvements in application to more complex structures.

Since in materials context the bulk properties are the basis of the economic value of the material, we have to be able to deal with bulk materials simulated by a large collection of molecules. In the next section we describe an efficient distance geometry algorithm capable of dealing with such materials.

7 Periodic boundary conditions and modelling bulk materials

The methodology described in the previous section is well suited for the analysis of the behaviour of a single chain isolated in space. However, in the case of bulk polymers and composites where the effect of entangled networks is markedly pronounced (see Section 4), the

efficiency of this methodology decreases. We recall that in order to produce near-equilibrium structures for dense bulk materials we can apply [42]

- Lattice construction technique;
- Van der Vegt compression box technique;
- Experimental density simulation-box technique.

The lattice construction technique is based on (a) meshing the 3D space of the experimental polymer volume with regular tetrahedrons, (b) placing tetravalent atoms (such as carbon or silicon) in the centre of a regular tetrahedron where their 4 bonds are perpendicular to the faces of this tetrahedron, and (c) constructing the polymer chains by generating random walks following adjacent tetrahedrons. Although this technique is claimed to scale linearly with the number of atoms [42], its application is limited to several classes of materials only.

The Van der Vegt compression box technique starts with a dilute model polymer and compresses it slowly until the target experimental density is achieved. In fact, the algorithm described in the previous section is relatively easy amenable to this technique, although such amendments come at a high computational cost. In dealing with bulk materials we can constraint a representative portion of the bulk molecular system to a given shape and then replicate it periodically in the 3D space. The constraint procedure can be built into one of the error functions (6.1), (6.5) by the definition a new, so-called “shape-function”

$$\tilde{h}(\mathbf{x}) = \begin{cases} \tilde{h} = 0, & \text{if } \mathbf{x} \text{ is on the boundary surface,} \\ \tilde{h} > 0, & \text{if } \mathbf{x} \text{ is outside of boundary surface,} \\ \tilde{h} < 0, & \text{if } \mathbf{x} \text{ is inside the boundary surface.} \end{cases} \quad (7.1)$$

Indeed, to constraint a representative portion of the bulk material we have to add to the error function the extra term

$$\max(0, \tilde{h}(\mathbf{x})), \quad (7.2)$$

and hence, from a mathematical point of view we reduce a constraint optimisation problem to a series of unconstraint problems.

In the remainder of this report we deal with the experimental density simulation-box technique consisting, as the name suggests, of packing chains into a simulation box at the experimental density. Instead of picking up distances randomly (between lower and upper bounds) as we did in the previous section, we pick up 3D coordinates within a chosen periodic cell creating by periodic boundary conditions. In this case we do not compress the structure, as in the classical embedding algorithms, due to its projection from $N - 1$ dimensions to three dimensions. To achieve a higher homogeneity of the resulting structure, we prescribe minimum and maximum bounds on the distance between *each* pair of atoms. These bounds are used then to define bond lengths and bond angles between neighbouring atoms, and the minimum separation of non-bonded atoms. In constructing the code, we assumed that atoms in different molecules can be treated as hard spheres and they have a minimum separation corresponding to a non-overlap condition which leads to the obvious condition for default lower bounds. The default maximum separation can be essentially arbitrary, but in some cases the assumption that all random points are confined to an “amorphous” cube can lead to “natural” default upper bounds (typically taken as the half of a diagonal of the cubic

cell). Atoms in the same molecules separated by many bonds (≥ 3) were treated in the same way as atoms in different molecules (i.e. non-bonded) which leads naturally to the situation where in some case nonbonded neighbours were closer than bonded neighbours. By creating a representative cell of the system and replicating it periodically in all direction in space [17], we begin simulations with random atom coordinates within this “central” cell with periodic boundary conditions. The size of the cell, taken as cubic, was determined by the density of material.

The basic idea of our algorithm is similar to classical procedures [10, 11, 6] and consists of minimising the error function by generating random conformers and then refining the obtained coordinates against an error (objective) function chosen here as

$$\tilde{f}_2(\mathbf{x}) = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \epsilon_{ij}^5, \quad \text{where } \epsilon_{ij}^5 = \max(l_{ij} - d_{ij}, d_{ij} - u_{ij}). \quad (7.3)$$

The main difference of the proposed algorithm compared to classical DG procedures is in a number of steps were taken to achieve savings in computational time for producing better starting conformations, in the sense that subsequent minimisation of the error function more often succeeds in reaching a value close to zero. Compared to classical distance geometry algorithms such as the metric matrix DG our algorithm is considerably simpler. In fact, we do not based our constructions on computing (three) largest positive eigenvalues and corresponding eigenvectors of the trial matrices which correspond to some conformations in \mathbb{R}^n and converting such matrices into a set of trial coordinates in \mathbb{R}^3 .

The main steps of our algorithm are as follows

- We select an atom from the given list at random (let it be, say, atom i);
- We calculate the distance between this atom and *each* other atoms;
- If the calculated distance falls outside of the prescribed bounds, we calculate the difference, i.e. the bounds error computed according to (7.3), and identify the largest of such bounds error (attained, say, for atom j);
- Then we move atom i along the path $i - j$ in order to satisfy the prescribed bounds by sampling the new distance randomly from a quadratic distribution between the minimum and maximum bounds (i.e. using a uniform distribution in \mathbb{R}^3);
- We repeat the above procedure for all atoms until all bounds are satisfied.

This algorithm has been tested for a number of bcc and fcc lattices and allow to obtain quickly initial topologies for structures many times larger than those reported in the previous section. Typical computational results for long linear alkanes, confined to a periodic cell, i.e. bulk polyethylene (solid or liquid) are presented in Fig. 2 for a chain of 1000 backbone atoms. In this series of experiments the size of a periodic cell was determined from the prescribed density of polyethylene at $\rho = 0.8\text{g/cm}^3$ and temperature 300°K . In particular, in Fig. 2 the size of the cubic cell is 32.1430\AA , while for a 2500 backbone atom structure it was 43.6305\AA (we computed structures with 2500 backbone carbon atoms and larger systems). The apparent separation of several atoms noticeable in this figure is due to the imposed periodic boundaries in this case.

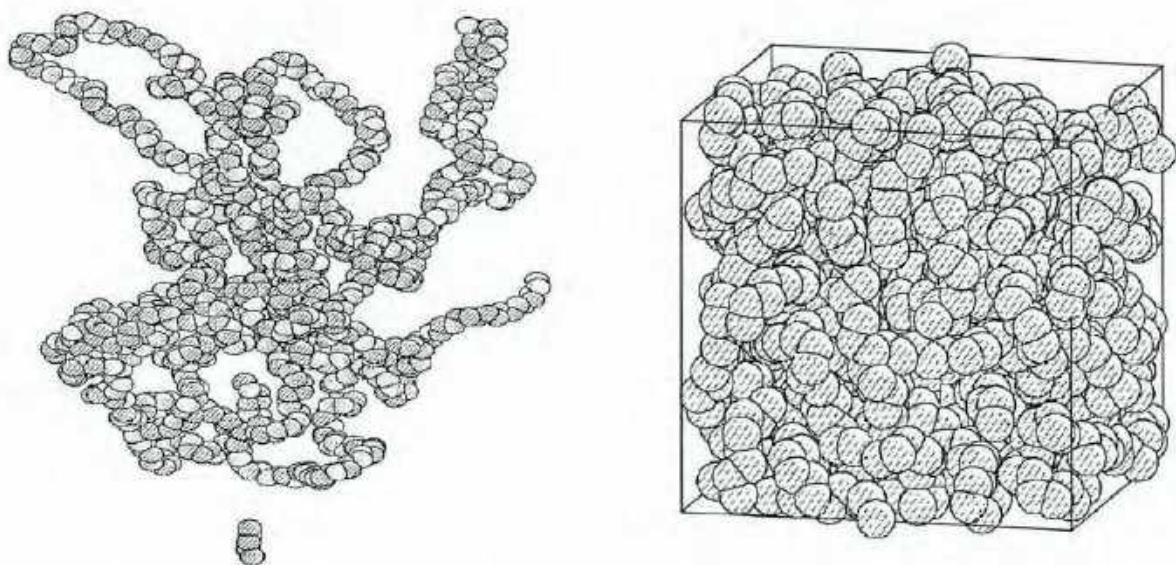


Figure 2: Structure from the representative sample for a 1000-atom alkane chain with periodic boundary conditions (all constraints); unpacked (left) and packed (right)

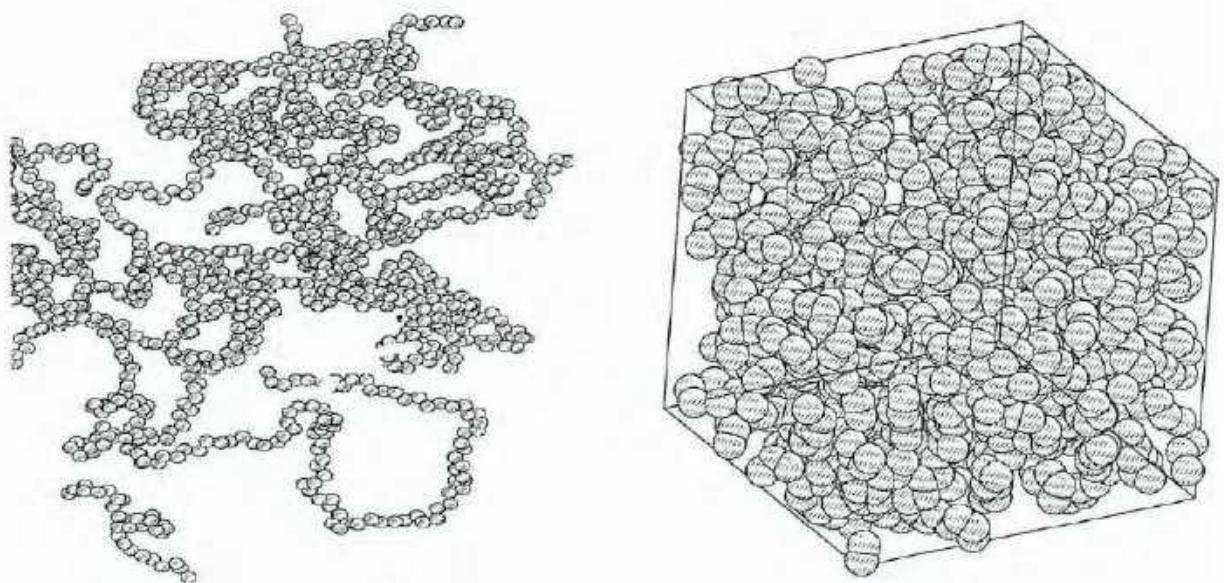


Figure 3: Structure 1 of the representative sample for 10 alkane chains 150 atoms each (periodic boundary conditions, all constraints); unpacked (left) and packed (right)

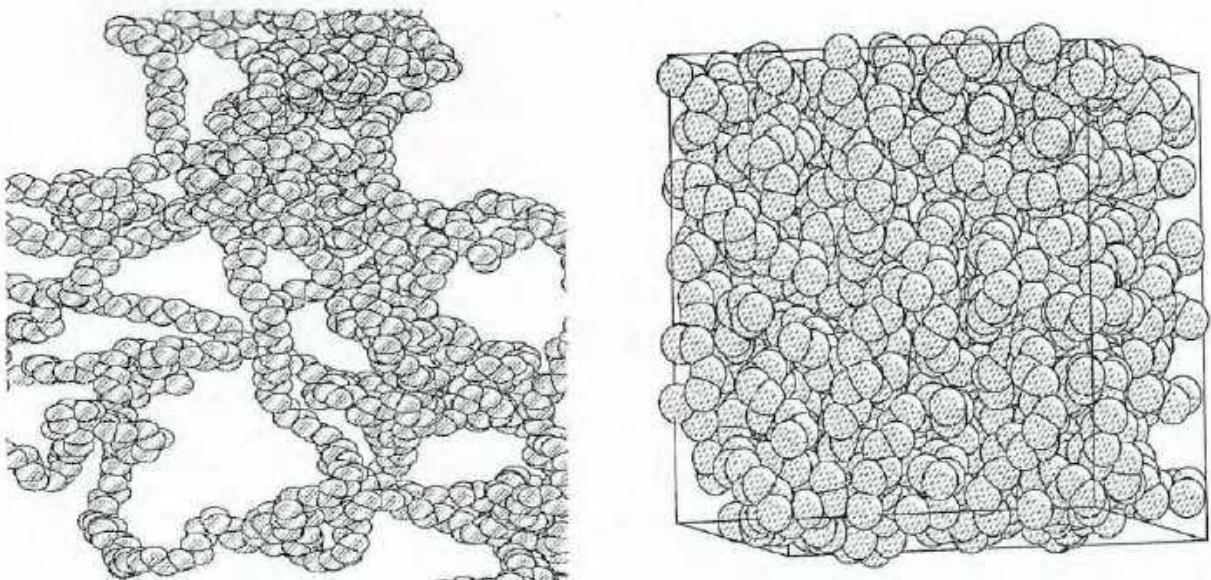


Figure 4: Structure 8 of the representative sample for 10 alkane chains 150 atoms each (periodic boundary conditions, all constraints); unpacked (left) and packed (right)

Since for many realistic polymer and composite systems, we have to deal with many chains confined to one molecular structure, in figure 3 and 4 we present two typical members of our representative sample for bulk linear polyethylene constructed from 10 chains of 150 backbone carbon atoms each. The cell box in this case was determined as 36.799\AA . It can be seen that all presented structures are sufficiently homogeneous to be relaxed efficiently with molecular dynamics codes, and we shall present the results of the MD refinement of these structures elsewhere.

Finally, in Fig. 5 we demonstrate the results of computation for bulk decane (decane molecule is ethylene's derivative with the degree of polymerisation equal 5) $C_{10}H_{22}$ simulating with 30 molecules confined to a periodic cell 21.3331\AA at density $\rho = 0.73\text{g/cm}^3$ and temperature 300°K . Although hydrogen atoms are often omitted from consideration by assuming that their positions can be well approximated and they can be added latter by "decorating" the structure (this idea was used in the previous groups of experiments), we took hydrogen atoms into account via packing constraints for this particular case. Since decane can produce a wide spectrum of volatile hydrocarbon products on degradation [3], we plan to investigate the effect of inclusion of hydrogen atoms into the model further.

The obtained conformations are now being analysed using the concepts of pair correlation (i.e. radial distribution functions), dihedral distribution, radius of gyration, cohesive energy, free volume, and Voronoi statistics [43, 37]. In particular, the radial distribution function

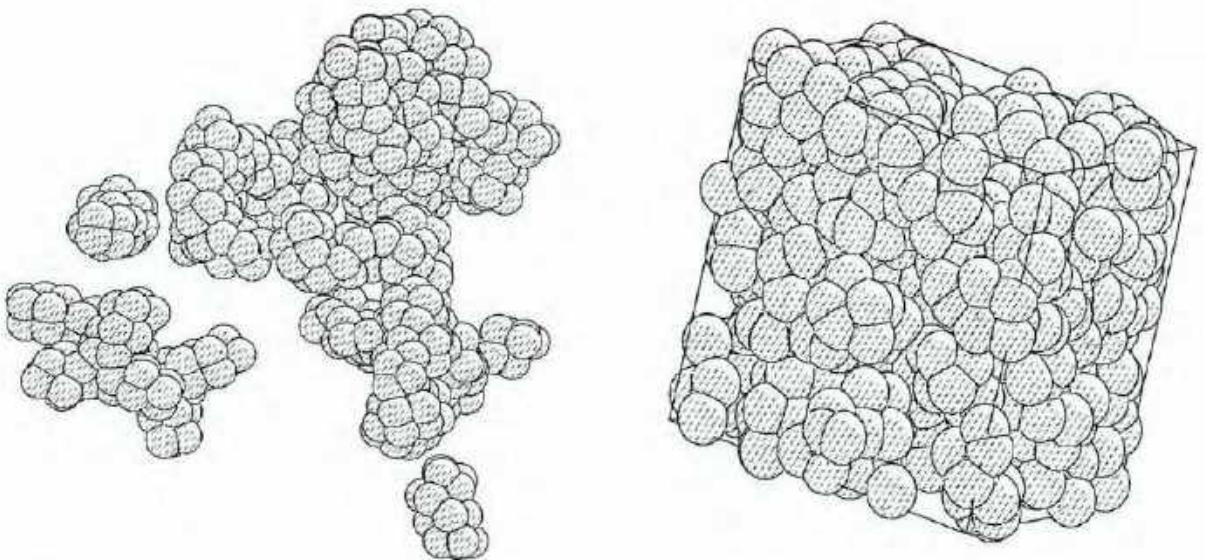


Figure 5: A member of the representative sample for bulk decane simulated by 30 molecules (periodic boundary conditions, all constraints); unpacked (left) and packed (right)

for the conformation shown in Fig. 5 confirms the presence of 2 peaks, one at 1.54 Å, and the other at 2.52 Å, as expected.

In modelling complex polymer and composite systems it is important to achieve a high level of vectorisation of computational procedures. Using NEC SX-4, we monitored the level of vectorisation of the algorithm presented in this section in a series of model experiments reported here. In particular, generating a representative sample of 10 structures for a 1500-atom model (10 chains, 150 backbone carbon atoms each) took 1302 sec with the level of vectorisation exceeding 99%. A similar computation for the decane model required 4876 sec with vectorisation 99.5%.

8 Conclusions and future directions

In this report we considered efficient distance geometry algorithms capable of dealing effectively with individual polymer chains isolated in space, as well as with large material structures, including bulk materials. By providing good approximations to topological molecular structures of materials such algorithms can increase the efficiency of available molecular modelling codes and therefore can lead to the possibility of calculating structural and mechanical properties of polymer and composite structures using available methodologies [18]. In this sense, the report is a contribution to the development of effective computational mathematics methodologies to design, characterise, and optimise polymers and composites before undertaking expensive experimental work.

As a development of the presented work we are now undertaking a comparison of the performance of several distance geometry algorithms for simulating bulk molecular cross-linked structures using end-to-end distributions, radii of gyration, and radial distribution functions. It is important to get further insight on the role of entanglement when the degree

of chemical cross-linking goes from high (where most of entanglement are locked) to relatively small. A related avenue of our current and future research lies with a detailed investigation of the dependence of distance geometry algorithms on the total numbers of atoms and on chain lengths for bulk materials, including high molecular weight polymers where the concept of entanglement becomes especially important.

References

- [1] Alfakih, A.Y., Khandani, A., and H. Wolkowicz, Solving Euclidean Distance Matrix Completion Problems via Semidefinite Programming, *Computational Optimization and Applications*, **12**, 1999, 13–30.
- [2] S. Al-Homidan, R. Fletcher, *Hybrid methods for finding the nearest Euclidean distance matrix*, in Recent Advances in Nonsmooth Optimization, Eds. D.-Z. Du, L. Qi and R.S. Womersley, World Scientific, 1995, pp. 1–17.
- [3] Allen, N.S. and Edge, M., Fundamentals of Polymer Degradation and Stabilisation, Elsevier, London, 1992.
- [4] S. Barnett, *Matrices: Methods and Applications*, Clarendon Press, Oxford, 1992.
- [5] R. Bhatia, *Matrix Analysis*, Springer, N.Y., 1997.
- [6] Blaney, J. M. and Dixon, J.S., Distance Geometry in Molecular Modeling, in Review in Computational Chemistry, Vol. V, Eds. K. B. Lipkowitz, D. B. Boyd, VCH Publishers, N. Y., 1994, 299 –335.
- [7] Blumenthal, L. M., The Theory and Applications of Distance Geometry, CHELSEA, 1970.
- [8] Bornemann, F., Homogenization in Time of Singularly Perturbed Mechanical Systems, Springer-Verlag, Berlin, 1998.
- [9] Comba, P. and Hambly, T. W., Molecular Modeling of Inorganic Compounds, VCH, Weinheim, 1995.
- [10] Crippen, G.M. and Havel, T.F., Distance Geometry and Molecular Conformation, John Wiley & Sons, N.Y., 1988.
- [11] Crippen, G.M., Smellie, A.S. and W. W. Richardson, Conformational sampling by a general linearised embedding algorithm, *Journal of Computational Chemistry*, Vol. 13, No. 10, 1992, 1262–1274.
- [12] Doi, M. and Edwards, S.F., The Theory of Polymer Dynamics, Clarendon Press, Oxford, 1995.
- [13] Duering, E. R., Kremer, K., and Grest, G. S., Relaxation of randomly cross-linked polymer melts, *Physical Review Letters*, **67**, 1991, 3531–3534.
- [14] Duering, E. R., Kremer, K., and Grest, G. S., Structure and relaxation of end-linked polymer networks, *J. Chem. Phys.*, **101**, 1994, 8169–8192.
- [15] Everaers, R. and Kremer, K., Test of the foundations of classical rubber elasticity, *Macromolecules*, **28**, 1995, 7291–7294.
- [16] Everaers, R. and Kremer, K., Topological interactions in model polymer networks, *Physical Review E*, **53**, 1996, R37-R40.
- [17] Gelin, B.R., Molecular Modeling of Polymer Structures and Properties, Hanser Publisher, Munich, 1994.
- [18] W.A. Goddard III et al, *Atomistic simulation of materials*, in Molecular Modeling, Ed.: M.A. Chaer Nascimento, World Scientific, Singapore, 1994, pp. 65–130.

- [19] Havel, T. F., Kuntz, I.D. and G.M. Crippen, The theory and practice of distance geometry, *Bulletin of Mathematical Biology*, Vol. 45, 1983, 665–720.
- [20] Hendrickson, B., The molecule problem: exploiting structure in global optimisation, *SIAM J. Optimization*, 5, 1995, 835–857.
- [21] Hodsdon, M. E., Ponder, J.W. and Cistola, D. P., The NMR solution structure of intestinal fatty acid-binding protein complexed with palmitate: application of a novel distance geometry algorithm, *Journal of Molecular Biology*, V. 264, No. 3, 1996, 585–602.
- [22] J.D. Honeycutt, *A general simulation method for computing conformational properties of single polymer chains*, Computational and Theoretical Polymer Science, 8 (1998), pp. 1–8.
- [23] Huisenga, W. et al, From simulation data to conformational ensembles: structure and dynamics based methods, ZIB, Preprint SC 98-36, 1998.
- [24] Karasawa, N., Goddard III, W. A., Force fields, structures, and properties of poly(vinylidene) crystals, *Macromolecules*, Vol. 25, No. 26, 1992, 7268–7281.
- [25] Karplus, M., Molecular dynamics: applications to proteins, in *Modelling of Molecular Structures and Properties*, Ed. Rivail, J.-L., Studies in Physical and Theoretical Chemistry, Vol. 71, 1990, Elsevier, 427–461.
- [26] Kremer, K., Numerical studies of polymer networks and gels, *Computational Materials Science*, 10, 1998, 168–174.
- [27] Kroger, M., Voigt, H., On a quantity describing the degree of chain entanglement in linear polymer system, *Macromol. Theory Simul.*, Vol. 3, No. 4, 1994, 639–647.
- [28] E. Leontidis et al, *Monte Carlo algorithms for the atomistic simulation of condensed polymer phases*, J. Chem. Soc. Faraday Trans., 91 (1995), pp. 2355–2368.
- [29] V. G. Mavrantzas et al, *End-bringing Monte Carlo: a fast algorithm for atomistic simulation of condensed phases of long polymer chains*, Macromolecules, 32 (1999), pp. 5072–5096.
- [30] Melnik, R.V.N., On consistent regularities of control and value functions, *Numer. Funct. Anal. and Optimiz.*, 18, 1997, 401–426.
- [31] Moré, J. J. and Wu, Z., Global Continuation for distance geometry problems, *SIAM J. Optim.*, Vol. 7, No. 3, 1997, pp. 814–836.
- [32] M. Murat and K. Kremer, *From many monomers to many polymers*, Journal of Chemical Physics, 108 (1998), pp. 4340–4348.
- [33] Newell, W. R. et al, Construction of genetic maps using distance geometry, *Genomics*, V. 30, No. 1, 1995, 59–70.
- [34] Nielsen, L. E. and Landel, R.F., Mechanical Properties of Polymers and Composites, Marcel Dekker, N.Y., 1994.
- [35] Nyden, M.R. and Gilman, J. W., Molecular dynamics simulations of the thermal degradation of nano-confined polypropylene, *Computational and Theoretical Polymer Science*, 7, 1997, 191–198.
- [36] Ogden, R. W., Elastic Deformation of Rubberlike Solids, In: H. G. Hopkins and M.J. Sewell, eds.; *Mechanics of Solids: the Rodney Hill 60th Anniversary Volume*, Pergamon Press, Oxford, 1982, 499–537.
- [37] Schweizer, K. S. and Curro, J.G., PRISM theory of the structure, thermodynamics, and phase transitions of polymer liquids and alloys, *Advances in Polymer Science*, 116, 319–377.

- [38] Shenkin, P. S. and D. Q. McDonald, Cluster analysis of molecular conformations, *Journal of Computational Chemistry*, 15, 1994, 899–916.
- [39] D. C. Spellmeyer et al, *Conformational analysis using distance geometry methods*, *Journal of Molecular Graphics and Modeling*, 15 (1997), pp. 18–36.
- [40] Stepto, R.F.T. (Ed.) *Polymer Networks: Principles of their formation, structure and properties*, Blackie Academic & Professional, London, 1998.
- [41] Theodorou, D.N. and U. W. Suter, Detailed molecular structure of a vinyl polymer glass, *Macromolecules*, 18, 1985, 1467–1478.
- [42] Tokarski, J. S. et al., Molecular modelling of polymers 17. Simulation and QSPR analyses of transport behavior in amorphous polymeric materials, *Computational and Theoretical Polymer Science*, 7, 1997, 199–214.
- [43] Polymer: User Guide, Part 1–3, Biosym/MSI, San Diego, 1995.
- [44] Young, R.J. and Lovell, P.A., *Introduction to Polymers*, Chapman & Hall, 1991, London.

USQ



TOOWOOMBA

**DYNAMICS OF SHAPE-MEMORY-
ALLOYS: A REDUCTION PROCEDURE
FOR 3D MODELS**

R V N Melnik, A J Roberts, K A Thomas
Department of Mathematics & Computing, USQ
Faculty of Sciences Working Paper Series
SC-MC-9906
4 May 1999

THE UNIVERSITY OF
SOUTHERN
QUEENSLAND

Working Papers

FACULTY
OF
SCIENCES

**DYNAMICS OF SHAPE-MEMORY-
ALLOYS: A REDUCTION PROCEDURE
FOR 3D MODELS**

R V N Melnik, A J Roberts, K A Thomas
Department of Mathematics & Computing, USQ
Faculty of Sciences Working Paper Series
SC-MC-9906
4 May 1999

ECCM '99

European Conference on
Computational Mechanics

August 31 – September 3
München, Germany

DYNAMICS OF SHAPE-MEMORY-ALLOYS: A REDUCTION PROCEDURE FOR 3D MODELS

R.V.N. Melnik, A. J. Roberts, K. A. Thomas

Key words: Shape Memory Alloys, Centre Manifold Technique

Abstract. Starting from a general Landau-type 3D model for the shape memory alloy dynamics we have developed a new mathematical "slow manifold" model that allows us to describe effectively the main features of the thermomechanical behaviour of CuAlNi alloys. A robust numerical procedure, developed in our earlier papers for the Falk model, has been generalised to this case. Results of the mathematical modelling of the thermomechanical fields in CuAlNi shape memory alloys are discussed with numerical examples.

1 Introduction

Shape-memory alloys are an intrinsic part of smart material and structure (SMS) technology with a wide variety of applications in such areas as aerospace engineering, medicine, manufacturing (including automotive industry and consumer products), civil infrastructure systems, biomechanics, oceanographic research, vibration control and suppression.

A smart structure is a non-biological physical structure having a definite purpose, means and imperatives to achieve that purpose and a biological pattern of functioning [25]. Smart materials are assumed to be a subset of smart structures considered at the microscopic or mesoscopic scales. Materials such as an optical fiber or a surface acoustic wave device (having the sensing function), piezoelectric devices or shape memory alloys (having the actuating function) are in this category. Closely related to SMS technology is the concept of intelligent materials. These materials are often defined as materials which respond to environmental changes, adjust themselves toward the optimum conditions, and manifest their functions according to these changes [27]. In addition to sensor and actuator functions, an intelligent material has to have "build-in" processor and feedback functions. Key areas of industrial research in the SMS technology are centred around:

- piezoelectric materials;
- shape-memory materials;
- magnetostrictive materials;
- electro- and magneto-rheological fluids;
- optical fibres and feedback control systems.

In this paper we address the adequate mathematical description of the dynamic behaviour of shape-memory alloys, materials with a rapidly expanding market of commercial exploitations [16]. Shape memory alloys are sensitive to both temperature and applied stress, and their thermomechanical properties are effectively used in the production of medical/biomedical devices (such as superelastic orthodontic archwires and optic lens holders), in minimally invasive surgery, control devices for robotics, sensor/actuator functional elements and many other applications in SMS technology [28].

The mathematical modelling of shape-memory materials and the analysis of associated models have become areas of increasing interest in the engineering, computational mechanics, and applied mathematics communities [17, 28, 12, 1, 4, 15, 6, 2, 9, 10]. This interest has two major sources: on the one hand, a wide range of important industrial applications of shape-memory materials; and on the other hand, a class of very challenging mathematical problems that arises in modelling the dynamics of these materials.

We organise this paper as follows.

- In Section 2 we describe main properties of shape memory alloys that are of industrial importance with the emphasis on copper-based alloys and explain the major difficulties in computational studies of these materials.
- In Section 3 we provide the analysis of different scales used in the modelling of shape memory alloy dynamics.
- Section 4 provides the reader with a three-dimensional Landau-type model used as a core model in our analysis.
- In Section 5 we apply a computer algebra implementation of slow manifold analysis to give a systematic approach for the modelling of thermomechanical behaviour of shape memory alloys on the mesoscopic scale.
- Some computational results obtained with our new model are presented Section 6.
- Conclusion and future directions are discussed in Section 7.

2 Copper-based shape memory alloys: thermomechanical properties and difficulties of their computer analysis

Under the action of thermal, mechanical, magnetic, hydrostatic or other fields some materials may restore their original shapes after being deformed. This property is usually termed the shape-memory effect and has been observed in a number of material systems such as metals, ceramics and polymers.

Although a large variety of materials can exhibit the shape memory effect, only those that

- can recover a substantial amount of strain or
- generate significant force upon changing shape

are of current industrial interest. The key in the wide applicability of shape-memory alloys is in a displacive diffusionless process, called the first order martensitic phase transformation, which leads to an internal structural change in the material where in certain temperature regimes two different phases (austenite and martensite) may coexist. We aim for the adequate description of the dynamics of this transformation using tools of mathematical modelling and computational experiment. At present, amongst the shape memory alloy family nickel-titanium (NiTi) and copper-base alloys are the most important. Mechanical properties of these materials vary greatly over the temperature range spanning their transformation. This brings difficulties in determining thermomechanical characteristics of shape memory alloys. For example, if the temperature is slightly above the transformation temperature for this material, we observe a nonlinear pseudoelasticity effect. In this case the material becomes extremely elastic and the elastic characteristics of the material (such as Young's modulus) become strongly dependent on both temperature

and strain deformation. Below we give a brief overview of thermomechanical characteristics of copper-based alloys, in particular copper-aluminium-nickel.

Despite its wide commercial exploitation, NiTi in finished form is very expensive, and in many applications Cu-based alloys such as CuZnAl and CuAlNi provide a more economical alternative to NiTi [29]. Compared to NiTi, a lower recoverable strain of the copper-based shape-memory alloys (around 4% compared to 8.5% for NiTi) has been making these alloys very attractive for the design of different types of actuators.

Perhaps one of the most important advantage of copper-based alloys lies with the fact that they have transformation temperatures well above NiTi alloys. Therefore, where higher temperature actuation is required, CuAlNi alloys are usually preferred since they can give a recovery temperature of up to 190°–200°C (the melting temperature is 1000°–1050°C) [22]. According to Hodgson, Wu and Biermann (see <http://www.sma-inc.com>), with the density 7.12 g/cm³ commercially available CuAlNi alloys have a thermal conductivity coefficient within 30 – 43 W/(m × °C) and a heat capacity 373–574 J/(kg × °C). These alloys have the yield strength around 400 MPa in the β parent phase and 130 MPa in the martensite (Young's modulus is 85 GPa in the β parent phase and 80 GPa in the martensite), whereas their ultimate tensile strength is up to 500–800 MPa.

Thermomechanical characteristics of CuAlNi have made these shape memory alloys the most appropriate for switching elements in circuit breakers and many other applications [21]. Copper-based shape-memory alloys are becoming more and more important in consumer goods manufacturing and are now used in fire protection devices, in actuators for anti-scald safety valves etc. However, note that these materials are quite sensitive to brittleness (at low temperatures) and instability (they are metastable in nature). The resistivity of these alloys is only 11–13 $\mu\text{Om} \times \text{cm}$ compared to 100 $\mu\text{Om} \times \text{cm}$ in the austenite state (70 $\mu\text{Om} \times \text{cm}$ in the martensite state) for nickel-titanium alloys. This leads to considerable difficulties in modelling the dynamics of CuAlNi alloys undergoing thermally- and mechanically-induced thermoelastic martensitic transformations. Due to its intrinsic metastability, in practical applications this material often requires training (i.e. a progressive modification of the admissible mixtures of martensites and austenite, produced by thermomechanical treatments of the alloy) in order to retain the parent β -phase for shape-memory effects.

Our main results in this paper concern copper-based shape memory alloys, in particular copper-aluminium-nickel alloys.

3 Spatial scales for modelling cubic-to-monoclinic phase transformations in copper-based shape memory alloys

The first step in the construction of mathematical and computational models for the analysis of thermomechanical behaviour of shape memory alloys is the choice of an appropriate spatial length scale. In principal, such models can be constructed on any of the atomic-, meso- or macro-scale levels. However, the deformation process of the phase transformation

simulated on the computer will be strongly dependent on the length of observation. In the majority of current applications the required length of observation may vary from a few nanometers to hundreds of micrometers ($\approx 10^{-9}$ – 10^{-4} m).

Using the Landau-Devonshire phenomenology, established by Falk on the mesoscopic scale, in our earlier papers [14, 15] we performed a computational analysis of the austenitic-martensitic phase transitions of shape memory alloys described by the following model

$$\left\{ \begin{array}{l} C_v \left[\frac{\partial \theta}{\partial t} + \tau_0 \frac{\partial^2 \theta}{\partial t^2} \right] - k_1 \left[\theta \frac{\partial u}{\partial x} \frac{\partial^2 u}{\partial t \partial x} + \tau_0 \frac{\partial}{\partial t} \left(\theta \frac{\partial u}{\partial x} \frac{\partial^2 u}{\partial t \partial x} \right) \right] - \mu \left[\left(\frac{\partial^2 u}{\partial t \partial x} \right)^2 + \right. \\ \left. \tau_0 \frac{\partial}{\partial t} \left(\frac{\partial^2 u}{\partial t \partial x} \right)^2 \right] - \nu \left[\frac{\partial \theta}{\partial t} \frac{\partial^2 u}{\partial t \partial x} + \tau_0 \frac{\partial}{\partial t} \left(\frac{\partial \theta}{\partial t} \frac{\partial^2 u}{\partial t \partial x} \right) \right] - \frac{\partial}{\partial x} \left(k \frac{\partial \theta}{\partial x} \right) = G, \\ \rho \frac{\partial^2 u}{\partial t^2} - \frac{\partial}{\partial x} \left[k_1 \frac{\partial u}{\partial x} (\theta - \theta_1) - k_2 \left(\frac{\partial u}{\partial x} \right)^3 + k_3 \left(\frac{\partial u}{\partial x} \right)^5 \right] - \mu \frac{\partial^3 u}{\partial x^2 \partial t} - \nu \frac{\partial^2 \theta}{\partial x \partial t} = F, \end{array} \right. \quad (1)$$

where u is the displacement field, θ is the temperature field, τ_0 is the thermal relaxation time, k is the thermal conductivity of the material, C_v is the specific heat constant of the material, μ and ν are material-specific coefficients that characterise the dependency of the stress on the rate of the deformation gradient and temperature respectively, ρ is the density of the material, θ_1 is a positive constant that characterises a critical temperature of the material, and k_i , $i = 1, 2, 3$ are material-specific constants that characterise the material's free energy. The right-hand sides of system (1), F and G , represent distributed mechanical and thermal loadings of the body.

Although this model was developed on the mesoscale level, it should be noted that the connection between different levels of descriptions manifests itself in the constitutive theories where one has to couple (a) the free energy, (b) the stress, and (c) the heat flow of the crystal with deformation, temperature and, when necessary, with their temporal and spatial gradients [14]. When only equilibrium properties of the material are of interest, such theories can be simplified with the stress represented by the derivative of the free energy function with respect to deformation. In model (1) we used the phenomenological Landau-Devonshire free energy function and the Cattaneo-Vernotte model for heat conduction. The shear strain on the habit plane (the contact plane between both phases) was taken as the basic deformation variable. In some cases we have to deal with habit planes of discontinuous deformation (for example, considering domain walls between martensitic variants, nucleation phenomena, or studying the interface between martensite and austenite). One way of dealing with such situations is to incorporate the Ginzburg term into the model (modify the Landau-Devonshire free energy function by adding the couple stresses term). Apart from the fact that this approach is disputable [26], it appeared that with reported values of the Ginzburg coefficient (10^{-10} – 10^{-12}) the gradient strain term has a negligible effect in the class of experiments we performed [14]. As we shall see in Section 5 a similar situation arises when the 3D Landau-type model is reduced to one dimension.

The model (1) was completed by appropriate initial and boundary conditions and was solved with respect to (u, θ) in the spatial-temporal region $Q = \{(x, t) | 0 \leq x \leq L, 0 \leq$

$t \leq T_f\}$, where L is the length of the structure and T_f is the required time of observation. The initial conditions for the model (1) were taken in the following form

$$u(x, 0) = u^0(x), \quad v(x, 0) = \frac{\partial u}{\partial t}(x, 0) = u^1(x), \quad \theta(x, 0) = \theta^0(x), \quad \frac{\partial \theta}{\partial t}(x, 0) = \theta^1(x), \quad (2)$$

with specified functions $u^0, u^1, \theta^0, \theta^1$. Boundary conditions are problem-specific [14]. In our experiments: mechanical boundary conditions were either specified stress or specified displacement,

$$s(0, t) = s_1(t), \quad s(L, t) = s_2(t), \quad \text{or} \quad u(0, t) = u_1(t), \quad u(L, t) = u_2(t); \quad (3)$$

thermal boundary conditions are those of specified heat flux

$$\frac{\partial \theta}{\partial x}(0, t) = \bar{\theta}_1(t), \quad \frac{\partial \theta}{\partial x}(L, t) = \bar{\theta}_2(t), \quad (4)$$

where functions $s_i(t)$ (or $u_i(t)$) and $\bar{\theta}_i(t)$, $i = 1, 2$ were given.

In this paper we continue investigation of meso-scale models of the Landau-type with the emphasis on the modelling of the copper-based shape memory alloy copper-aluminium-nickel. Recall that in the parent high temperature phase (austenite) the copper-based SMAs (such as CuZnAl and CuAlNi) has an ordered body-centred cubic (BCC) lattice. In the phase transformation “each” cubic cell of austenite is transformed into a tetragon, forcing the parent phase to be transformed to an ordered and twinned martensite. Therefore, it is fundamental to this type of transformations that both the parent austenitic phase and the martensitic product are ordered. In the general case one has to deal with 24 crystallographically equivalent martensitic variants (due to the transformation from the 48th order cubic symmetry group of the parent phase to the 2nd order monoclinic group of the product phase). The twinning property of the martensite is the result of minimisation of the stress on the habit plane. Therefore, as follows from the Falk deformation theory [6], due to high interface mobilities between martensite variants and between austenite and martensite variants, for the shape memory effect to occur this plane should be invariant. However, on the *micro-deformation level* (where only the lattice deformation, i.e. the Bain strain, is taken into account) the invariant habit plane does not exist [6]. Indeed, in the martensitic phase transformation the atomic displacements are very small and one can assume that no atomic migration is required for this transformation. Atoms cooperatively rearrange to form a new crystal structure. As a result of this assumption on the *mesoscale deformation level* we consider a coupled effect of the Bain strain and the lattice-invariant shearing mechanism. In the copper-based alloys of interest (CuZnAl and CuAlNi) the lattice-invariant strain is realised as a regular twinning on layers resulting from {110}-planes (called twinning or basal planes) of the BCC lattice (every third layer twins to form a large monoclinic martensite unit cell). It is this coupling effect between the Bain strain and the regular twinning that leads to “almost” invariant plane strain, i.e. a stress-free interface between austenite and martensite.

Now, when the habit plane is stress-free the sheared martensitic volume element will not fit into the region it occupied when it was in the austenitic state [6]. However, this

problem is easily solved on the *macroscopic deformation level* where we assume that the martensitic inclusion does not form as a single martensitic variant. A self-accommodating group of martensitic variants is formed into an aggregate of several martensitic variants, the so-called domain (sometimes called a microstructure) [6, 13]. For such domains the average deformation of their components cancel out. Due to this, on cooling the phase transformation from austenite to martensite occurs without a macroscopic shape change (scale of 10^{-4} m or larger) exhibiting a "ferroelastic" type of stress-strain curves. However, when the transformation is induced by a stress, in the temperature range where the stress-free austenite is stable, we have to deal with pseudoelastic stress-strain curves. One of the approaches that allow work on the macroscopic scale (where the deformation defined on a length scale of $100 \mu\text{m}$ or more) stems from Frémond's work (see references in [2]) and is based on the introduction of an internal variable that characterises fractions of austenite and martensite variants in shape memory alloys. We shall not consider this approach in this paper. We only note that one may expect that the Landau theory on the mesoscale level used in this paper is a consequence of the Landau theory on the microscale level and the constitutive relations of the macroscopic theories (including the Frémond theory) have to follow from the Landau theory considered on the mesoscale level. However, the construction of the hierarchy

Microscale Landau theory \Rightarrow Mesoscale Landau theory \Rightarrow Macroscopic theories

is just at the beginning of its development [6]. Hereafter we concentrate on a particular Landau-type models developed on the mesoscale level [5].

4 Three-dimensional Falk-Konopka free energy function and a Landau-type model for the dynamics of shape memory alloys

Numerical analysis of Landau-type models for the description of shape-memory-alloy dynamics has been typically restricted to the one-dimensional case [17, 9, 10]. Three-dimensional modelling in this field is traditionally associated with the application of Frémond's type models [2]. At the time of writing this paper we are not aware of any computational results obtained with the later models for physically realistic data. The Frémond models consider the phase proportions as thermodynamic quantities and typically include strain gradient terms to account for interfacial energies. This leads to a smoothing term in the momentum balance equation and simplifies analytical analysis of the models (see also Section 3). In the terminology discussed in Section 3, Frémond's models are considered on the macroscopic scale with temperature, macroscopic strain, strain gradient and the volumetric proportions of austenite and martensites (usually restricted by only two variants) being state variables. Such macroscale models ignore the internal atomic and mesoscale structures of a material using the assumption that the phases simultaneously present at each point of the material with appropriate proportions. In this case we have to deal with self-accommodating groups of martensite as a result of the deformation of a macroscopic sample, as explained in the previous section. We are

concerned with a smaller scale, which however is larger than the scale of the monoclinic lattice (i.e. the Bain strain is beyond of the resolution of our models). More precisely, our models describe the phase-transition between body-centred cubic austenite and monoclinic martensite variants on the mesoscopic scale. In order to describe the macroscopic behaviour of shape-memory-alloys an additional averaging (over the different martensite variants forming the macroscopic sample) procedure is required (see [5] and references therein).

Mathematical description of the appropriate mesoscale measure starts from the approximation by a polynomial (with coefficients depending on temperature) with respect to an order parameter characterising the phase transformation as follows [5, 14]

$$\Psi(\epsilon, \theta) = \psi^0(\theta) + \sum_{i=1}^{\infty} \psi^i(\epsilon, \theta) \quad (5)$$

where independent material parameters of the n -th order for $n = 1, 2, \dots$ are determined through strain invariants, \mathcal{I}_j^n as follows

$$\psi^n = \sum_{j=1}^{j^n} \psi_j^n \mathcal{I}_j^n \quad \text{and} \quad \psi^0(\theta) = \psi_0^0(\theta). \quad (6)$$

To make the free energy function invariant with respect to the symmetry group of austenite, the upper limit of the sum in (6), j^n , is chosen as the number of all invariant directions associated with a representation of the 48th order cubic symmetry group of the parent (austenite) phase.

For the copper-based alloys it is possible to reduce the number of required parameters in (5) to only 10 material constants (in the general case, temperature dependent) [5]

$$\Psi = \psi^0(\theta) + \sum_{i=1}^3 \psi_i^2 \mathcal{I}_i^2 + \sum_{i=1}^5 \psi_i^4 \mathcal{I}_i^4 + \sum_{i=1}^2 \psi_i^6 \mathcal{I}_i^6, \quad (7)$$

where the strain invariants \mathcal{I}_i^n of second, forth and sixth orders of the 48th order cubic symmetry group of the parent phase are determined as follows

$$\begin{aligned} \mathcal{I}_1^2 &= \frac{1}{9}(\text{tr}(\epsilon_{ij}))^2, \quad \mathcal{I}_2^2 = \frac{1}{12}(2\epsilon_{33} - \epsilon_{11} - \epsilon_{22})^2 + \frac{1}{4}(\epsilon_{11} - \epsilon_{22})^2, \\ \mathcal{I}_3^2 &= \epsilon_{23}^2 + \epsilon_{13}^2 + \epsilon_{12}^2, \quad \mathcal{I}_1^4 = (\mathcal{I}_2^2)^2, \quad \mathcal{I}_2^4 = \epsilon_{23}^4 + \epsilon_{13}^4 + \epsilon_{12}^4, \quad \mathcal{I}_1^6 = (\mathcal{I}_2^2)^3 \\ \mathcal{I}_3^4 &= (\mathcal{I}_3^2)^2, \quad \mathcal{I}_4^4 = \mathcal{I}_2^2 \mathcal{I}_3^2, \quad \mathcal{I}_5^4 = \epsilon_{23}^2 \left[\frac{1}{6}(2\epsilon_{33} - \epsilon_{11} - \epsilon_{22}) - \frac{1}{2}(\epsilon_{11} - \epsilon_{22}) \right]^2 + \\ &\quad \epsilon_{13}^2 \left[\frac{1}{6}(2\epsilon_{33} - \epsilon_{11} - \epsilon_{22}) + \frac{1}{2}(\epsilon_{11} - \epsilon_{22}) \right]^2 + \frac{1}{9}\epsilon_{12}^2(2\epsilon_{33} - \epsilon_{11} - \epsilon_{22})^2, \\ \mathcal{I}_2^6 &= \frac{1}{36}(2\epsilon_{33} - \epsilon_{11} - \epsilon_{22})^2 \left(\frac{1}{36}(2\epsilon_{33} - \epsilon_{11} - \epsilon_{22})^2 - \frac{1}{4}(\epsilon_{11} - \epsilon_{22})^2 \right)^2. \end{aligned} \quad (8)$$

Due to the symmetry, the strain tensor ϵ_{ij} , $i, j = 1, 2, 3$ forms a 6 dimensional space with the basis ϕ_K , $K = 1, \dots, 6$ [11]

$$\boldsymbol{\epsilon} = \sum_{K=1}^6 \epsilon_K \phi_K, \quad (9)$$

where

$$\epsilon_1 = \epsilon_{11}, \quad \epsilon_2 = \epsilon_{22}, \quad \epsilon_3 = \epsilon_{33}, \quad \epsilon_4 = 2\epsilon_{23}, \quad \epsilon_5 = 2\epsilon_{13}, \quad \epsilon_6 = 2\epsilon_{12}. \quad (10)$$

We assume that the strain tensor is coupled to the spatial displacements $\mathbf{u} = (u_1, u_2, u_3)$ of the material point with Lagrangian coordinates $\mathbf{x} = (x_1, x_2, x_3)$ at time t by the Cauchy relation

$$\epsilon_{ij}(\mathbf{x}, t) = \frac{1}{2} \left[\frac{\partial u_i(\mathbf{x}, t)}{\partial x_j} + \frac{\partial u_j(\mathbf{x}, t)}{\partial x_i} \right], \quad i, j = 1, 2, 3. \quad (11)$$

Finally, to complete the definition of the free energy function (7) we specify the parameters ψ_i^j ($j = 0, i = 0; j = 2, i = 1, 2, 3; j = 4, i = 1, \dots, 5; j = 6, i = 1, 2$) which for the copper-aluminium-nickel alloys (Cu is 14 wt% and Al is 3 wt%) are

$$\begin{aligned} \psi_1^2 &= 5.92 \times 10^6 \text{ g/(ms}^2\text{cm}), & \psi_2^2 &= (1.41 \times 10^5 + 46(\theta - 300)) \text{ g/(ms}^2\text{cm}), \\ \psi_3^2 &= (1.48 \times 10^6 + 940(\theta - 300)) \text{ g/(ms}^2\text{cm}), & \psi^0 &= -\alpha_1 \theta \ln[(\theta - \theta_0)/\theta_0] \text{ g/(ms}^2\text{cm}), \\ \psi_1^4 &= (-1.182 \times 10^8 + 3.55 \times 10^5(\theta - 300)) \text{ g/(ms}^2\text{cm}), \\ \psi_2^4 &= 3.13 \times 10^9 \text{ g/(ms}^2\text{cm}), & \psi_3^4 &= 1.64 \times 10^9 \text{ g/(ms}^2\text{cm}), \\ \psi_4^4 &= -5.53 \times 10^8 \text{ g/(ms}^2\text{cm}), & \psi_5^4 &= -4.27 \times 10^8 \text{ g/(ms}^2\text{cm}), \\ \psi_1^6 &= 3.35 \times 10^{10} \text{ g/(ms}^2\text{cm}), & \psi_2^6 &= 3.71 \times 10^{11} \text{ g/(ms}^2\text{cm}), \end{aligned} \quad (12)$$

where α_1 is the heat capacity of the material.

Having the free energy function, we define the shear stress by its three components: the quasi-conservative component, \mathbf{s}^q , the stress component due to mechanical dissipation, \mathbf{s}^m , and the stress component due to thermal dissipations, \mathbf{s}^t , (the latter is assumed to be negligible),

$$\mathbf{s} = \mathbf{s}^q + \mathbf{s}^m + \mathbf{s}^t, \quad \text{with} \quad s^q = \rho \frac{\partial \Psi}{\partial \boldsymbol{\epsilon}}, \quad s^m = \rho \mu \frac{\partial \boldsymbol{\epsilon}}{\partial t}, \quad s^t = 0. \quad (13)$$

If we assume that $\mu = 0$, then the relationship $\mathbf{s} = 0$ (i.e. the derivative of the free energy function with respect to $\boldsymbol{\epsilon}$) gives the necessary conditions for the minima of Ψ . Although we have 6 such conditions (see the representation (9)), we are only interested in the conditions associated with austenite and martensite phases. The first such conditions,

$\epsilon = 0$, is associated with the austenite phase which is stable when the Hessian of Ψ (computed with respect to ϵ) is positive definite. The second condition, derived in [5],

$$\epsilon = \begin{pmatrix} 2\alpha & \beta & 0 \\ \beta & 0 & -\beta \\ 0 & -\beta & -2\alpha \end{pmatrix} \quad (14)$$

corresponds to a monoclinic spontaneous strain with temperature-dependent parameters α and β subject to the following system of equations [5]

$$\begin{cases} 48\alpha^4\psi_1^6 + 8\alpha^2\psi_1^4 + 2\beta(\psi_4^4 + \psi_5^4) + \psi_2^2 = 0, \\ 4\alpha^2(\psi_4^4 + \psi_5^4) + 2\beta^2(\psi_2^4 + 2\psi_3^4) + \psi_3^2 = 0. \end{cases} \quad (15)$$

Given α^2 , from (15) we deduce β^2 and then the magnitude of the shear strain vector by the formula [5]

$$|\epsilon|^2 = 2 \left[1 - \sqrt{1 - 8(2\alpha^2 + \beta^2)} \right]. \quad (16)$$

The strain invariants in this situation are

$$\begin{aligned} I_1^2 &= 0, & I_2^2 &= 4\alpha^2, & I_3^2 &= 2\beta^2, & I_1^4 &= 16\alpha^4, & I_2^4 &= 2\beta^4, \\ I_3^4 &= 4\beta^4, & I_4^4 &= 8\alpha^2\beta^2, & I_5^4 &= 8\alpha^2\beta^2, & I_1^6 &= 64\alpha^6, & I_2^6 &= 0. \end{aligned} \quad (17)$$

For the given temperature these relations allow us to define the stress and the free energy function as functions of the monoclinic spontaneous strain defined by the matrix (14).

Now we are in the position to describe the dynamics of shape-memory alloys in the 3D case by the coupled system that consists of the equation of motion and the energy balance equation

$$\begin{cases} \rho \frac{\partial^2 \mathbf{u}}{\partial t^2} = \nabla_{\mathbf{x}} \cdot \mathbf{s} + \mathbf{F}, \\ \rho \frac{\partial e}{\partial t} - \mathbf{s}^T : (\nabla \mathbf{v}) + \nabla \cdot \mathbf{q} = g, \end{cases} \quad (18)$$

where \mathbf{F} and g are given distributed mechanical and thermal loadings of the body and $\mathbf{a}^T : \mathbf{b} = \sum_{i,j=1}^3 a_{ij} b_{ij}$ is the standard notation for the rank 2 tensors \mathbf{a} and \mathbf{b} . The velocity function \mathbf{v} , the internal energy function e and the heat flux \mathbf{q} are defined as follows

$$\mathbf{v} = \frac{\partial \mathbf{u}}{\partial t}, \quad e = \Psi - \theta \frac{\partial \Psi}{\partial \theta}, \quad \mathbf{q} = -k \nabla \theta - \alpha \frac{\partial k \nabla \theta}{\partial t}, \quad (19)$$

where the last expression is an approximation to the solution of the 3D Cattaneo-Vernotte equation [14] (computational experiments were performed with the classical Fourier law where $\alpha = 0$).

The embedding of constitutive equations that couples the free energy, the stress and the heat flow with the state variables into the basic laws of mechanics (18) provides the

foundation for the mathematical modelling of shape-memory-alloy dynamics. Of course, the specification of constitutive equations for shape memory materials is far from being unique and the development of thermodynamic constitutive models for these materials is an active area of research [4, 1, 8, 12]. Recently, Lurier [12] proposed a quite general approach to the construction of constitutive models for shape memory alloys (a number of earlier developed structural-analytical, phenomenological and micromechanical models follow from his model as special cases). Although we do not consider these constitutive models here, in the next section we develop a computer-algebra-based technique which, with appropriate modifications, is capable of effectively incorporating those models.

5 The construction of low-dimensional slow manifold models for the dynamics of shape memory alloys

We consider a shape memory alloy slab which is very large in the $x = x_1$ direction compared to its thickness of $2b$ in the $y = x_2$ direction ($-b < y < b$) and neglect any motion and dependence in the x_3 direction. In this section, we reduce the 3D model for the dynamics of shape-memory-alloys described in Section 4 to a simpler model expressed in terms of the amplitudes of cross-sectional averages of critical quantities. The basic idea of our approach is to express the physical fields in terms of asymptotic sums in these amplitudes and their longitudinal gradients. The asymptotic approximation is found to solve the system (18) describing the dynamics of SMA. This technique is at the heart of centre manifold theory, originated by Pliss [18]. During the last decade this technique has been successfully linked to computer algebra approaches and apply to a number of problems in continuum mechanics [3, 20, e.g.]. It is this linkage that makes the centre manifold method a powerful tool in the analysis of complex mathematical models. We refer the reader to [23, 7] (and the references therein) for the rigorous analysis of the reduction procedure of a system onto a centre manifold. Here, using similar techniques, we develop the slow, sub-centre manifold [24, §7] model by adapting the analysis of beam theory developed by [19] to the nonlinear dynamics of SMA. We show, using the computer algebra package **REDUCE**, how to derive systematically (up to the arbitrary order of accuracy) an accurate low-dimensional model for the description of thermomechanical behaviour of thin slabs in shape memory alloy materials.

For the unforced dynamics ($F = 0, g = 0$) we derive a model that describes the dynamics of slowly-varying (along the slab) modes. We note that in the presence of a time-dependent forcing, the system may substantially deviate from the slow manifold and the use of geometric projection of initial conditions is required to determine the forcing appropriate for the model. As for the boundary conditions, theory typically employs arguments based upon the spatial evolution away from the boundary which are applied to the original model and its approximation (see [20] and references therein). Here, we use "zero-stress" & "thermal-insulation" boundary conditions specified on $y = \pm b$, and "pinned" & "insulating ends" boundary conditions provide a leading approximation in the "long" direction (this, however, requires a quite delicate analysis which is outside the scope of this paper

[20]).

We model the long-wavelength, small-wavenumber modes along the slab. For simplicity in this first approach, we also assume that the effect of dissipation processes is negligible (we set $\alpha = \mu = 0$). Then, the eigenvalue analysis of the cross-slab modes shows that generally there is a zero eigenvalue of multiplicity five which corresponds to large-scale longitudinal waves, large-scale bending, and one heat mode (since dissipation effects were neglected, all other eigenvalues are pure imaginary). Thus there exists a slow sub-centre manifold based upon these five modes. This fact allows us to construct a new model on the sub-centre manifold (although the model will not have an assured guarantee of asymptotic completeness [23, 7, 3]).

Since the leading order structure of the critical eigenmodes are constant across the slab, the amplitudes of the critical modes are chosen to be the cross-sectional averages

$$U_i(x, t) = \bar{u}_i, \quad V_i(x, t) = \bar{v}_i, \quad \Theta'(x, t) = \bar{\theta}', \quad (20)$$

where an overbar denotes the y -average quantity and $\theta' = \theta - \theta_0$ ($\theta_0 = 300^\circ K$). Our low-dimensional model is written in terms of these parameters. We seek the low-dimensional invariant manifold upon which these amplitudes evolve slowly:

$$u_i = U_i(\mathbf{U}, \mathbf{V}, \Theta'), \quad v_i = V_i(\mathbf{U}, \mathbf{V}, \Theta'), \quad i = 1, 2, \quad \theta = \mathcal{T}(\mathbf{U}, \mathbf{V}, \Theta'), \quad (21)$$

$$\text{where } \frac{\partial U_i}{\partial t} = V_i, \quad \frac{\partial V_i}{\partial t} = g_i(\mathbf{U}, \mathbf{V}, \Theta'), \quad \frac{\partial \Theta'}{\partial t} = g_\theta(\mathbf{U}, \mathbf{V}, \Theta'). \quad (22)$$

The expressions (21) are substituted into the model (18) and the resulting system is solved to some order in the small parameters ∂_x , $E = \|\mathbf{U}_x\| + \|\mathbf{V}_x\|$ and $\vartheta = \|\Theta'\|$ with the computer algebra package REDUCE. Thus, here we treat the strains as small, as measured by E , while permitting asymptotically large displacements and velocities.

We use $\mathcal{O}(E^p + \partial_x^q + \vartheta^r)$ to denote an error in the model obtained by neglecting all terms involving $\partial_x^{\beta_1} E^{\beta_2} \vartheta^{\beta_3}$ such that $\beta_1/p + \beta_2/q + \beta_3/r \geq 1$. Then with errors $\mathcal{O}(E^5 + \partial_x^{5/2} + \vartheta^{5/2})$ the displacement and temperature fields of the slow manifold, in terms of the amplitudes and the scaled transverse coordinate $Y = y/b$, are

$$u_1 \approx U_1 - YbU_{2x} + 0.15(3Y^2 - 1)b^2U_{1xx}, \quad (23)$$

$$u_2 \approx U_2 - (0.9 - 3.05e-5\Theta')YbU_{1x} + 0.15(3Y^2 - 1)b^2U_{2xx} \\ - 141YbU_{1x}^3 + 1.00e-4(3Y - Y^3)b^3V_{1x}^2U_{1x}, \quad (24)$$

$$\theta \approx 300 + \Theta' - 2.43e6(3Y - Y^3)b^3(V_{1x}U_{2xx} + U_{1x}V_{2xx}) \\ - 25.1(7 - 30Y^2 + 15Y^4)V_{1x}^3U_{1x}. \quad (25)$$

As we noted in [14], the mechanical and thermal field approximations represented by (23)–(25) have cross-slab structure. In particular, the sideways deformation u_2 (which is a nonlinear function of the longitudinal strains) of the shape memory alloy feed back at higher order to contribute to and complicate the longitudinal and thermal dynamics.

The computer algebra code to derive these expressions is available upon request to the authors. Finally, with errors $\mathcal{O}(E^8 + \partial_x^4 + \vartheta^4)$ the model for the longitudinal dynamics on this slow manifold is

$$\left. \begin{aligned} \rho \frac{\partial V_1}{\partial t} &= c_1 U_{1xx} + \gamma_1 b^2 U_{1xxxx} \\ &\quad + \partial_x \left[(c_2 \Theta' - c_3 \Theta'^2) U_{1x} - (c_4 - c_5 \Theta') U_{1x}^3 + c_6 U_{1x}^5 \right. \\ &\quad \left. + (c_7 - c_8 \Theta') b^2 V_{1x}^2 U_{1x} + c_9 b^4 V_{1x}^4 U_{1x} - c_{10} b^2 V_{1x}^2 U_{1x}^3 \right], \\ \rho \frac{\partial V_2}{\partial t} &= -\gamma_2 b^2 U_{2xxxx}, \\ C_v \frac{\partial \Theta'}{\partial t} &= \kappa \Theta'_{xx} + (c_{11} + c_{12} \Theta' - c_{13} \Theta'^2) U_{1x} V_{1x} \\ &\quad + (c_{14} + c_{15} \Theta') V_{1x} U_{1x}^3 + c_{18} V_{1x} U_{1x}^5 \\ &\quad - (c_{16} + c_{17} \Theta') b^2 V_{1x}^3 U_{1x} - c_{19} b^2 V_{1x}^3 U_{1x}^3 - c_{20} b^4 V_{1x}^5 U_{1x} \\ &\quad + c_{21} b^2 U_{1xx} V_{1xx} + c_{22} b^2 U_{2xx} V_{2xx} + \partial_x^2 \left[-c_{23} b^2 U_{1x} V_{1x} \right], \\ \frac{\partial U_i}{\partial t} &= V_i, i = 1, 2, \end{aligned} \right\} \quad (26)$$

where coefficients c_k , $k = 1, \dots, 23$, are positive material constants.

In the first equation of this system the first line in the right-hand side describes linear dispersive elastic waves along the slab, whereas the second line gives the temperature dependent quintic stress-strain relation of the shape memory alloy (the analogue of the classical Falk representation). The remaining terms describe the effects due to rates of change of the strain [14]. The first two lines in the right-hand side of the third equation of the system describes the diffusion of heat generated/absorbed by mechanical strains, $\Theta U_{1x}^p V_{1x}$ terms. The remaining lines shows the effects of the internal pattern of mesoscopic strain. Finally, note that since there is no coupling of the longitudinal dynamics to the bending modes of the slab (to this order of truncation), the second equation of the system is the classic beam equation.

6 Numerical examples

Computational experiments reported in this section were performed for a thin single-crystal CuAlNi slab. The thermomechanical parameters for this material were chosen to be (see Section 2)

$$\rho = 7.12 \text{ g/cm}^3, \quad k = 0.0030 \text{ cmg}/(\text{ms}^3 \text{K}), \quad C_v = \rho \alpha_1 = 26.6 \text{ g}/(\text{ms}^2 \text{cmK}). \quad (27)$$

By assuming the slab is thin enough so that in effect $b = 0$ we reduce our model (26) to

the following differential algebraic system

$$\left. \begin{aligned} \rho \frac{\partial V_1}{\partial t} &= \frac{\partial s}{\partial x} + F, \\ C_v \frac{\partial \Theta'}{\partial t} &= k \frac{\partial^2 \Theta'}{\partial x^2} + (c_{11} + c_{12}\Theta' - c_{13}(\Theta')^2) \frac{\partial U_1}{\partial x} \frac{\partial V_1}{\partial x} + \\ &\quad + (c_{14} + c_{15}\Theta') \frac{\partial V_1}{\partial x} \left(\frac{\partial U_1}{\partial x} \right)^3 + c_{18} \frac{\partial V_1}{\partial x} \left(\frac{\partial U_1}{\partial x} \right)^5 + g, \\ s &= (c_1 + c_2\Theta' - c_3(\Theta')^2) \frac{\partial U_1}{\partial x} - (c_4 - c_5\Theta') \left(\frac{\partial U_1}{\partial x} \right)^3 + c_6 \left(\frac{\partial U_1}{\partial x} \right)^5, \\ \frac{\partial U_1}{\partial t} &= V_1. \end{aligned} \right\} \quad (28)$$

This system is then discretised in space using second-order approximations in space on staggered grids and solved with the differential-algebraic solver described in our earlier papers [14, 15]. The coefficients in the model (28) are obtained directly from the model (26) and for CuAlNi alloys have the following numerical values

$$\begin{aligned} c_1 &= 1.91 \times 10^6, & c_2 &= 592, & c_3 &= 0.00931, & c_4 &= 2.75 \times 10^9, & c_5 &= 8.42 \times 10^6, \\ c_6 &= 4.56 \times 10^{11}, & c_{11} &= 1.78 \times 10^5, & c_{12} &= 586, & c_{13} &= 5.94, \\ c_{14} &= 2.53 \times 10^9, & c_{15} &= 8.11 \times 10^6, & c_{18} &= 1.08 \times 10^{12}, & \gamma_1 &= 5.15 \times 10^5. \end{aligned} \quad (29)$$

If the thickness of the slab is not negligible compared to its length (taken to be 1 cm in our experiments), then the inclusion of all “*b*”-dependent terms in the model (28) is required. The corresponding coefficients for CuAlNi have the following values

$$\begin{aligned} c_7 &= 1811, & c_8 &= 5.64, & c_9 &= 0.728, & c_{10} &= 2.51 \times 10^3, & \gamma_2 &= 6.36 \times 10^5 \\ c_{16} &= 36.8, & c_{17} &= 0.00761, & c_{18} &= 1.08 \times 10^{12}, & c_{19} &= 1.016 \times 10^6, \\ c_{20} &= 0.0116, & c_{21} &= 1.05 \times 10^4, & c_{22} &= 5.92 \times 10^4, & c_{23} &= 5228. \end{aligned} \quad (30)$$

In the first experiment we study the dynamics of a shape memory alloy sample under the uniform mechanical forcing $F = -1 \times 10^{-3} \text{ g}/(\mu\text{s}^2\text{cm}^2)$ and time-dependent distributed heating/cooling according to the rule $g = 1.15 \times 10^{-3} \pi \sin^3(t\pi/15) \text{ g}/(\mu\text{s}^3\text{cm})$. The initial displacement field was taken to be

$$u^0(x) = \begin{cases} 0.0636x, & 0 \leq x \leq 1/2, \\ 0.0636(1-x), & 1/2 \leq x \leq 1. \end{cases} \quad (31)$$

We used “pinned ends” and “thermal insulation” boundary conditions. With the initial temperature 350°K the dynamics of the sample on the time interval $0 \leq t \leq 30 \mu\text{s}$ is presented in Figure 1. Upon heating, the martensite phase becomes metastable. Under appropriate thermomechanical conditions, this metastability may lead to the transformation of martensite phase into austenite. An attempt of this transformation is accompanied by a certain decrease in temperature, clearly visible in Figure 1. This decrease compensates for the given thermal forcing and the resulting temperature is insufficient to

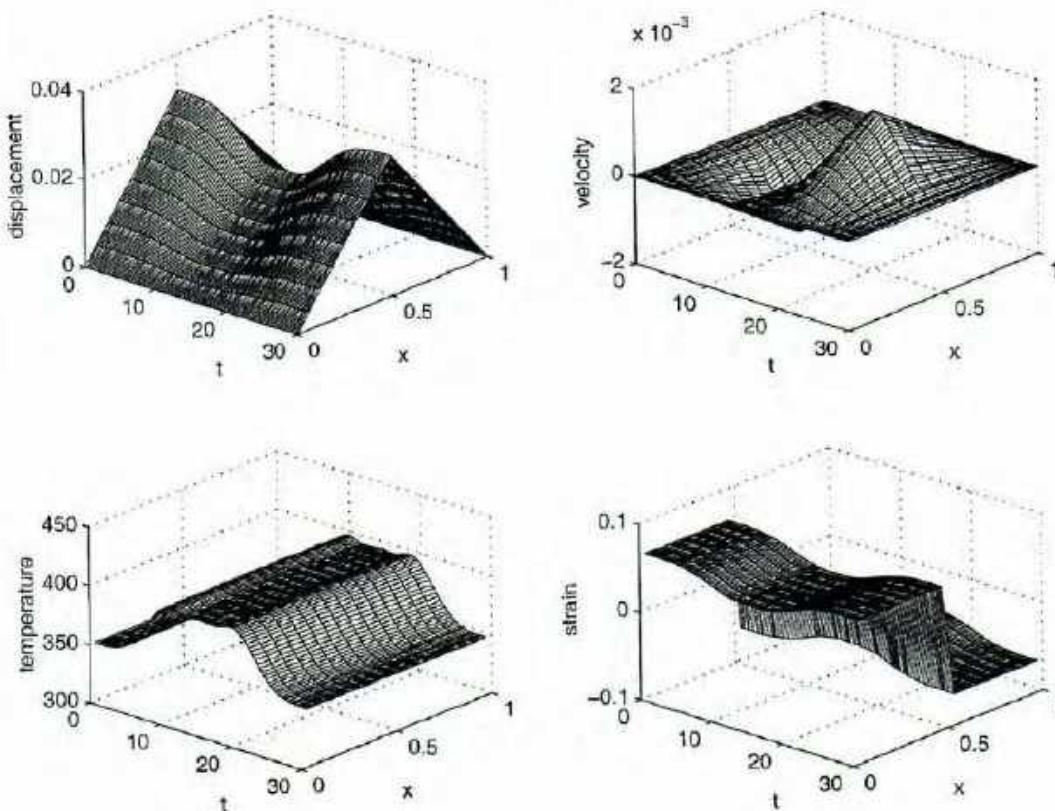


Figure 1: Thermal forcing of a CuAlNi rod.

transform martensite into austenite on this scale of observation. Therefore, upon cooling, the martensite phase returns to its original stable state.

The second experiment uses a purely mechanical ($g = 0$) time-dependent forcing given by the formula $F = 0.03 \sin^3(\pi t/15) \text{ g}/(\text{cm}^2 \mu\text{s}^2)$. We use the same boundary conditions as before. Starting with the initial configuration $u^0 = 0$ at the temperature 350°K we observe a strong coupling phenomenon between thermal and mechanical fields (see Figure 2). This phenomenon reminds us of the result of Experiment 2 from [14], where the dynamics of AuCuZn alloys was investigated. It is evident that it is much more difficult to retain the austenite phase in copper-aluminium-nickel compared to AuCuZn alloys and the computational analysis of stability of the parent phase for CuAlNi requires further investigation.

7 Conclusions

In this paper we presented results of an investigation into the thermomechanical behaviour of a copper-aluminium-nickel shape memory alloy. Using the three dimensional Landau theory describing the martensitic phase transformation of shape memory alloys we developed a systematic approach to the shape memory alloy modelling based on centre manifold techniques implemented in computer algebra packages. In particular, we

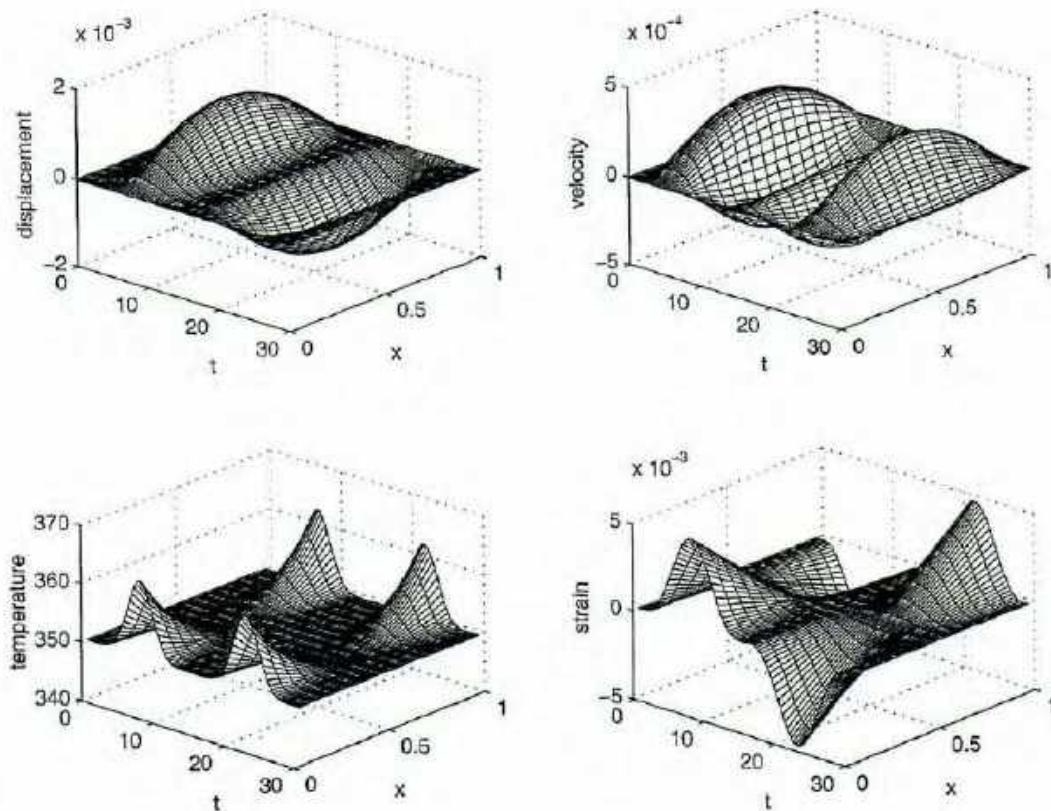


Figure 2: Mechanical forcing of a CuAlNi rod.

constructed a new model for the description of the shape memory alloy dynamics that allows us to describe essential dynamic behaviour of the system by determining the evolution on the crucial subset of all possible modes. The numerical scheme based on an effective differential-algebraic integrator provides a robust approach to the analysis of shape memory alloy dynamics. This analysis has been used in computational studies of thermomechanical behaviour of CuAlNi alloys.

The implementations of computational results into engineering applications are impeded by the absence of a theory that would allow us to compute thermomechanical characteristics of polycrystallines using such characteristics for single crystal. Effective averaging procedures over the grains are required in order to transfer highly nonlinear stress-strain curves for single crystals to polycrystals [6, 13]. Using such procedures, the technique proposed in our paper can be effectively adapted to the analysis of shape-memory alloy dynamics in polycrystalline structures as well as to other problems on structural phase transitions.

References

- [1] J.G. Boyd and D.C. Lagudas: *A thermodynamic constitutive model for shape memory materials, Part 1.* Int. J. Plasticity, **12**, No. 6, 805, (1996).

- [2] P. Colli and J. Sprekels: *Global Existence for a Three-Dimensional Model for the Thermo-Mechanical Evolution of Shape Memory Alloys.* Nonlinear Analysis: TMA, **18**, 873 – 888, (1992).
- [3] S.M. Cox and A.J. Roberts: *Centre manifolds of forced dynamical systems.* J. Austral. Math. Soc. Ser. B, **32**, 401–436, (1991).
- [4] V.G. DeGiorgi and H. Saleem: *A comparison of a few shape memory alloy constitutive models.* In: V.V. Varadan (ed.), Mathematics and Control in Smart Structures, Proc. of SPIE Vol. 3667, (1999).
- [5] F. Falk and P. Konopka: *Three-dimensional Landau theory describing the martensitic phase transformation of shape-memory-alloys.* J. Phys.: Condens. Matter, **2**, 61–77, (1990).
- [6] F. Falk: *On constitutive theories of shape memory alloys undergoing a structural phase transformation.* Material Science Forum, **123-125**, 91–100, (1993).
- [7] Th. Gallay: *A center-stable manifold theorem for differential equations in Banach spaces.* Commun. Math. Phys., **152**, 249–268, (1993).
- [8] E.J. Graesser, and F.A. Cozzarelli: *A proposed three-dimensional constitutive model for shape memory alloys.* Journal of Intelligent Material Systems and Structures, **5**, 78 – 89, (1994).
- [9] K.-H. Hoffmann and J. Zou: *Finite Element Approximations of Landau-Ginsburg's Equation Model for Structural Phase Transitions in Shape Memory Alloys.* M²AN, **29**, 629–655, (1995).
- [10] O. Klein: *Stability and uniqueness results for a numerical approximation of the thermomechanical phase transitions in SMA.* Advances in Mathematical Sciences and Applications (Tokyo), **5**, No. 1, 91 – 116, (1995).
- [11] P. Konopka and F. Falk: *Three-Dimensional Landau Theory Describing the Martensitic Phase Transformation of Shape-Memory Alloys.* Materials Science Forum, **123-125**, 113 – 122, (1993).
- [12] S.A. Lurier: *On thermodynamical constitutive relations for shape memory materials.* Mechanics of Solids, No. **5**, 110–122, (1997).
- [13] M. Luskin: *On the computation of crystalline microstructure.* Acta Numerica, **5**, 191–257, (1996).
- [14] R.V.N. Melnik and A.J. Roberts: *Approximate Models of Dynamic Thermoviscoelasticity Describing Shape-Memory-Alloy Phase Transitions.* to appear in the Proceedings of the NEMACOM'98, Centre for Mathematics and its Applications, Australian National University, (1999).
- [15] R.V.N. Melnik, A.J. Roberts and K.A. Thomas: *Modelling dynamics of shape-memory-alloys via computer algebra.* In: V.V. Varadan (ed.), Mathematics and Control in Smart Structures, Proc. of SPIE Vol. 3667, (1999).
- [16] N.B. Morgan and C.M. Friend: *Market Strategies for the Commercial Exploitation of Shape Memory Alloys.* J. Phys. IV France (Colloque C5), **7**, 615 – 620, (1997).

- [17] M. Niezgodka and J. Sprekels: *Convergent Numerical Approximations of the Thermomechanical Phase Transitions in Shape Memory Alloys.*, Numer. Math., **58**, 759–778, (1991).
- [18] V.A. Pliss: *A reduction principle in the theory of stability of motion.* Izv. Akad. Nauk. SSSR Ser. Mat., **28**, 1297–1324, (1964).
- [19] A.J. Roberts. *The invariant manifold of beam deformations. Part 1:the simple circular rod.* J. Elas., **30**, 1–54, (1993).
- [20] A.J. Roberts: *Low-dimensional modelling of dynamics via computer algebra.* Comput. Phys. Comm., **100**, 215–230, (1997).
- [21] E. Runtsch: *Shape memory actuators in circuit breakers.* In: T.W. Duering (ed.), Engineering Aspects of Shape Memory Alloys, Butterworth-Heinemann (1990), 330–337.
- [22] L. McDonald Schetky: *Shape memory alloy applications in space systems.* In: T.W. Duering (ed.), Engineering Aspects of Shape Memory Alloys, Butterworth-Heinemann (1990), 170–177.
- [23] L. P. Shilnikov et al: *Canter Manifold. Local Case.* In: L. P. Shilnikov et al, Methods of Qualitative Theory in Nonlinear Dynamics. Part 1, World Scientific (1990), 269–323.
- [24] J. Sijbrand. *Properties of centre manifolds.* Trans. Amer. Math. Soc., **289**, 431–469, (1985).
- [25] W.B. Spillman, J.S. Sirkis and P.T. Gardiner: *Smart materials and structures: what are they?* Smart Mater. Struct., **5**, 247–254, (1996).
- [26] J. Sprekels: *Shape memory alloys: mathematical models for a class of first order solid-solid phase transitions in metals.* Control and Cybernetics, **19**, No. 3-4, 287 – 308, (1990).
- [27] T. Takagi: *Recent research on Intelligent Materials.* Journal of Intelligent Material Systems and Structures, **7**, 346 – 352, (1996).
- [28] J. Van Humbeeck: *Shape Memory Materials: State of the Art and Requirements for Future Applications.*, J. Phys. IV France (Colloque C5), **7**, 3 – 12, (1997).
- [29] M. H. Wu: *Cu-based shape memory alloys.* In: T.W. Duering (ed.), Engineering Aspects of Shape Memory Alloys, Butterworth-Heinemann (1990), 69–88.

USQ



TOOWOOMBA

**ANALYSIS OF APPROXIMATE MODELS
FOR NONLINEAR CONTROL OF NON-
SMOOTH DYNAMIC SYSTEMS**

R V N Melnik
Department of Mathematics & Computing, USQ
Faculty of Sciences Working Paper Series
SC-MC-9905
26 March 1999

THE UNIVERSITY OF
SOUTHERN
QUEENSLAND

Working Papers

FACULTY
OF
SCIENCES

ANALYSIS OF APPROXIMATE MODELS FOR NONLINEAR CONTROL OF NON- SMOOTH DYNAMIC SYSTEMS

R V N Melnik

Department of Mathematics & Computing, USQ

Faculty of Sciences Working Paper Series

SC-MC-9905

26 March 1999

ANALYSIS OF APPROXIMATE MODELS FOR NONLINEAR CONTROL OF NON-SMOOTH DYNAMIC SYSTEMS

R.V.N. Melnik

Keywords : nonlinear control, non-reflexive Banach spaces, Hamilton-Jacobi-Bellman equations.

Abstract

In many applications of control and games theory we often encounter evolutionary partial differential equations, solutions of which are not smooth enough to satisfy these equations in the classical sense. A classical example is provided by Hamilton-Jacobi-Bellman equations that describe dynamics of value/cost functions which may not be differentiable everywhere. This paper is devoted to the analysis of such situations in both deterministic and stochastic cases.

1 Introduction

Non-smooth dynamic systems appear naturally and frequently in the control field [15]. Typically non-smooth systems fail the Lipschitz continuity requirement that is critical for the definition of classical solutions. Starting from Filippov's works, many fruitful ideas have been developed in this field including the Dini derivative technique, Clarke's generalised gradients, and the viscosity solution theory.

Until recently, the analysis of associated differential equations was predominantly conducted in reflexive Banach spaces. However, non-reflexive Banach spaces, such as L^1 , play an increasingly important role in the control context. This has been demonstrated by the introduction of L^1 control theory [14] as well as by an increasing number of engineering applications (see [7, 11] and references therein). The objective of this paper is to contribute to the construction and the analysis of approximate PDE models for the description of non-smooth control systems that work in a dynamic environment.

Three main approaches to the solution of optimal control problems have been reported. These are Pontryagin's maximum principle, Bellman's dynamic programming approach and the Markov control policy approach. Historically, deterministic control problems have been solved with the maximum principle approach whereas stochastic control problems have been solved with the dynamic programming approach. During recent years much effort has

been made to clarify the connection between these two approaches in the non-smooth case [5, 1, 16, 17]. Our current deliberation is based on the observation that both these approaches are ultimately connected with Markov control policy techniques [9]. A rigorous justification of this connection is well established only in the case when the value function is a classical solution of a HJB-type PDE [16, 8]. A recent breakthrough in this field was achieved with the viscosity solution theory (see [5, 1] and references in [8, 9]). Nevertheless some important questions remain open and they will be addressed in this paper.

2 Hamilton-Jacobi-Bellman Equation and Nonhomogeneous Conservation Laws

A close connection between Hamilton-Jacobi-Bellman-type equations and optimal control theory is well known [12]. Indeed, in order to derive a partial differential equation which describes the evolution of the optimal value/cost function, we can effectively use Bellman's dynamic programming approach. The resulting equation, considered in the open time-space domain $Q_T = I \times \Omega$ ($\Omega \subseteq B_1$, where B_1 is a given Banach space, $I = (t_0, T)$), has much in common with the nonhomogeneous conservation law (or the conservation law with source)

$$\frac{\partial u}{\partial t} + \frac{\partial F(t, x, u)}{\partial x} = G(t, x, u), \quad (2.1)$$

where u is the unknown function from the given Banach space B_2 , F and G are given functions such that $I \times B_1 \times B_2 \rightarrow \mathbb{R}$, $T > t_0$ ($t_0 \geq 0$) is the given number (the possibility of $T = \infty$ is not excluded). One can think of u as the conserved quantity, subject to the appropriate definition of its flux F and the source term G , the initial and boundary conditions.

By setting in (2.1) $F = -uf(t, x, u)$, $G = -u\frac{\partial f}{\partial x}$, we obtain the following equation

$$\frac{\partial u}{\partial t} - f(t, x, u) \frac{\partial u}{\partial x} = 0. \quad (2.2)$$

Apart from the fact that this equation (under appropriate assumptions) can be interpreted as the equation for

the governing dynamics of a control system, formally it is a special case of a more general partial differential equation, known as the Hamilton-Jacobi-Bellman (HJB) equation. The later equation describes the evolution of the value/cost function of the control system

$$\frac{\partial V}{\partial t} + H(t, x, V, DV) = 0, \quad (2.3)$$

where $V(t, x)$ denotes the value/cost function that is connected with function $u(t, x)$ by the control goal, $DV(t, x)$ denotes the Fréchet derivative with respect to x , and H (often interpreted as the system Hamiltonian) is the given function. Even in this relatively simple case classical global solutions for this equation may not exist and one has to specify in what sense we have to understand the solution of (2.3). Moreover, if the class of generalised solutions is specified, one may expect non-uniqueness of such solutions. A well-known non-uniqueness example provides the functional class $W_{loc}^{1,\infty}(Q_T) = \{V, DV \in L_{loc}^\infty(Q_T)\}$ (i.e. V and DV are measurable bounded on each open set, the closure of which is in Q_T) with $\Omega = \mathbb{R}$.

Another difficulty intrinsic to equation (2.3) manifests itself in the ‘curse of dimensionality’ when attempts to approximate this equation are made. The importance of approximations of Hamilton-Jacobi-Bellman equations for control applications is well known [2]. Along with the classical artificial/vanishing viscosity method and its modifications, one of the main bases for the development of effective numerical procedures for the solution of the HJB equation has been the conservation laws [6, 10]. In the one-dimensional case the analogy between HJB equation (2.3) and the homogeneous conservation law (i.e. equation (2.1) with $G = 0$) is straightforward [6]. This analogy becomes blurred if we recall that in practical applications only an approximation of the function H is available. This fact leads to the conclusion that the key role in the construction of numerical procedures for HJB systems has to be assigned to the non-homogeneous conservation law (2.1), rather than its homogeneous counterpart. The source term G in (2.1) can be interpreted as the perturbation source for the HJB system. Indeed, by applying Bellman’s dynamic programming approach in a heuristic manner we can only derive a HJB equation that gives an approximation to the dynamics of the value/cost function. Of course, in some cases we may be able to establish verification theorems that are useful in finding an optimal feedback. However, in the final analysis the quality of this feedback is subject to the quality of approximation of the function H which is associated with the system Hamiltonian. Such an approximation will be denoted by H_ϵ^δ [8]. Similar to the nonhomogeneous conservation laws, the natural space for perturbations in the systems control context is L^1 rather than L^2 [4]. Indeed, L^1 perturbations of control problems is the largest and the most reasonable functional space for perturbations in such problems where control systems work in a dynamic and uncertain environment. This situation arises in a number of important

applications of control systems in oceanographic research, aerospace engineering and smart structure technology.

3 Non-smooth Deterministic Control

Three of the most important concepts in the development of control theory have been the Pontryagin maximum principle, the Wiener-Hopf-Kalman H^2 optimal control theory and H^∞ robust control theory [2]. Although based on different ideas for disturbance rejection (a stochastic white noise disturbance model and a deterministic disturbance model respectively), both H^2 and H^∞ theories involve L^2 -measures for the controller performance (in order to quantify disturbance rejections in the frequency domain). However, when the control system works in a dynamic uncertain environment, it is important to be able to capture the worst-case peak amplitude response. In such cases L^1 theory becomes more appropriate [14, 2]. Another technique for disturbance rejection is implicitly implemented in the Pontryagin maximum principle. The rigorous logical basis of the Pontryagin maximum principle technique, which leads to a system of ordinary differential equations (rather than to a partial differential equation formally obtainable using the Bellman dynamic programming approach) comes at a cost of certain assumptions which we analyse below.

The application of Pontryagin’s maximum principle is quite natural for the solution of the following control problem

$$\begin{aligned} J(t, x; u) &= \int_{t_0}^T f_0(\tau, x(\tau), u(\tau, x(\tau))) d\tau + \\ &\quad \alpha g(T, x(T)) \rightarrow \min \end{aligned} \quad (3.1)$$

with the dynamics governed by the equation

$$\frac{dx}{dt} = f(t, x, u) \text{ a.e. in } [t_0, T]; \quad (3.2)$$

$$x(t_0) = x_0, \quad u \in U. \quad (3.3)$$

The problem (3.1) – (3.3) is the Bolza problem, where U is the given set of admissible controls, f_0, f are given maps on $I \times B_1 \times B_2$, g is the given map on $I \times B_1$, and α is the given real number.

We introduce Pontryagin-Hamilton’s function [13]

$$P(t, x, u, \psi^p, \delta) = -\delta f_0 + f \psi^p, \quad \delta \geq 0, \quad (3.4)$$

where δ is the scaling factor and ψ^p is the adjoint function. A standard practical recipe that follows from the Pontryagin maximum principle is to consider function $P(t, x, u, \psi^p, \delta)$ as a function of $u(\cdot, \cdot)$ taking all other variables (t, x, ψ^p, δ) as parameters. Then for each fixed set

(t, x, ψ^p, δ) we have to solve the following optimisation problem

$$P(t, x, u, \psi^p, \delta) \rightarrow \sup, \quad u \in U \quad (3.5)$$

The supremum in (3.5) will be denoted by

$$H_\epsilon^\delta(t, x, \psi^p, \delta) = \sup_{u \in U} P(t, x, u, \psi^p, \delta), \quad (3.6)$$

where the definition of the adjoint function ψ^p is subject to the solution of the linear ordinary differential equation

$$\frac{d\psi^p}{dt} = \delta \frac{\partial f_0}{\partial x} - \frac{\partial f}{\partial x} \psi^p. \quad (3.7)$$

Functions f_0 and f are coupled by control u . In the general case this coupling may be exhibited for all values of $t \in (t_0, T)$. This fact makes it difficult to construct a general model for the adjoint system dynamics. Indeed, one of the main premises for the Pontryagin maximum principle is the Hamiltonian system paradigm

$$\frac{dx}{dt} = \frac{\partial P}{\partial \psi}, \quad \frac{d\psi}{dt} = -\frac{\partial P}{\partial x}, \quad (3.8)$$

where one assumes that function P is a function of t, x and the generalised impulse ψ . However this analogy between the generalised impulse ψ and the adjoint function ψ^p may not be appropriate. Let $P = -\delta f_0 + f \psi$ be the Hamiltonian of the control system described by (3.1) – (3.3). Then both functions f and f_0 become dependent on ψ (due to their coupling via control), which makes the first equality in (3.8) only approximate (assuming small rates of change for f and f_0 with respect to ψ). Further we notice that

$$\frac{d\psi}{dt} = -\frac{\partial P}{\partial x} = \delta \frac{\partial f_0}{\partial x} - \frac{\partial f}{\partial x} \psi - f \frac{\partial \psi}{\partial x}. \quad (3.9)$$

If we assume that $f \frac{\partial \psi}{\partial x} = 0$, then equations (3.7) and (3.9) become identical, and hence $\psi^p = \psi$. However, in the general case the gradient $\frac{\partial \psi}{\partial x}$ may be arbitrarily large. This fact leads to essential difficulties in the application of Pontryagin's maximum principle to a number of control problems.

4 Connection between the Pontryagin Maximum Principle and the Bellman Dynamic Programming Approach

A natural question to ask is how to apply Bellman's dynamic programming approach to (3.1)–(3.3). It is also quite natural to ask how to apply Pontryagin's maximum principle to the solution of stochastic optimal control problems.

One of the key difficulties in answering these questions lies in the fact that a formal derivation of the HJB equation for deterministic control problems requires Taylor's expansion of the value/cost function. This leads to *a priori* excessive assumptions on smoothness of this function. We propose to relax these assumptions by applying Steklov's operator technique.

We recall that the Bellman approach is well suited to stochastic optimal control problems, i.e. for systems whose dynamics are described by the following governing equation

$$\begin{aligned} x(t) &= x(t_0) + \int_{t_0}^t f(\tau, x(\tau), u(\tau)) d\tau + \\ &\quad \int_{t_0}^t \sigma(s, x(s), u(s, x(s))) d\omega(s) \end{aligned} \quad (4.1)$$

with the requirement

$$E_{t,x}^u[J(t, x; u)] \rightarrow \min, \quad (4.2)$$

where functional $J(t, x; u)$ is defined by (3.1), $u(\cdot, \cdot)$ is the control function valued (as above) in subset U , $E_{t,x}^u[J(t, x; u)]$ is the expectation conditioned on $u(t, x)$, $x(\cdot)$ is a \mathbb{R} -valued process, f and σ are given functions that serve as drift and diffusion of this process, $\omega(\cdot)$ is a \mathbb{R} -valued process which serves as a "driving noise" for the control system. Since values of function u at time t may depend on information about the past states $x(\cdot)$ prior to t , function $u(t, x(\tau))$, $0 \leq \tau < t$ is often denoted simply as $u(\cdot)$. We prefer to explicitly indicate the state-dependency.

The result of the application of Bellman's dynamic programming approach to the problem (4.1)–(4.2) is a second order PDE (see [8, 17] and references therein). Formally, one can also write a PDE associated with deterministic optimal control problems such as the Bolza problem (3.1)–(3.3)

$$\frac{\partial V}{\partial t} + H_\epsilon^\delta \left(t, x, \frac{\partial V}{\partial x} \right) = 0, \quad (4.3)$$

where V is the value/cost function defined as

$$V(t, x) = \inf_{u(\cdot, \cdot) \in U(t, x)} J(t, x; u) \quad (4.4)$$

and

$$H_\epsilon^\delta \left(t, x, \frac{\partial V}{\partial x} \right) = \inf_{u \in U} \left(f_0 + f \frac{\partial V}{\partial x} \right). \quad (4.5)$$

The connection between the described approaches is well established only in the classical situation when $V \in C^{1,2}(Q_T)$. In this case

$$\frac{\partial V}{\partial x}(t, x^*(t)) = \psi^p(t), \quad (4.6)$$

where $x^*(\cdot)$ denotes the optimal path of the value/cost function. Of course, it is well known that the assumption

on continuous differentiability of the functional (3.1), required for this connection, does not hold in the simplest cases. However, this assumption can be essentially relaxed by using the viscosity solution theory [5, 1] (or equivalent approaches [12]). In this “relaxed” framework (for the non-smooth deterministic case) the connection between the Pontryagin maximum principle and the Bellman dynamic programming approach can be interpreted in the following way (see [16] and references therein)

$$D_x^- V(t, x^*(t)) \subset \{\psi^p\} \subset D_x^+ V(t, x^*(t)); \quad (4.7)$$

$$D_{t,x}^- V(t, x^*(t)) \subset \{H_i^\delta\} \subset D_{t,x}^+ V(t, x^*(t)), \quad (4.8)$$

where $D_{t,x}^\pm V$ denotes superdifferential/subdifferential in the (t, x) -variables and $D_x^\pm V$ denotes the superdifferential/subdifferential in the x -variable for each fixed value of $t \in (t_0, T)$. Inclusions (4.7), (4.8) indicate major difficulties with both classical approaches, namely the choice of the equation adequately describing the dynamics of the adjoint process, and infinite dimensionality of the HJB equation. These difficulties become obvious when the control problem is solved numerically. We also note that in the general stochastic case inclusion (4.8) may be violated [17]. In the next sections we develop a unified treatment that allows us to effectively deal with both deterministic and stochastic control.

5 Generalized PDEs of the HJB-type in Control Theory

Non-reflexive Banach spaces, such as L^1 , reflect the nature of control problems in the most complete way and allow us to explore further the interplay between optimal, robust and adaptive control paradigms in control theory. In what follows we consider Sobolev spaces $W_l^1(\Omega)$. We recall that a measurable function $f(x)$ belongs to $W_l^1(\Omega)$ (l is a positive integer) if $f \in L^1(\Omega)$, and if Sobolev generalised derivative $f^{(l)}$ of order l exists and $f^{(l)} \in L^1(\Omega)$. The classes W_l^1 are converted into Banach spaces by the introduction of the norm

$$\|f\|_{W_l^1} = \|f\|_1 + \|f^{(l)}\|_1, \quad \|\cdot\|_1 \equiv \|\cdot\|_{L^1}. \quad (5.1)$$

For an elementary time-space region

$$Q_0 = \{(t', x') : t \leq t' \leq t + \Delta t, x \leq x' \leq x + \Delta x\} \quad (5.2)$$

we introduce Steklov's averaging operators as follows

$$S^t V(t, x) = \int_t^{t+\theta_1 \Delta t} V(\eta, x) d\eta, \quad (5.3)$$

$$S^x V(t, x) = \int_x^{x+\theta_2 \Delta x} V(t, \mu) d\mu, \quad (5.4)$$

where $0 \leq \theta_i \leq 1, i = 1, 2$. Choices of $\omega_1 = \Delta x$ and $\omega_2 = \Delta t$ are coupled to initial conditions (t, x) and value/cost

function $V(t, x)$. Using the Steklov operator technique we have obtained a Local Optimality Principle that governs the evolution of the value/cost function.

Theorem 5.1 *If $V \in W_1^{1,1}(Q_0)$, then the evolution of the value/cost function is described by the following integro-differential equation*

$$\begin{aligned} & \omega_1^2 \frac{\partial V}{\partial x} + \omega_2^2 \frac{\partial V}{\partial t} + \omega_1 \omega_2 \left[\frac{\partial V}{\partial t} + \frac{\partial V}{\partial x} \right] + \\ & \omega_1 \omega_2 \left[\omega_1 \frac{\partial^2 V}{\partial t \partial x} + \omega_2 \frac{\partial^2 V}{\partial x \partial t} \right] + \\ & \frac{1}{2} \left\{ \int_x^{x+\Delta x} \frac{\partial^2}{\partial \xi^2} (S^x V)(x + \Delta x - \xi) d\xi \Big|_{t+\Delta t} + \right. \\ & \left. \int_t^{t+\Delta t} \frac{\partial^2}{\partial \eta^2} (S^t V)(t + \Delta t - \eta) d\eta \Big|_{x+\Delta x} \right\} + \\ & \omega_2 (\omega_1 + \omega_2) f_0 = 0. \end{aligned} \quad (5.5)$$

We emphasise that this result is independent of the type of governing equation for the system dynamics and include both cases (3.1)–(3.3) and (4.1)–(4.2).

Our approach to the construction of a general model for the evolution of the value/cost function has much in common with the robust sample-data control theory. It leads to the rapprochement of (worst-case) L^1 theory and statistical approaches in control problems [7]. Indeed, using available statistical information we can construct such functional approximations for ω_1 and ω_2 that allow efficient continuation procedures for the equation (5.5) to the region Q_T . One can view such continuations as supervisory on-line control procedures (see [8] and references therein). The necessity of such supervisory procedures can be explained from the game theoretic point of view where the model uncertainty and the control can be seen as strategies employed by opposing players in a game (control is taken to minimise value/cost function, while the uncertainty opposes it by trying to maximise this function [3]). In this context we recall that both classical approaches (the Pontryagin maximum principle and the Bellman dynamic programming) can be directly applied to controlled systems with a complete model description. This description is unavailable when the systems work in a dynamic environment in the presence of any sort of uncertainty such as parametric type, unmodelled dynamics or external perturbations [3]. Nonreflexive Banach spaces are the most suitable functional spaces for the construction of control models in such situations.

Now we are in a position to derive some important special cases of (5.5).

If $V \in W_1^{2,2}(Q_0)$ then the stochastic control problem

(4.1)–(4.2) is equivalent to the following second order PDE

$$(1+v) \left[\frac{\partial V}{\partial x} + \frac{1}{v} \left(\frac{\partial V}{\partial t} + f_0 \right) \right] + \omega_1 \frac{\partial^2 V}{\partial t \partial x} + \omega_2 \frac{\partial^2 V}{\partial x \partial t} = \sigma_1 \frac{\partial^2 V}{\partial x^2} + \sigma_2 \frac{\partial^2 V}{\partial t^2}, \quad (5.6)$$

where $v = \omega_1/\omega_2$ and σ_1, σ_2 are σ -dependent diffusion functions. The later functions vanish in the case of partial differential equations that deal with non-smooth deterministic controls. More precisely, the following result holds.

Theorem 5.2 If $V \in W_1^{1,1}(Q_0)$, $\frac{\partial V}{\partial x} \in L^1(W_1^1(I), \Omega)$, $\frac{\partial V}{\partial t} \in L^1(I, W_1^1(\Omega))$, then the equation

$$(1+v) \left[\frac{\partial V}{\partial x} + \frac{1}{v} \left(\frac{\partial V}{\partial t} + f_0 \right) \right] + \omega_1 \frac{\partial^2 V}{\partial t \partial x} + \omega_2 \frac{\partial^2 V}{\partial x \partial t} = 0 \quad (5.7)$$

provides $\mathcal{O}(\Delta t + \Delta x)$ approximation of the Local Optimality Principle.

The most difficult case for analysis is the deterministic limit of control problems. Assuming $V \in L^1(Q_0)$, this case leads to the equation

$$(1+v) \left[\frac{\partial V}{\partial x} + \frac{1}{v} \left(\frac{\partial V}{\partial t} + f_0 \right) \right] = 0, \quad (5.8)$$

which (under certain conditions) coincides with the non-homogeneous conservation law.

6 Local Vector Fields in Solutions of Nonlinear Control Problems

Let $\vec{\psi} = (\psi_1, \psi_2)$, where

$$\psi_1(t, x) = \int_{x_0}^x V(t, \mu) d\mu, \quad \psi_2(t, x) = \int_{t_0}^t V(\eta, x) d\eta. \quad (6.1)$$

We introduce non-smooth parts of the Hamiltonian using a composition of Steklov's operators as a performance measure in Q_0 :

- non-smooth wrt t ($\forall x' \in [x, x+\Delta x]$ and a.e. in $t \in I$)

$$H_1 = S^t \otimes S^x \left[\int_t^{t+\Delta t} f_0 d\tau \right] + S^t \psi_1 \quad (6.2)$$

- non-smooth wrt x ($\forall t' \in [t, t+\Delta t]$ and a.e. in $x \in \Omega$)

$$H_2 = S^x \otimes S^t \left[\int_t^{t+\Delta t} f_0 d\tau \right] + S^x \psi_2, \quad (6.3)$$

where \otimes denotes the composition of the corresponding integral operators.

Then system dynamics in the non-smooth case can be described by the following coupled system of canonic equations

$$\begin{cases} \frac{\partial H_1}{\partial t} + \kappa_1 \operatorname{div} \vec{\psi} = 0, & \kappa_1 = \frac{1}{2} \left(\frac{\omega_1}{\omega_2} + \omega_1 \right), \\ \frac{\partial H_2}{\partial t} + \kappa_2 \operatorname{div} \vec{\psi} = 0, & \kappa_2 = \frac{1}{2} \left(\frac{\omega_2}{\omega_1} + \omega_2 \right). \end{cases} \quad (6.4)$$

Theorem 6.1 If $\omega_1 + \omega_2 = 2$, then a.e. in Q_0 , H_1 and H_2 are monotone functions in t and x respectively, simultaneously increasing or decreasing:

$$\frac{\partial H_1}{\partial t} \frac{\partial H_2}{\partial x} = (\operatorname{div} \vec{\psi})^2. \quad (6.5)$$

For a local optimum point we have

$$\operatorname{div} \vec{\psi} = \operatorname{div} \vec{H} = 0, \quad \frac{\partial^2 H_1}{\partial t \partial x} = \frac{\partial^2 H_2}{\partial x \partial t}. \quad (6.6)$$

7 Weak Solutions of Generalized PDEs of the HJB-type

Let $\kappa_{Q_0}(x, t)$ be the characteristic function of set Q_0 . Then in the non-smooth deterministic case we define generalised solutions of the PDE associated with the nonlinear control problem (3.1) – (3.3) as follows

Definition 7.1 A function $V(t, x) \in W_1^{1,1}(Q_0)$ is called the generalised solution of the non-smooth deterministic control problem if the integral identity

$$\iint_{Q_0} \left\{ S^x \otimes S^t \left[(1+v) \left(\frac{\partial V}{\partial x} + \frac{1}{v} \left(\frac{\partial V}{\partial t} + f_0 \right) \right) + \omega_1 \frac{\partial^2 V}{\partial t \partial x} + \omega_2 \frac{\partial^2 V}{\partial x \partial t} \right] \right\} \kappa_{Q_0}(x, t) dx dt = 0 \quad (7.1)$$

is satisfied.

Theorem 7.1 Let $f_0, g, f \in L^1(Q_0)$. If $\tilde{F} \equiv f_0(1+1/v) \in L_{\text{Lip}}^1(Q_0)$, then there exists a unique generalised solution of the non-smooth deterministic control problem. It has mixed derivatives if

$$\frac{\partial V}{\partial x} \in L^1(W_1^1(I), \Omega), \quad \frac{\partial V}{\partial t} \in L^1(I, W_1^1(\Omega)). \quad (7.2)$$

An analogous result was obtained in the stochastic case, where the generalised solution of (4.1)–(4.2) is defined as

Definition 7.2 A function $V(t, x) \in W_1^{1,1}(Q_0)$ is called the generalised solution of the stochastic control problem if the integral identity

$$\iint_{Q_0} \left\{ S^x \otimes S^t \left[(1+v) \left(\frac{\partial V}{\partial x} + \frac{1}{v} \left(\frac{\partial V}{\partial t} + f_0 \right) \right) + \right. \right. \quad (7.3)$$

$$\begin{aligned} & \omega_1 \frac{\partial^2 V}{\partial t \partial x} + \omega_2 \frac{\partial^2 V}{\partial x \partial t} - \\ & \left. \sigma_1 \frac{\partial^2 V}{\partial x^2} - \sigma_2 \frac{\partial^2 V}{\partial t^2} \right] \kappa_{Q_0}(x, t) dx dt = 0 \end{aligned} \quad (7.3)$$

is satisfied.

Theorem 7.2 Let $f_0, g, f \in L^1(Q_0)$. If $\tilde{F} \equiv f_0(1+1/v) \in L^1_{\text{Lip}}(Q_0)$, then there exists a unique generalized solution of the stochastic control problem. It has mixed and second order derivatives if

$$V(t, x) \in W_1^{2,2}(Q_0). \quad (7.4)$$

The application of the methodology described above to the analysis of the HJB-type equation in the deterministic case,

$$\frac{\partial V}{\partial x} + \frac{1}{v} \left(\frac{\partial V}{\partial t} + f_0 \right) = 0, \quad (7.5)$$

requires the locally relaxed Lipschitz condition:

$$\begin{aligned} \|S^x \otimes S^t (\tilde{F}(x, t, V') - \tilde{F}(x, t, V''))\|_{L^1(Q_0)} \leq \\ q \|S^x \otimes S^t (V' - V'')\|_{L^1(Q_0)}. \end{aligned} \quad (7.6)$$

The existence and uniqueness theorem for equation (7.5) has been constructively proved using Fejér's sums.

8 Conclusions and Future Directions

In this paper two major approaches for the solution of optimal control problems have been thoroughly analyzed in the framework of generalised solutions. We derived generalised partial differential equations for value/cost functionals in cases that are not covered by standard diffusion processes. Existence and uniqueness theory for these equations has been developed in nonreflexive Banach spaces. The approach proposed in this paper leads to the construction of approximating Markov chains for the derived equations [9]. A deeper investigation of the connection between non-smooth deterministic control problems and stochastic control problems in the class $W_1^{1,1}(Q_0)$ remains a challenging task for future work.

Acknowledgements

The work was supported by Australian Research Council Grant 17906. The author is grateful to Dr R. Watson for his helpful assistance at the final stage of preparation of this paper.

References

- [1] Barron, E. N., Jensen, R., "The Pontryagin maximum principle from dynamic programming and viscosity solutions to first-order partial differential equations", *Transactions of the American Mathematical Society*, **298**, No. 2, 635–641, (1986).
- [2] Beard, R. W., McLain, T.W., "Successive Galerkin approximation algorithms for nonlinear optimal and robust control", *Int. J. Control.*, **71**, No. 5, 717–743, (1998).
- [3] Boltyansky, V.G., Poznyak, A.S., "Robust maximum principle in minimax control", *Int. J. Control.*, **72**, No. 4, 305–314, (1999).
- [4] Borwein, J. M., Zhu, Q.J., "Variational analysis in nonreflexive spaces and applications to control problems with L^1 perturbations", *Nonlinear Analysis: TAM*, **28**, No. 5, 889–915, (1997).
- [5] Crandall, M.G., Lions, P.-L., "Viscosity solutions of Hamilton-Jacobi equations", *Transactions of the American Mathematical Society*, **277**, No. 1, 1–42, (1983).
- [6] Crandall, M.G., Lions, P. L., "Two approximations of solutions of Hamiltonian-Jacobi equations", *Mathematics of Computation* **43**, No. 167, 1–19, (1984).
- [7] Makila, P.M., "On robust control-oriented identification of discrete and continuous-time systems", *Int. J. Control.*, **70**, 319–335, (1998).
- [8] Melnik, R.V.N., "On Consistent Regularities of Control and Value Functions", *Numerical Functional Analysis and Optimization*, **18** (3&4), 401 – 426, (1997).
- [9] Melnik, R.V.N., "Dynamic System Evolution and Markov Chain Approximation", *Discrete Dynamics in NS, Gordon & Breach*, **2**, 7–39, (1998).
- [10] Osher, S., Shu, C-W., "High-order essentially nonoscillatory schemes for Hamilton-Jacobi equations", *SIAM J. Numer. Anal.*, **28**, No. 4, 907–922, (1991).
- [11] Partington, J.R., *Interpolation, Identification and Sampling*, Oxford University Press, (1997).
- [12] Subbotin, A. I., *Generalised Solutions of First-Order PDEs. The Dynamical Optimization Perspective*, Birkhäuser, (1995).
- [13] Vasiliev, F.P., *Numerical Methods for solving extremal problems*, Nauka, (1988).
- [14] Vidyasagar, M., "Optimal rejection of persistent bounded disturbances", *IEEE Trans. Automat. Control*, **31**, 527–534, (1986).
- [15] Wu, Q. et al "On construction of smooth Lyapunov functions for non-smooth systems", *Int. J. Control.*, **69**, 443–457, (1998).
- [16] Zhou, X.Y., "Maximum principle, dynamic programming, and their connection in deterministic control", *Journal of Optimization Theory and Applications*, **65**, 363–373, (1990).
- [17] Zhou, X.Y., "The connection between the maximum principle and dynamic programming in stochastic control", *Stochastics and Stochastics Report*, **31**, 1–13, (1990).

USQ



TOOWOOMBA

**MODELLING DYNAMICS OF SHAPE-
MEMORY-ALLOYS VIA COMPUTER
ALGEBRA**

R V N Melnik, A J Roberts, K A Thomas
Department of Mathematics & Computing, USQ
Faculty of Sciences Working Paper Series
SC-MC-9904
2 February 1999

THE UNIVERSITY OF
SOUTHERN
QUEENSLAND

Working Papers

FACULTY
OF
SCIENCES

**MODELLING DYNAMICS OF SHAPE-
MEMORY-ALLOYS VIA COMPUTER
ALGEBRA**

R V N Melnik, A J Roberts, K A Thomas
Department of Mathematics & Computing, USQ
Faculty of Sciences Working Paper Series
SC-MC-9904
2 February 1999

Modelling dynamics of shape-memory-alloys via computer algebra

R.V.N. Melnik^{*a}, A.J. Roberts^a, K.A. Thomas^a

^aDepartment of Mathematics and Computing,
University of Southern Queensland, QLD 4350, Australia

ABSTRACT

In this paper we present results on numerical studies of the martensitic-austenitic phase transition mechanism in a large shape-memory-alloy rod. Three groups of experiments are reported. They include results on stress- and temperature-induced phase transformations as well as the analysis of the hysteresis phenomenon. All computational experiments are presented for Cu-based structures.

Keywords: Shape memory alloys, phase transition, hysteresis, computer algebra

1. INTRODUCTION

The dynamics of martensitic-austenitic transformations has been investigated experimentally in a wide range of materials, in particular in metallic alloys such as NiTi, CuZnAl, CuZnGa, CuZn, NiAl, CuAlNi, and AgCd. Subject to appropriate thermomechanical conditions, these dynamics often exhibit a hysteretic behaviour accompanied by shape-memory effects. For example, if an unstressed shape-memory-alloy wire has been stretched at low temperature, it can be returned to its initial condition upon heating. Upon cooling it can be again returned to its stretched form. In other words, the materials under consideration can be "imprinted" with a shape that they "remember". Not surprisingly these effects have a wide variety of applications ranging from heat engines and different types of actuators to robotics, oceanographic and aerospace industries.

Over recent years, the interest in modelling the dynamics of shape-memory-alloys has been dramatically increased. This includes experimental,¹ theoretical,²⁻⁴ and computational works.⁵⁻⁸ Current and emerging applications of shape memory alloys require a deeper understanding of structural phase transitions in solids and provide new challenging problems in applied mathematics.

Many smart materials display a strong dependence of load deformation upon temperature. Therefore, in order to adequately model the dynamics of these materials it is important to account for the coupling of stresses, deformation gradients and displacements to the thermal field. Such a coupling is critical in the description of many phenomena that are becoming increasingly important in a wide range of applications of smart materials and structures. The strong nonlinear coupling between thermal and mechanical fields provides an important key to a better understanding of hysteresis-type phenomena. Since these phenomena and the associated shape-memory effects are very difficult to control experimentally, the tools of mathematical modelling and computational experiment play an increasing role in the investigation of thermally and mechanically induced hysteresis in viscoelastic and pseudoplastic materials.

Our main focus in this paper is the adequate description of thermomechanical behaviour of a large shape-memory-alloy rod in the martensitic-austenitic phase transition. Since the description of the mechanism of this transition requires nonconvex free energy functions, this leads to serious mathematical modelling difficulties even in low-dimensional cases.

This paper is a development of our earlier paper,⁸ where we derived the two mathematical models used here. In this paper we present three groups of experiments. In the first group of experiments we investigate the mechanical control of phase transitions in shape-memory-alloys. While in Ref. 8 we were interested in phase transitions activated by distributed loading, in this paper we are interested in the dynamics of shape-memory alloys when phase transitions are activated by applied stresses at the body boundary. In the second group of experiments we investigate thermally

^{*}Correspondence: Email: melnik@usq.edu.au; WWW: <http://www.sci.usq.edu.au/staff/melnik/>; Telephone: +61 7 46312632; Fax: +61 7 46311775

induced phase transitions and demonstrate the combined effect of distributed heating and boundary stresses on the development of phase transformations. Finally, we present computational results on and the analysis of hysteresis phenomena for different types of thermomechanical conditions.

We organised the rest of this paper as follows.

- Section 2 gives an overview of the mathematical models used in this paper.
- In Section 3 we provide the reader with the main ideas of our numerical approximations.
- The main part of the paper is Section 4, where we report computational results from the investigation of martensitic-austenitic phase transitions and thermally induced hystereses.
- Finally, in Section 5 we discuss future directions of the presented work.

2. MATHEMATICAL MODELS FOR SHAPE-MEMORY-ALLOY DYNAMICS

The starting point of our present deliberation is this system derived in Ref. 8:

$$\begin{cases} C_v \left[\frac{\partial \theta}{\partial t} + \tau_0 \frac{\partial^2 \theta}{\partial t^2} \right] - k_1 \left[\theta \frac{\partial u}{\partial x} \frac{\partial^2 u}{\partial t \partial x} + \tau_0 \frac{\partial}{\partial t} \left(\theta \frac{\partial u}{\partial x} \frac{\partial^2 u}{\partial t \partial x} \right) \right] - \mu \left[\left(\frac{\partial^2 u}{\partial t \partial x} \right)^2 + \right. \\ \left. \tau_0 \frac{\partial}{\partial t} \left(\frac{\partial^2 u}{\partial t \partial x} \right)^2 \right] - \nu \left[\frac{\partial \theta}{\partial t} \frac{\partial^2 u}{\partial t \partial x} + \tau_0 \frac{\partial}{\partial t} \left(\frac{\partial \theta}{\partial t} \frac{\partial^2 u}{\partial t \partial x} \right) \right] - \frac{\partial}{\partial x} \left(k \frac{\partial \theta}{\partial x} \right) = G, \\ \rho \frac{\partial^2 u}{\partial t^2} - \frac{\partial}{\partial x} \left[k_1 \frac{\partial u}{\partial x} (\theta - \theta_1) - k_2 \left(\frac{\partial u}{\partial x} \right)^3 + k_3 \left(\frac{\partial u}{\partial x} \right)^5 \right] - \mu \frac{\partial^3 u}{\partial x^2 \partial t} - \nu \frac{\partial^2 \theta}{\partial x \partial t} = F, \end{cases} \quad (1)$$

where u is the displacement field, θ is the temperature field, τ_0 is the thermal relaxation time, k is the thermal conductivity of the material, C_v is the specific heat constant of the material, μ and ν are material-specific coefficients that characterise the dependency of the stress on the rate of the deformation gradient and temperature respectively, ρ is the density of the material, θ_1 is a positive constant that characterises a critical temperature of the material, and k_i , $i = 1, 2, 3$ are material-specific constants that characterise the material's free energy. The right-hand side parts of system (1), F and G , represent the distributed mechanical and thermal loadings of the body.

System (1) is completed by appropriate initial and boundary conditions and has to be solved with respect to (u, θ) in the spatial-temporal region $Q = \{(x, t) : 0 \leq x \leq L, 0 \leq t \leq T_f\}$, where L is the length of the structure and T_f is the required time of observation. The initial conditions for the model (1) are taken in the following form

$$u(x, 0) = u^0(x), \quad v(x, 0) = \frac{\partial u}{\partial t}(x, 0) = u^1(x), \quad \theta(x, 0) = \theta^0(x), \quad \frac{\partial \theta}{\partial t}(x, 0) = \theta^1(x), \quad (2)$$

with specified functions $u^0, u^1, \theta^0, \theta^1$. Boundary conditions are problem-specific.⁸ In all experiments reported in this paper mechanical boundary conditions are either specified stress or specified displacement:

$$s(0, t) = s_1(t), \quad s(L, t) = s_2(t), \quad \text{or} \quad u(0, t) = u_1(t), \quad u(L, t) = u_2(t); \quad (3)$$

thermal boundary conditions are those of specified heat flux

$$\frac{\partial \theta}{\partial x}(0, t) = \bar{\theta}_1(t), \quad \frac{\partial \theta}{\partial x}(L, t) = \bar{\theta}_2(t), \quad (4)$$

where functions $s_i(t)$ (or $u_i(t)$) and $\bar{\theta}_i(t)$, $i = 1, 2$ are given.

2.1. The free energy function as a coupling mechanism between mechanical and thermal fields

Temperature plays a critical role in the dynamics of shape-memory-alloys. Under different thermal conditions such materials can exhibit qualitatively different behaviour. For example, at low temperature it is reasonable to expect that shape-memory-alloy materials exhibit *ferroelastic behaviour*. At intermediate temperature they may behave like a *pseudoelastic* material, while at high temperature they behave similar to *elastic* materials.⁹ From the mathematical point of view, such a wide range of qualitatively different behaviours is adequately described only if the thermal and mechanical fields are considered together. This coupling is determined by the free energy function. A "slight" change in this function (such as the account for [or omitting of] the viscous or coupled stresses, thermal memory terms, etc²) may require completely different mathematical arguments in the analysis of the well-posedness of the model and in the construction of numerical schemes for its solution.

The coupling of thermal and mechanical fields is a key component of the models we deal with in this paper. The description of phase transformations with the model (1)–(4) is based on a non-convex free energy function, which leads to a non-monotone load-deformation curve.¹⁰ This idea, often attributed to van der Waals, was further developed theoretically by Landau and in the context of shape-memory-alloys was pioneered by Falk.⁹ The model (1)–(4) incorporates the Helmholtz free energy in the Landau-Devonshire form. In a number of experiments we have also used a modified model (1)–(4) with the Landau-Devonshire-Ginzburg form for the free energy function which assumes the dependency of the free energy on a change of the curvature ϵ_x of the metallic lattice. The form of this dependency determines the coupled stress $\zeta = \zeta(\epsilon_x)$, an extra term in the free energy function, that typically takes the smoothing role of the viscous stress and simplifies the analysis of the mathematical model. The definition of this term varies in the literature with the most common taken to be a quadratic dependency $\gamma/2\epsilon_x^2$, known as the Ginzburg term (γ is the Ginzburg coefficient). This term leads to the appearance of an additional term, γu_{xxxx} , in the first equation of system (1). As we noted in Ref. 8, with reported values of the Ginzburg coefficient ($\gamma \sim 10^{-10} - 10^{-12}$) this term showed little influence on the dynamics of shape-memory alloys in the group of experiments we performed.

2.2. Low-Dimensional Modelling of Shape-Memory-Alloy Dynamics

The properties of shape-memory alloys may strongly depend on chemical composition and on the processing of the material. This often leads to difficulties in comparing different physical experiments. From the mathematical modelling point of view, these difficulties may show themselves in the choice of the free energy function, especially in the higher dimensional case. Indeed, using the analogy with the Landau theory one needs to determine at least 32 material parameters in order to describe thermomechanical behaviour of shape-memory-alloys in 3D.^{11,8}

Therefore, some *physically justified assumptions* may be very useful in the mathematical modelling of shape memory alloys. Using such assumptions one can reduce the number of required parameters for Cu-based shape memory alloys to 10. These parameters for $\text{Cu}_{14}\text{Al}_3\text{Ni}_{83}$ (see Ref. 11) have been used in the the construction of a low dimensional model for shape-memory-alloy dynamics in our recent paper.⁸ Using the computer algebra package REDUCE such a model has been derived from the 3D Falk-Konopka model¹¹ for modelling a shape-memory-alloy slab that has a very large extent in the x -direction compared to its thickness ($2b$) in the y -direction ($-b < y < b$). We have assumed that the essential dynamical behaviour of the slab can be determined by a subset of all possible modes. This is a key idea of a quite general methodology arising from centre manifold theory.^{12,13}

The model has been constructed with respect to the amplitudes of the critical modes, U_i , V_i ($i = 1, 2$) and θ' under the assumption that there exists a low-dimensional invariant (slow) manifold upon which these amplitudes evolve slowly. These amplitudes have been chosen as y -averages of u_i , $i = 1, 2$ (displacements in x - and y - directions respectively), v_i , $i = 1, 2$ (velocities in x - and y - directions respectively) and $\theta' = \theta - \theta_0$ with θ_0 taken 300°K . The

model for the longitudinal unforced dynamics of Cu₁₄Al₃Ni₈₃ shape-memory-alloy slab has the following form:

$$\left\{ \begin{array}{lcl} \rho \frac{\partial V_1}{\partial t} & = & 2.97e5 U_{1xx} + 8.03e5 b^2 U_{1xxxx} \\ & & + \partial_x [(922 \Theta' - 0.0145 \Theta'^2) U_{1x} - (4.28e9 - 1.31e7 \Theta') U_{1x}^3 + 7.12e11 U_{1x}^5 \\ & & + (2820 - 8.80 \Theta') b^2 V_{1x}^2 U_{1x} + 1.24 b^4 V_{1x}^4 U_{1x} - 5.42e4 b^2 V_{1x}^2 U_{1x}^3], \\ \rho \frac{\partial V_2}{\partial t} & = & -9.91e5 b^2 U_{2xxxx}, \\ C_v \frac{\partial \Theta'}{\partial t} & = & \kappa \Theta'_{xx} + (2.77e5 + 914 \Theta' - 9.25 \Theta'^2) U_{1x} V_{1x} \\ & & + (3.94e9 + 1.26e7 \Theta') V_{1x} U_{1x}^3 - (57.3 + 0.0117 \Theta') b^2 V_{1x}^3 U_{1x} \\ & & + 1.68e12 V_{1x} U_{1x}^5 - 1.58e6 b^2 V_{1x}^3 U_{1x}^3 - 0.0203 b^4 V_{1x}^5 U_{1x} \\ & & + 1.63e4 b^2 U_{1xx} V_{1xx} + 9.22e4 b^2 U_{2xx} V_{2xx} + \partial_x^2 [-8151 b^2 U_{1x} V_{1x}]. \end{array} \right. \quad (5)$$

Some features of this model are similar to model (1). For example, in the second line of the first equation recognise the temperature dependent stress-strain relation of the shape memory alloy. The model (5) is exact up to the eighth order with respect to the norm $\|U_x\| + \|V_x\|$ and the fourth order with respect to two other small parameters chosen in our expansion, ∂_x and $\|\Theta'\|$. Displacements and velocities on the slow manifold are approximately determined using the solution of system (5) as follows (see details in Ref. 8)

$$u_1 \approx U_1 - Y b U_{2x} + 0.15(3Y^2 - 1)b^2 U_{1xx}, \quad (6)$$

$$u_2 \approx U_2 - (0.9 - 3.05e-5 \Theta') Y b U_{1x} + 0.15(3Y^2 - 1)b^2 U_{2xx} \\ - 141 Y b U_{1x}^3 + 1.00e-4(3Y - Y^3)b^3 V_{1x}^2 U_{1x}, \quad (7)$$

$$\theta \approx 300 + \Theta' - 2.43e6(3Y - Y^3)b^3(V_{1x} U_{2xx} + U_{1x} V_{2xx}) \\ - 25.1(7 - 30Y^2 + 15Y^4)V_{1x}^3 U_{1x}, \quad (8)$$

where $Y = y/b$ is the scaled transverse coordinate. Unfortunately, as stated in Ref. 11 there has been few experiments on martensitic elastic moduli of shape memory alloys. Perhaps, one of the most widely cited materials is Cu₁₄Al₃Ni₈₃ (see Ref. 11). At the time of writing of this paper, we are not aware of parameters for this material for the 1D model. This fact makes it difficult to compare results obtained with low-dimensional and one-dimensional models such as (5)–(8) and (1)–(4). At present, we are using the available parameters for Au₂₃Cu₃₀Zn₄₇ (see Ref. 8).

Computer algebra provides a powerful tool in the analysis of shape-memory-alloy dynamics. Indeed, using computer algebra, the model (5)–(8) can be modified to take into account mechanical and/or thermal forcing as well as higher order terms. Currently model (5)–(8) is under our scrutiny.

3. NUMERICAL APPROXIMATIONS

The system (1) is a strongly nonlinear system of partial differential equations that couples hyperbolic and parabolic modes of the dynamics in a unified whole. In the general case, the coupling pattern between thermal and mechanical fields is difficult to analyse. However, if we simplify the system by assuming $\tau_0 = \mu = \nu = 0$, it can be seen that the main terms responsible for the coupling phenomenon are

$$k_1 \theta \frac{\partial u}{\partial x} \frac{\partial v}{\partial x} \quad \text{and} \quad \frac{\partial}{\partial x} \left(k_1 \frac{\partial u}{\partial x} (\theta - \theta_1) \right) \quad (9)$$

in the first and the second equation respectively. As shall be seen in Section 4, these terms provide the basis for the analysis of the phase transition mechanism.

Several numerical procedures have been reported in the literature for the solution of systems of PDEs describing shape-memory-alloy dynamics (see, for example, Ref. 6,14 and references therein). Our approach is different from those previously reported. The main idea of our approach is a transformation of the problem (1)–(4) into a system of differential-algebraic equations with respect to (u, v, θ, s) . Then we solve this system using second-order accurate spatial differences on staggered grids. The developed MATLAB code is simple, robust and easy to implement.

Taking into account the above assumptions the transformed system for all computational experiments described in Section 4 has the following form

$$\begin{cases} \frac{\partial u}{\partial t} = v, \\ \rho \frac{\partial v}{\partial t} = \frac{\partial s}{\partial x} + F, \\ C_v \frac{\partial \theta}{\partial t} = k \frac{\partial^2 \theta}{\partial x^2} + k_1 \theta \frac{\partial u}{\partial x} \frac{\partial v}{\partial x} + G, \\ s = k_1 (\theta - \theta_1) \frac{\partial u}{\partial x} - k_2 \left(\frac{\partial u}{\partial x} \right)^3 + k_3 \left(\frac{\partial u}{\partial x} \right)^5. \end{cases} \quad (10)$$

The term $\frac{\partial u}{\partial x} = \epsilon$ in (10) is the linearised strain that plays the role of the order parameter in the Landau theory.

The initial and boundary conditions for this model are problem-specific and will be defined in the next section. All experiments reported here were performed for a $\text{Au}_{23}\text{Cu}_{30}\text{Zn}_{47}$ rod of the length $L = 1\text{cm}$. For the $\text{Au}_{23}\text{Cu}_{30}\text{Zn}_{47}$ material all necessary parameters were first specified by Y. Murakami (see ref. in Falk⁹). In the context of system (10) we use (see also Ref. 6,8)

$$\begin{aligned} k &= 1.9 \times 10^{-2} \text{ cmg}/(\text{ms}^3\text{K}), & \rho &= 11.1 \text{ g/cm}^3, & C_v &= 29 \text{ g}/(\text{ms}^2\text{cmK}), & \theta_1 &= 208\text{K}, \\ k_1 &= 480 \text{ g}/(\text{ms}^2\text{cmK}), & k_2 &= 6 \times 10^8 \text{ g}/(\text{ms}^2\text{cmK}), & k_3 &= 4.5 \times 10^8 \text{ g}/(\text{ms}^2\text{cmK}). \end{aligned}$$

4. COMPUTATIONAL EXPERIMENTS

Due to a wide range of industrial applications and associated challenging applied mathematics problems, the control of phase transitions in shape memory alloys has recently become a topic of considerable interest.³ In this section we investigate different options for the control of shape memory alloys, including stress-boundary control, distributed heating and their combination. Special attention is given to the computational analysis of hysteresis.

4.1. Stress Induced Phase Transitions

The stress-induced phase transition exhibits a hysteresis, provided that we have thermodynamic barriers to prevent an equilibrium phase transition.⁹ In the following two experiments the role of these barriers is played by the boundary stress chosen as the main control variable.

Experiment 1.1. First, in the spirit of Ref. 7, let us consider the following initial conditions

$$u(x, 0) = \epsilon_0 x, \quad v(x, 0) = 0, \quad \theta(x, 0) = 230. \quad (11)$$

Choosing $\epsilon_0 = 0.106051$ the rod is initially in martensitic phase M_+ . For the given initial temperature this state is stable and we need to change thermomechanical conditions of the rod to induce a phase transition (for such a low temperature the same is also true for M_- martensitic phases).

We assume no distributed loading (i.e. $F = G = 0$). Instead, for the first 6 ms we load the rod at the boundaries (which are assumed to be thermally insulated) with the compressive load $-7000 \sin^3(\pi t/6) \text{ g}/(\text{ms}^3\text{cm})$. Then, for the next 6 ms we do not apply any force, and for the next 6 ms after that we apply a tensile load $7000 \sin^3(\pi t/6) \text{ g}/(\text{ms}^3\text{cm})$. Therefore, we have the following boundary conditions on $x = 0$ and L :

$$\frac{\partial \theta}{\partial x} = 0, \quad s = \begin{cases} -7000 \sin^3(\pi t/6), & 0 \leq t \leq 6, \\ 7000 \sin^3(\pi t/6), & 12 \leq t \leq 18, \\ 0, & \text{otherwise.} \end{cases} \quad (12)$$

During the initial period, observe the phase transformation $M_+ \rightarrow M_-$ in Fig. 1. Then, observe the appearance of two regions with a slight increase/decrease in displacement (upward and downward "humps"). These regions demonstrate *thermomechanical coupling* effects between the two phases in the period when the temperature pattern

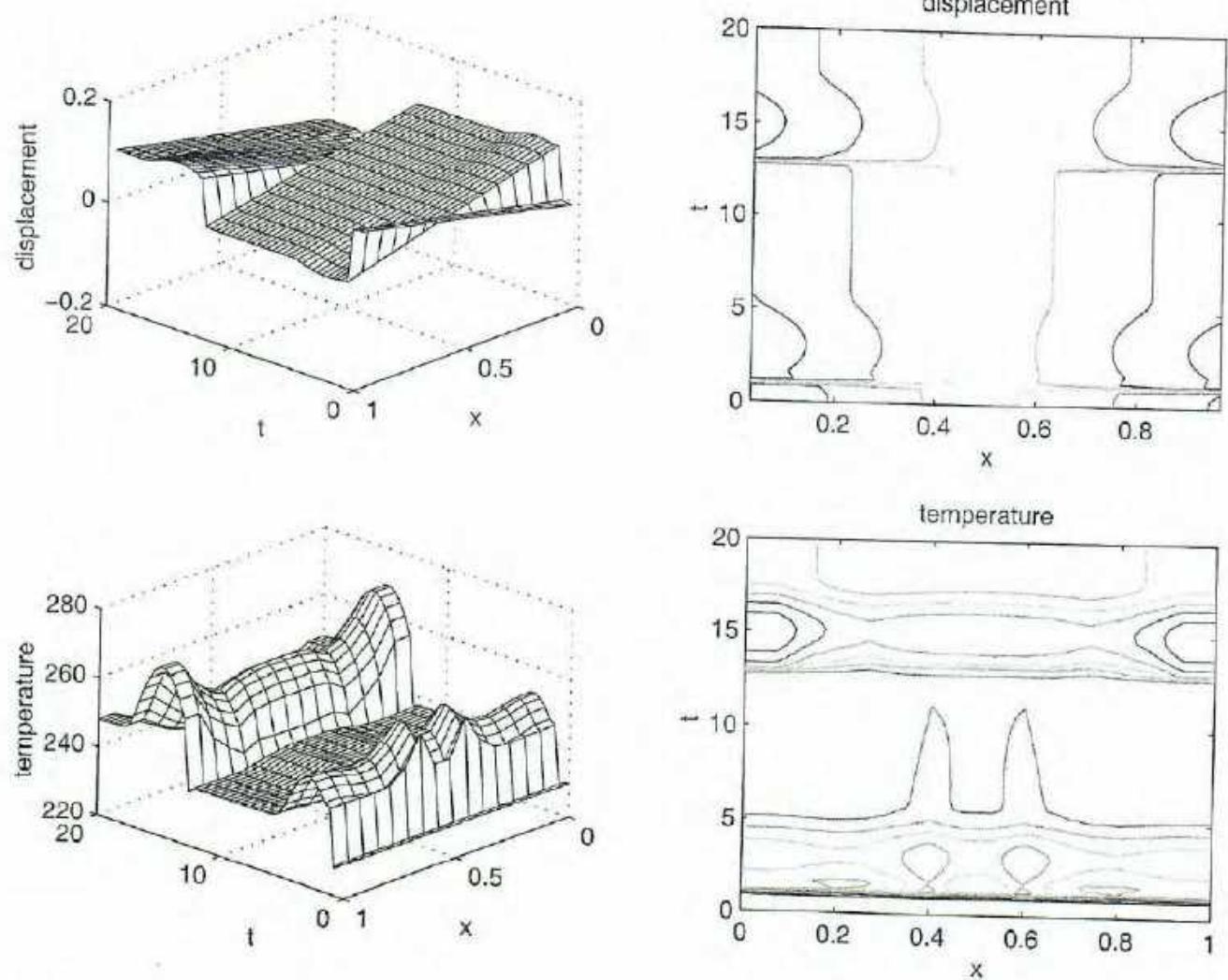


Figure 1. Stress induced phase transition: the tensile load exceeds the yield limit.

changes (see temperature plots in Fig. 1). After a while, one observes that these regions vanish and we have only the M_- phase which is in stable equilibrium. Finally, observe a reverse phase transformation $M_- \rightarrow M_+$ (due to the tensile load) according to a pattern which is analogous to that described above.

The behaviour of this type is typical for ferro-elastic materials. When subjected to low initial temperatures, shape memory alloys do not produce an intermediate austenite phase under the above thermomechanical conditions.

Experiment 1.2. In the previous example we saw that after the transformation $M_+ \rightarrow M_-$ takes place, the reverse transformation $M_- \rightarrow M_+$ is possible under fairly high tensile loading exceeding the yield limit. If the last condition is not satisfied, the rod would remain in the M_- phase. This is demonstrated by the next experiment, where the tensile loading was set 10 times lower than in the previous example:

$$s = \begin{cases} -7000 \sin^3(\pi t/6), & 0 \leq t \leq 6, \\ 700 \sin^3(\pi t/6), & 12 \leq t \leq 18, \\ 0, & \text{otherwise.} \end{cases} \quad (13)$$

All other conditions remained the same as in Experiment 1.1. The effect of thermomechanical coupling (after the transition $M_+ \rightarrow M_-$) is observed only for a short period of time. It results in small perturbations visible on Fig. 2 as the regions with two "humps". After that the rod returns to the M_- stable equilibrium.

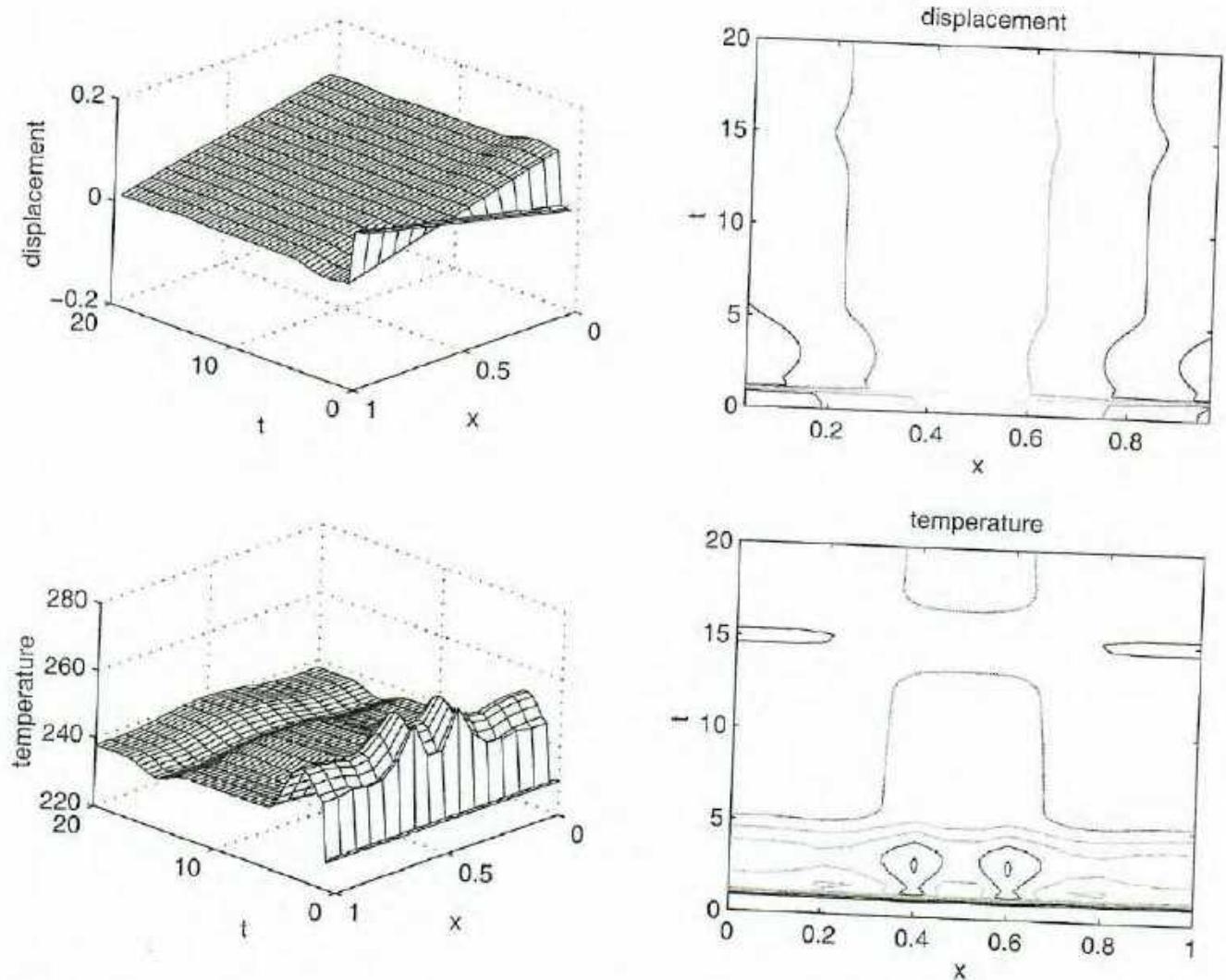


Figure 2. Absence of thermodynamic barriers to prevent an equilibrium phase transition; the tensile load is less than the yield limit.

The situation will not qualitatively change if we keep a constant load (say, $s = 100$) for all the times when the compressive/tensile load at the boundary is absent, i.e. when

$$s = \begin{cases} -7000 \sin^3(\pi t/6), & 0 \leq t \leq 6, \\ 700 \sin^3(\pi t/6), & 12 \leq t \leq 18, \\ 100, & \text{otherwise.} \end{cases} \quad (14)$$

The almost-linear behaviour of displacements on the second stage of this experiment suggests that the rod exhibits elastic properties under the given thermomechanical conditions.

With spatial and temporal steps 0.66 cm and $9.26 \times 10^{-4}\text{ ms}$ respectively, experiments 1.1 and 1.2 take on average 10–11 minutes to complete on a Digital Alpha 255 station (300 MHz).

4.2. Temperature Induced Phase Transitions

The purely mechanical control of phase transitions discussed in Section 4.1 may not be efficient. In the next experiment we show how to control this process by temperature. More precisely, we explore the range of temperature

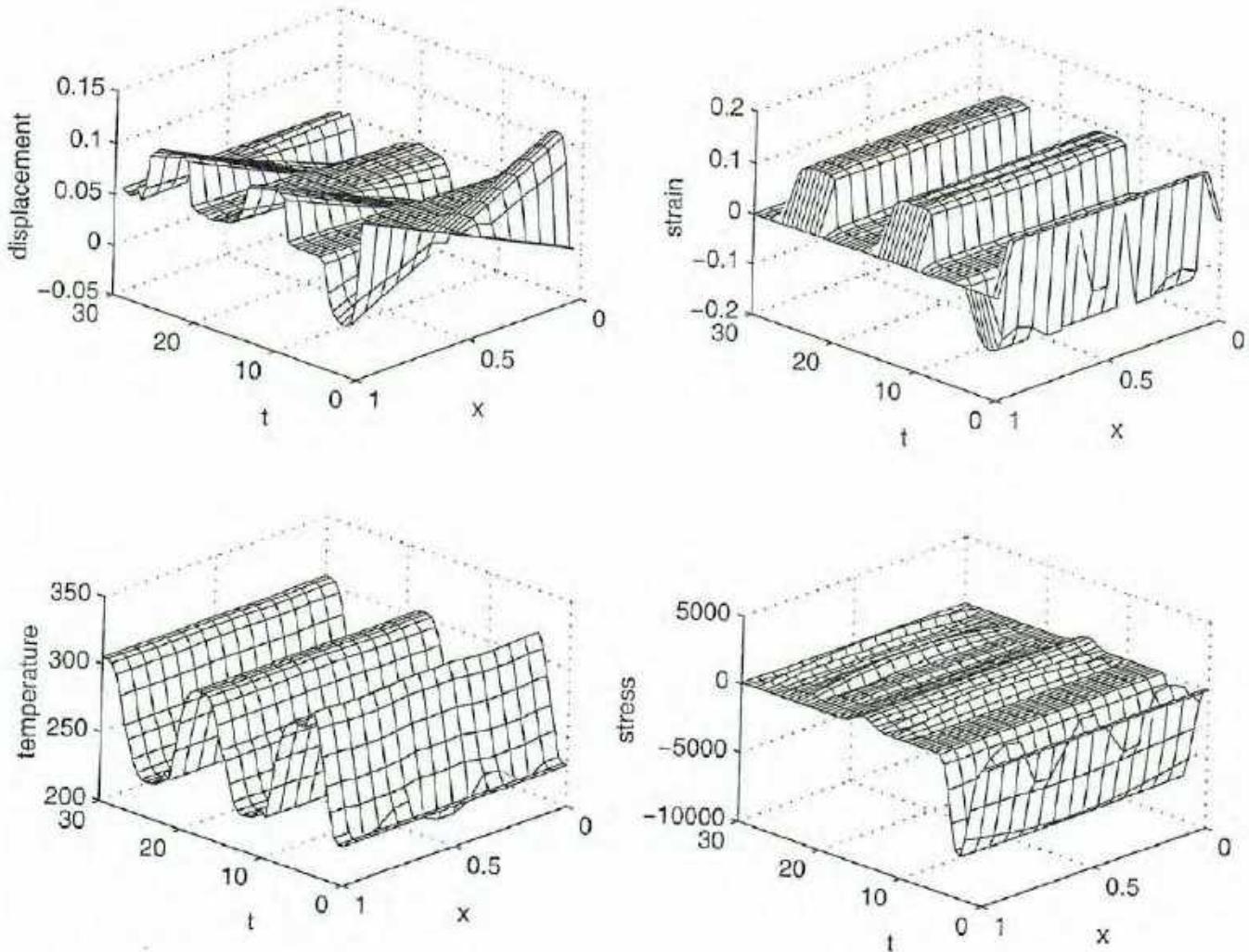


Figure 3. Temperature controls thermodynamic barriers allowing phase transitions even if the tensile load is less than the yield limit.

for which on cooling the formation of martensite starts; and conversely, the range of temperature, for which on heating, the formation of austenite starts.

The driving force for this type of transition is the difference between the free energies of both phases. We control this difference by controlling the distributed heating/cooling pattern.

Experiment 2.1. Consider our last experiment where we were unable to produce a phase transition with the given loading pattern. Let us keep the stress on the boundary such that (14) is satisfied and all other conditions, except for the distributed thermal loading, remain the same as in Experiment 1.2. The distributed heating/cooling used is

$$G = 375 \sin^2(\pi t/6) \text{ g}/(\text{ms}^3 \text{cm}). \quad (15)$$

Then, after the initial switch $M_+ \rightarrow M_-$ (due to insufficiently low initial temperature), we observe phase transitions (see Fig. 3)

$$M_- \rightarrow A, \text{ then } A \rightarrow M_+, \quad M_+ \rightarrow A \quad \text{etc.} \quad (16)$$

The hyperbolic features of these transitions are demonstrated by the stress plot presented on Fig. 3.

This experiment shows that in those cases when the mechanical load alone is not sufficient to induce the phase transition in the rod, we can effectively use distributed heating/cooling to control the phase transition process.

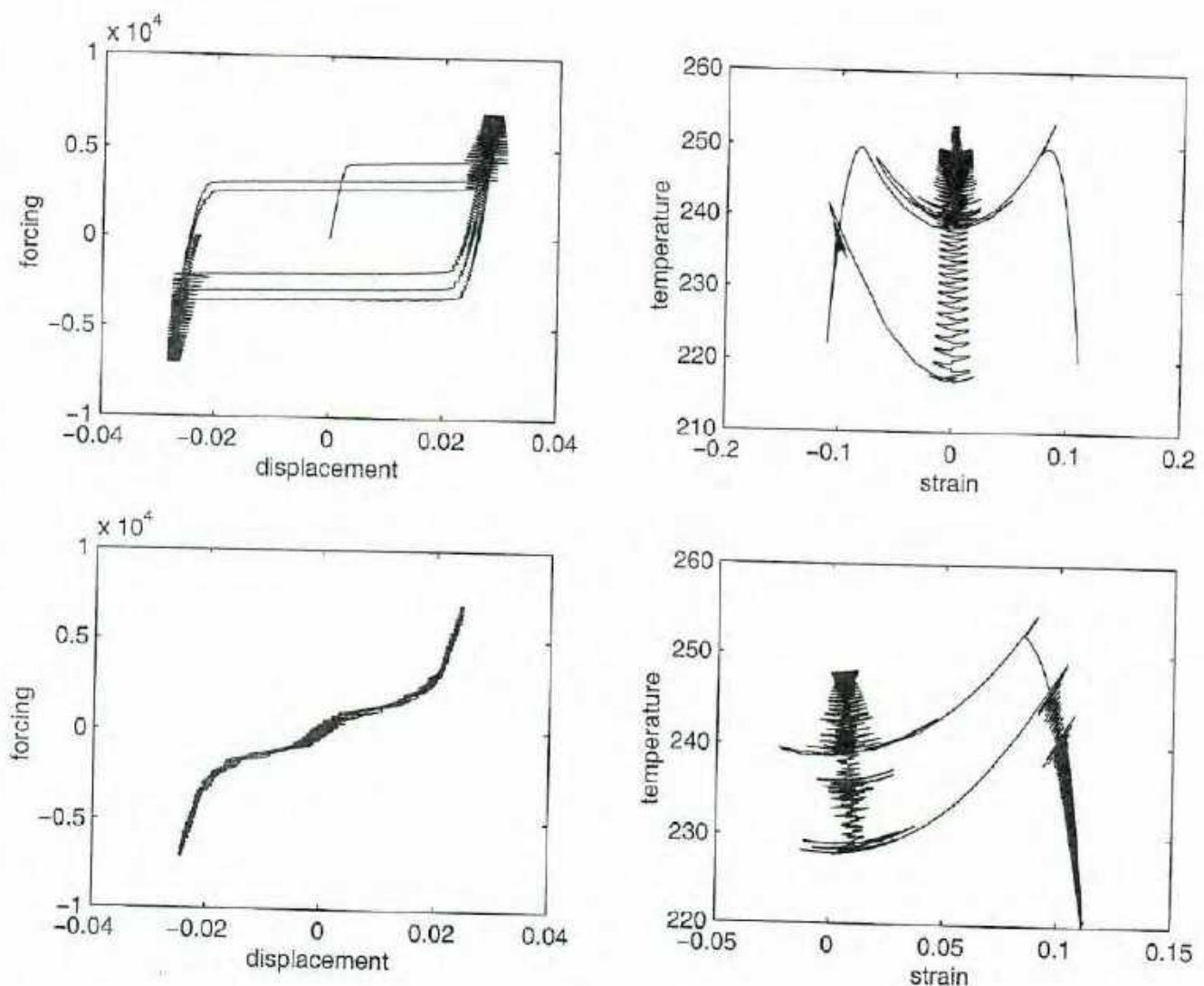


Figure 4. Mechanically and thermally induced hystereses.

4.3. Hysteresis Analysis

The non-convexity and the non-linear thermal dependency of the free energy function, used in the derivation of mathematical models such as (1)–(4) and (5)–(8), play key roles in the understanding of hysteresis-type phenomena. Indeed, in the general case these characteristics imply non-monotone stress-stress relationships (or load-deformation diagrams) which depend on the temperature change in a complicated nonlinear manner. We explore such situations below.

Experiment 3.1. Again consider the case of mechanical loading in the low-temperature regime. However, in contrast to Experiment 1.1, we start with two M_+ and one M_- martensites, so that the initial conditions for this experiment are

$$\theta^0 = 220, \quad u^0 = \begin{cases} 0.11869x, & 0 \leq x \leq 0.25, \\ 0.11869(0.5 - x), & 0.25 \leq x \leq 0.75, \\ 0.11869(x - 1), & 0.75 \leq x \leq 1, \end{cases} \quad v^0 \equiv u^1 = 0. \quad (17)$$

Another difference from previously considered examples is the pinned mechanical boundary conditions

$$u = 0, \quad \frac{\partial \theta}{\partial x} = 0. \quad (18)$$

In the absence of thermal loading ($G = 0$) we assume the following time-dependent distributed mechanical loading

$$F = 7000 \sin^3\left(\frac{\pi t}{2}\right) \text{ g}/(\text{ms}^3\text{cm}). \quad (19)$$

Since under these thermomechanical conditions shape-memory-alloys behave like a ferroelastic material, one may expect a hysteresis loop to be observed.^{15,5} This phenomenon is clearly demonstrated by the upper left plot on Fig. 4.

Experiment 3.2. Under intermediate-temperature conditions, shape-memory alloys behave like pseudoelastic materials. In this case, in spite of the difference in loading/unloading stress-strain curves, the mechanical loading does not lead to a residual strain.⁹

In our next experiment we start from the martensitic state M_+ given by the following initial conditions:

$$u^0 = 0.11869x, \quad v^0 = 0, \quad \theta^0 = 270. \quad (20)$$

The thermal and mechanical distributed loading are the same as in Experiment 3.1, but we also apply stress at the thermally insulated boundaries according to the following rule

$$s = \begin{cases} -1000 \sin^3(\pi t/6), & 0 \leq t \leq 6, \\ 0, & \text{otherwise}, \end{cases} \quad \frac{\partial \theta}{\partial x} = 0. \quad (21)$$

The given temperature is on the border of the "pseudoelastic" range and the two symmetric loops, that are typical for these thermomechanical conditions, are very small (see the lower-left plot on Fig. 4). Further increase in temperature leads to the complete disappearance of hysteresis since in this case shape-memory-alloys start to behave like elastic materials.

The concluding two experiments are aimed at the description of hystereses in thermally-induced phase transformations.

Experiment 3.3. Two symmetric martensites are taken as the initial state for the next experiment, namely

$$u^0 = \begin{cases} 0.11869x, & 0 \leq x \leq 0.5, \\ 0.11869(1-x), & 0.5 \leq x \leq 1, \end{cases} \quad v^0 = 0, \quad \theta^0 = 220. \quad (22)$$

In this experiment we assume constant distributed loading of $F = 500 \text{ g}/(\text{ms}^3\text{cm})$ and the following thermal distributed loading

$$G = \frac{375}{2} \pi \sin^3\left(\frac{\pi t}{6}\right) \text{ g}/(\text{ms}^3\text{cm}). \quad (23)$$

On the boundary we assume that

$$u = 0, \quad \frac{\partial \theta}{\partial x} = 0. \quad (24)$$

The temperature-strain plot, presented on the right-upper plot of Fig. 4, demonstrates temperature-induced transformations between austenite and martensites. See that these transformations are approximately symmetric with respect to the sign of strain.

Experiment 3.4. Finally, we assume no mechanical distributed loading ($F = 0$). We start from an M_+ martensitic state defined by the following initial conditions:

$$u^0 = 0.11869x, \quad v^0 = 0, \quad \theta^0 = 220. \quad (25)$$

Boundary conditions are those of stress-free, insulated ends:

$$s = 0, \quad \frac{\partial \theta}{\partial x} = 0, \quad (26)$$

while the distributed thermal loading is assumed to follow the same pattern as in Experiment 3.3. The lower-right plot of Fig. 4 gives the temperature-strain curve of the phase transformation between M_+ and A states.

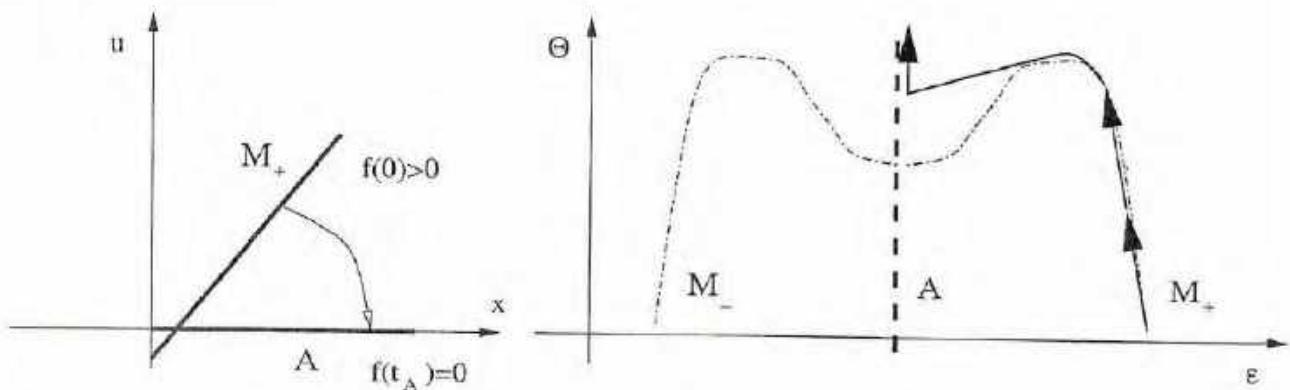


Figure 5. Schematic representations of thermally induced hystereses.

With spatial and temporal steps 0.0625 cm and 6.67×10^{-4} ms respectively, the last 4 experiments take on average 33–36 minutes to complete on a Digital Alpha 255 station.

The overall shape of the temperature-strain curves in Experiment 3.3 and 3.4 are explained using a simple approximate analysis that follows. Let us assume that the strain varies in time but is more-or-less uniform over the rod: $\frac{\partial u}{\partial x} \approx \zeta f(t)$ with a certain constant coefficient ζ (see Fig. 5 for a schematic representations of the transformation $M_+ \rightarrow A$). Then, integrating this relationship (assuming the mean displacement to be zero) we get

$$u = \zeta x f(t) \quad \text{and, hence,} \quad v = \zeta x \frac{\partial f}{\partial t}. \quad (27)$$

After differentiating (27) we have that $\frac{\partial v}{\partial x} = \zeta \frac{\partial f}{\partial t}$.

On the other hand, in the absence of diffusion ($k = 0$) when $\tau_0 = 0$ and $\mu = \nu = 0$, the thermal equation of system (1) is

$$\frac{\partial \theta}{\partial t} = k_c \theta \frac{\partial u}{\partial x} \frac{\partial v}{\partial x}, \quad \text{where} \quad k_c = k_1/C_v. \quad (28)$$

Using the above representations for $\frac{\partial u}{\partial x}$ and $\frac{\partial v}{\partial x}$ in terms of the function f , from (28) we get that

$$\frac{1}{\theta} \frac{\partial \theta}{\partial t} = k_c \zeta^2 f \frac{\partial f}{\partial t}. \quad (29)$$

Hence, taking into account that $f = \frac{1}{\zeta} \frac{\partial u}{\partial x}$, an approximation to the strain-temperature relationship is derived directly from (29) as

$$\ln \frac{\theta}{\theta_c} = \frac{k_c}{2} \zeta^2 f^2 \quad \text{or} \quad \ln \frac{\theta}{\theta_c} = \frac{k_c}{2} \left(\frac{\partial u}{\partial x} \right)^2, \quad (30)$$

where θ_c is a constant. The last relationship confirms the parabolic shape of the temperature-strain curves in the transitions from martensites to austenites as depicted on the right-hand side plots of Fig. 4.

5. FUTURE DIRECTIONS

Physical experiments suggest the possible existence of non-equilibrium states inside of hysteresis loops.¹⁵ Therefore, the geometry of hysteresis loops as a function of temperature requires further investigations. Mathematical modelling and computational experiments in such investigations provide a very useful and powerful tool. For a deeper understanding of hysteresis phenomenon in shape-memory-alloys further computational experiments are needed. An important direction of these experiments is the comparison of results obtained with different mathematical models such as (1)–(4) and (5)–(8).

Another aspect of future work follows from the evidence that the width of hysteresis may often be determined by an additional term which is responsible for the interfacial energy and when this term vanishes the phase transition may occur reversibly.¹⁰ Mathematical models that are derived using free energy functions that incorporate interfacial energy contributions and take into account the influence of mechanical and thermal dissipations (such as the latent heat) may prove to be helpful.

Strictly speaking the models presented in this paper are applicable only for single crystals. A much more complex task is to describe the dynamic of composite heterogeneous materials. Such materials often arise in structural engineering when dealing, for example, with polymer based composites, in biomechanics, and in food industries. Some composites such as fiber-reinforced polymers, are formed as a mixture of elastic and viscoelastic materials.¹⁶ Mathematical modelling of the dynamics of such materials, including their thermoviscoelastic contacts, present a challenge for future work.

ACKNOWLEDGMENTS

This work was supported by USQ-PTRP Grant 179452 and by Australian Research Council Grant 179406.

REFERENCES

1. H. Benzaoui, C. L'Excellent, N. Chaillet, B. Lang, and A. Bourjault, "Experimental study and modelling of a tini shape memory alloy wire actuator," *Journal of Intelligent Material Systems and Structures* 8, pp. 619–629, 1997.
2. J. Sprekels, "Global existence for thermomechanical processes with nonconvex free energies of ginzburg-landau form," *Journal of Mathematical Analysis and Applications* 141, pp. 333–348, 1989.
3. N. Bubner, J. Sokolowski, and J. Sprekels, "Optimal boundary control problems for shape memory alloys under state constraints for stress and temperature," *Numer. Funct. Anal. and Optimiz.* 19(5&6), pp. 489–498, 1998.
4. R. S. Anderssen, I. G. Gotz, and K.-H. Hoffmann, "The global behavior of elastoplastic and viscoelastic materials with hysteresis-type state equations," *SIAM J. Appl. Math.* 58(2), pp. 703–723, 1998.
5. O. Klein, "Stability and uniqueness results for a numerical approximation of the thermomechanical phase transitions in shape memory alloys," *Advances in Mathematical Sciences and Applications (Tokyo)* 5(1), pp. 91–116, 1995.
6. M. Niezgodka and J. Sprekels, "Convergent numerical approximations of the thermomechanical phase transitions in shape memory alloys," *Numer. Math.* 58, pp. 759–778, 1991.
7. H. Alt, K.-H. Hoffmann, M. Niezgodka, and J. Sprekels, "A numerical study of structural phase transitions in shape memory alloys," Tech. Rep. 90, Department of Mathematics, University of Augsburg, 1985.
8. R. V. N. Melnik and A. J. Roberts, "Approximate models of dynamic thermoviscoelasticity describing shape-memory-alloy phase transitions," in *New Methods in Applied and Computational Mathematics (NEMACOM'98)*, D. Stewart and S. Oliviera, eds., *Proc. of the Centre for Mathematics and its Applications* (to appear; see <http://www.sci.usq.edu.au/cgi-bin/wp/research/workingpapers>), 1998.
9. F. Falk, "Model free energy, mechanics, and thermomechanics of shape memory alloys," *Acta Metallurgica* 28, pp. 1773–1780, 1980.
10. I. Müller and H. Xu, "On the pseudo-elastic hysteresis," *Acta Metall. Mater.* 39(3), pp. 263–271, 1991.
11. F. Falk and P. Konopka, "Three-dimensional landau theory describing the martensitic phase transformation of shape-memory alloys," *J. Phys.: Condens. Matter.* 2, pp. 61–77, 1990.
12. J. Carr, *Applications of Centre Manifold Theory*, Springer, Berlin, 1981.
13. S. M. Cox and A. J. Roberts, "Centre manifolds of forced dynamical systems," *J. Austral. Math. Soc. Ser. B* 32, pp. 401–436, 1991.
14. K.-H. Hoffmann and J. Zou, "Finite element approximations of landau-ginzburg's equation model for structural phase transitions in shape memory alloys," *M²AN* 29(6), pp. 629–655, 1995.
15. M. Bornert and I. Müller, "Temperature dependence of hysteresis in pseudoelasticity," in *Free Boundary Value Problems*, K.-H. Hoffmann and J. Sprekels, eds., pp. 27–35, Birkhauser, 1990.
16. H. I. Ene, M. L. Mascarenhas, and J. S. J. Paulin, "Fading memory effects in elastic-viscoelastic composites," *M²AN* 31(7), pp. 927–952, 1997.

USQ



TOOWOOMBA

**USER'S GUIDE TO SCSIMU: A
PACKAGE FOR NUMERICAL
SIMULATION OF SEMICONDUCTOR
DEVICES WITH THE QUASI-
HYDRODYNAMIC MODEL**

Hao He
Department of Theoretical Physics, University of Sydney

THE UNIVERSITY OF
S O U T H E R N
Q U E E N S L A N D

Working Papers

FACULTY
OF
SCIENCES

**USER'S GUIDE TO SCSIMU: A
PACKAGE FOR NUMERICAL
SIMULATION OF SEMICONDUCTOR
DEVICES WITH THE QUASI-
HYDRODYNAMIC MODEL**

Hao He

Department of Theoretical Physics, University of Sydney

R V N Melnik

Department of Mathematics & Computing, USQ

Faculty of Sciences Working Paper Series

SC-MC-9903

2 February 1999

USER'S GUIDE TO SCSIMU: A PACKAGE FOR NUMERICAL SIMULATION OF SEMICONDUCTOR DEVICES WITH THE QUASI-HYDRODYNAMIC MODEL

Hao He

Department of Theoretical Physics,
School of Physics, University of Sydney, NSW 2006

R. V. N. Melnik

Department of Mathematics and Computing,
University of Southern Queensland, QLD 4350, Australia

Abstract

In this paper we present results of a numerical simulation of semiconductor devices using the quasi-hydrodynamic model. The results have been obtained with the package SCSIMU, a C++ based program in which efficient exponential difference schemes for the quasi-hydrodynamic model have been implemented. Testing procedures and modelling of realistic semiconductor devices such as ballistic and PIN diodes are discussed with numerical results.

Key words: semiconductor device simulation, quasi-hydrodynamic models, exponential difference schemes.

1 Introduction

The quasi-hydrodynamic model for semiconductor device simulation plays an intermediate role between Boltzman's type models and drift-diffusion models. While the later models are known to be insufficient for the description of nonlocal and non-equilibrium processes in semiconductor plasma, the high computational cost and "noisiness" (with a great deal of redundant information) of Boltzman models preclude their effective use in engineering practice. Therefore, the development of computational software for quasi-hydrodynamic type models constitutes an important and challenging task for the engineering simulation of semiconductor devices. The basis of our current deliberations is the following quasi-hydrodynamic model:

$$\left\{ \begin{array}{l} \partial_{xx}\varphi = q(n - p - N)/\epsilon\epsilon_0 , \\ \partial_t n - \partial_x J_n/q = F , \\ \partial_t p + \partial_x J_p/q = F , \\ \partial_t \bar{\mathcal{E}}_n + \partial_x Q_n = -J_n \partial_x \varphi + P_n , \\ \partial_t \bar{\mathcal{E}}_p + \partial_x Q_p = -J_p \partial_x \varphi + P_p , \end{array} \right. \quad (1.1)$$

where n and p are concentrations of the majority (electrons) and minority (holes) carriers respectively, φ is the electrostatic potential, F is the generation/recombination/ionisation term, P_n and P_p are the rates of energy loss (by scattering on the lattice) for electrons and holes respectively, J_n and J_p are the current densities, Q_n and Q_p are the energy fluxes, N is the doping density of the device (the summarised concentration of dopants), q is the electron charge, ϵ and ϵ_0 are the relative dielectric permittivity of the semiconductor material and of vacuum respectively, $\bar{\mathcal{E}}_n$ and $\bar{\mathcal{E}}_p$ are approximate energy densities of the carriers. The following relationships supplement system (1.1)

$$\bar{\mathcal{E}}_n = 1.5nT_n, \quad \bar{\mathcal{E}}_p = 1.5pT_p, \quad P_n = n(T_l - T_n)/\tau_\omega^n, \quad P_p = p(T_l - T_p)/\tau_\omega^p \quad (1.2)$$

$$D_n = \mu_n T_n, \quad D_p = \mu_p T_p, \quad \mu_n = \mu_n^0 \sqrt{T_n/T_l}, \quad \mu_p = \mu_p^0 \sqrt{T_p/T_l} \quad (1.3)$$

$$J_n = -qn\mu_n \partial_x \varphi + \partial_x(T_n \mu_n n), \quad J_p = -qp\mu_p \partial_x \varphi - \partial_x(T_p \mu_p p), \quad (1.4)$$

$$Q_n = \beta_n T_n n \mu_n \partial_x \varphi - \beta_n \partial_x[T_n D_n n]/q, \quad Q_p = -\beta_p T_p p \mu_p \partial_x \varphi - \beta_p \partial_x[T_p D_p p]/q, \quad (1.5)$$

where T_n and T_p are the carrier temperatures, T_l is the lattice temperature, τ_ω^n and τ_ω^p are average energy relaxation times for electrons and holes respectively, D_n (D_p) and μ_n (μ_p) are the diffusion and mobility coefficients. Note that the current densities can be equivalently defined as follows

$$J_n = -qn v_n, \quad \text{and} \quad J_p = qp v_p, \quad (1.6)$$

which provide us with formulae for computing carrier velocities, v_n and v_p . For simplicity, the Peltier coefficients, β_n and β_p are assumed to be constants equal to 2.5 and for the energy

relaxation times we use the following formulae

$$\tau_n^n = \frac{3\mu_n^0 \sqrt{T_l T_n}}{2q(v_s^n)^2}, \quad \mu_n^0 = 1400 \text{ cm}^2/(\text{Vs}) \quad \text{and} \quad \tau_p^p = \frac{3\mu_p^0 \sqrt{T_l T_p}}{2q(v_s^p)^2}, \quad \mu_p^0 = 400 \text{ cm}^2/(\text{Vs}), \quad (1.7)$$

where v_s^n and v_s^p are saturation velocities of carriers. The recombination term F is given by:

$$F(n, p) = \frac{pn - n_{ie}^2}{\tau_n(p + n_{ie}) + \tau_p(n + n_{ie})}, \quad (1.8)$$

where n_{ie} is the effective intrinsic concentration of carriers and the carrier life times are set as follows

$$\tau_n = 1.7 \times 10^{-5} \text{ s}, \quad \tau_p = 3.95 \times 10^{-4} \text{ s}.$$

Initial and boundary conditions for problem (1.1)–(1.8) are device specific. For the sake of technical convenience we write the problem in terms of dimensionless variables (keeping the same notation as in (1.1)–(1.8)):

$$\left\{ \begin{array}{l} \partial_{xx}\varphi = n - p - N, \\ \partial_t n - \partial_x J_n = F, \\ \partial_t p + \partial_x J_p = F, \\ 3/2\partial_t(nT_n) + \partial_x Q_n = -J_n\partial_x\varphi + P_n, \\ 3/2\partial_t(pT_p) + \partial_x Q_p = -J_p\partial_x\varphi + P_p, \end{array} \right. \quad (1.9)$$

where

$$J_n = -n\mu_n\partial_x\varphi + \partial_x(T_n\mu_n n), \quad J_p = -p\mu_p\partial_x\varphi - \partial_x(T_p\mu_p p), \quad (1.10)$$

$$Q_n = \beta_n T_n n \mu_n \partial_x \varphi - \beta_n \partial_x [T_n D_n n], \quad Q_p = -\beta_p T_p p \mu_p \partial_x \varphi - \beta_p \partial_x [T_p D_p p]. \quad (1.11)$$

2 Semiconductor Devices Used in Numerical Experiments

To solve the coupled system partial differential equations (1.9)–(1.11) we developed an efficient C++ code, which is easily adaptable for a wide range of semiconductor devices. In this paper we present results of the simulation for two types of devices. The typical representative of the first type is the silicon-based $p^+ - i - n^+$ (PIN) diode used extensively for microwave control applications such as microwave switches and for electronically steered phased-array antennas [3]. The length of the simulated diode is chosen to be $3.5 \mu\text{m}$ (see also [3]). The p^+ region of length $0.5 \mu\text{m}$ is doped at density -10^{18} cm^{-3} . The central region of length $2.0 \mu\text{m}$ is doped at density $3.5 \times 10^{14} \text{ cm}^{-3}$, while the n^+ region is doped at $2.4 \times 10^{19} \text{ cm}^{-3}$. A schematic doping profile of this device is shown in Figure 1(a).

The typical representative of the other type of devices is the $n^+ - n - n^+$ ballistic diode [2, 1]. This device is often used to model the $n^+ - n - n^+$ channel in MEtal-Semiconductor

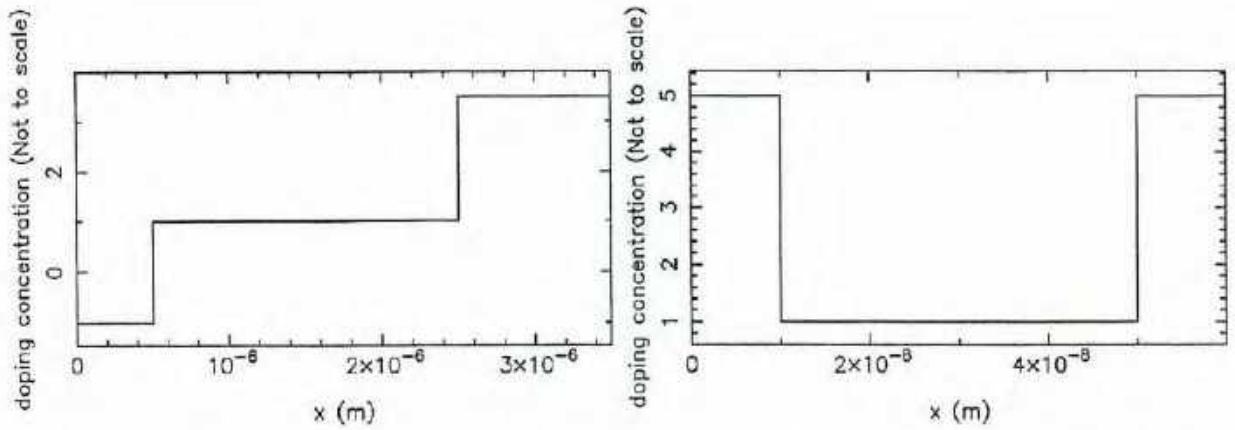


Figure 1: Doping distribution in semiconductor structure. (a) PIN diode; (b) ballistic diode.

Field-Effect Transistors (MESFET). The simulated diode has a central n region of length $0.4 \mu\text{m}$ bounded by two n^+ regions of length $0.1 \mu\text{m}$ each. The n^+ regions are doped at density $N = 5 \times 10^{17} \text{ cm}^{-3}$, while the n region is doped at $N = 2 \times 10^{15} \text{ cm}^{-3}$. We show the schematic doping profile for this device in Figure 1 (b).

3 Numerical Approximations

In this section we briefly describe the main ideas of our numerical scheme for the stationary case.

3.1 Discretization of the Poisson equation

For the solution of the Poisson equation,

$$\partial_{xx}\varphi = n - p - N, \quad (3.1)$$

we use the following discretisation

$$\begin{aligned} \mathcal{F}_i^\varphi(\varphi_{i-1}^{l+1}, \varphi_i^{l+1}, \varphi_{i+1}^{l+1}) &= (\varphi_{i+1}^{l+1} - \varphi_i^{l+1})/h_{i+1} - (\varphi_i^{l+1} - \varphi_{i-1}^{l+1})/h_i - \\ h_i^* [n_i^l \exp(\varphi_i^{l+1} - \varphi_i^l) - p_i^l \exp(-\varphi_i^{l+1} + \varphi_i^l) - N] &= 0. \end{aligned} \quad (3.2)$$

Note that in this discretisation we used the relationships $n^{l+1} = n^l \exp(\varphi^{l+1} - \varphi^l)$, and $p^{l+1} = p^l \exp(\varphi^l - \varphi^{l+1})$ (written in normalised units), easily obtainable under the assumption of constant carrier temperatures. Then, the Jacobian of system (3.2) is determined by the following partial derivatives:

$$\frac{\partial \mathcal{F}_i^\varphi}{\partial \varphi_{i-1}^{l+1}} = \frac{1}{h_i}, \quad \frac{\partial \mathcal{F}_i^\varphi}{\partial \varphi_{i+1}^{l+1}} = \frac{1}{h_{i+1}}, \quad (3.3)$$

$$\frac{\partial \mathcal{F}_i^\varphi}{\partial \varphi_i^{l+1}} = -\frac{1}{h_{i+1}} - \frac{1}{h_i} - h_i^* (n_i^l \exp(\varphi_i^{l+1} - \varphi_i^l) + p_i^l \exp(-\varphi_i^{l+1} + \varphi_i^l)). \quad (3.4)$$

The banded structure of this Jacobian permits an efficient computational implementation of the linearised system using a sparse solver. Indeed, approximations to each nonlinear equation of system (1.9) can be written in the form

$$\tilde{\mathcal{F}}(\mathbf{x}) = 0, \quad (3.5)$$

where $\tilde{\mathcal{F}}$ is an equation-specific discrete operator and \mathbf{x} is a quantity to be determined (say, φ). We start with an initial approximate solution \mathbf{x}' and solve the following system of linear equations:

$$\mathbf{J} \cdot \delta \mathbf{x} = -\tilde{\mathcal{F}}(\mathbf{x}'), \quad (3.6)$$

with the Jacobian of the system (that is a sparse matrix) given by

$$J_{ij} = \frac{\partial \mathcal{F}_i}{\partial x_j} \quad (3.7)$$

The values of \mathbf{x}' are then updated according to the standard methodology

$$\mathbf{x}'_{new} \leftarrow \mathbf{x}'_{old} + \delta \mathbf{x} \quad (3.8)$$

and steps (3.6)–(3.8) are repeated until the required accuracy is achieved.

3.2 Discretization of the continuity equations

We use exponential difference schemes for the discretisation of continuity equations. In the stationary case these equations have the form:

$$-\partial_x J_n = F, \quad \text{where } J_n = -n \mu_n \partial_x \varphi + \partial_x (T_n \mu_n n), \quad (3.9)$$

$$\partial_x J_p = F, \quad \text{where } J_p = -p \mu_p \partial_x \varphi - \partial_x (T_p \mu_p p). \quad (3.10)$$

Equation (3.9) is discretised as follows:

$$\tilde{\mathcal{F}}_i^n = A_i^n n_{i-1}^{l+1} + B_i^n n_{i+1}^{l+1} - C_i^n n_i^{l+1} + h_i^* F_i(n_i^{l+1}, p_i) = 0, \quad (3.11)$$

where

$$A_i^n = \frac{D_n[(T_n)_{i-1}^{l+1}]}{h_i} f \left(\frac{\varphi_i^{l+1} - \varphi_{i-1}^{l+1}}{(T_n)_{i-1}^{l+1}} \right), \quad (3.12)$$

$$B_i^n = \frac{D_n[(T_n)_{i+1}^{l+1}]}{h_{i+1}} f_1 \left(\frac{\varphi_{i+1}^{l+1} - \varphi_i^{l+1}}{(T_n)_{i+1}^{l+1}} \right), \quad C_i^n = A_{i+1}^n + B_{i-1}^n. \quad (3.13)$$

The Jacobian of system (3.11) is determined by:

$$\frac{\partial \mathcal{F}_i^n}{\partial n_{i-1}^{l+1}} = A_i^n, \quad \frac{\partial \mathcal{F}_i^n}{\partial n_i^{l+1}} = -C_i^n + h_i^* \frac{\partial F_i}{\partial n_i^{l+1}}, \quad \frac{\partial \mathcal{F}_i^n}{\partial n_{i+1}^{l+1}} = B_i^n. \quad (3.14)$$

Similarly, we obtain the discretised equation and the Jacobian for the continuity equation (3.10):

$$\mathcal{F}_i^p = A_i^p p_{i-1}^{l+1} + B_i^p p_{i+1}^{l+1} - C_i^p p_i^{l+1} + h_i^* F_i(n_i, p_i^{l+1}) = 0, \quad (3.15)$$

where

$$A_i^p = \frac{D_p[(T_p)_{i-1}^{l+1}]}{h_i} f_1 \left(\frac{\varphi_i^{l+1} - \varphi_{i-1}^{l+1}}{(T_p)_{i-1}^{l+1}} \right), \quad (3.16)$$

$$B_i^n = \frac{D_p[(T_p)_{i+1}^{l+1}]}{h_{i+1}} f \left(\frac{\varphi_{i+1}^{l+1} - \varphi_i^{l+1}}{(T_p)_{i+1}^{l+1}} \right), \quad C_i^p = A_{i+1}^n + B_{i-1}^n, \quad (3.17)$$

$$\frac{\partial \mathcal{F}_i^p}{\partial p_{i-1}^{l+1}} = A_i^p, \quad \frac{\partial \mathcal{F}_i^p}{\partial p_i^{l+1}} = -C_i^p + h_i^* \frac{\partial F_i}{\partial p_i^{l+1}}, \quad \frac{\partial \mathcal{F}_i^p}{\partial p_{i+1}^{l+1}} = B_i^n. \quad (3.18)$$

3.3 Discretization of the energy balance equations

The energy balance equations are also approximated by exponential difference schemes. In the stationary case these equations have the form

$$\partial_x Q_n = -J_n \partial_x \varphi + P_n, \quad \partial_x Q_p = J_p \partial_x \varphi + P_p. \quad (3.19)$$

The first equation in (3.19) is approximated by the following scheme (with respect to the time layer $l+1$):

$$\mathcal{F}_i^{\mathcal{E}_n} = \bar{A}_i^n (\mathcal{E}_n)_{i-1} + \bar{B}_i^n (\mathcal{E}_n)_{i+1} - \tilde{C}_i^n (\mathcal{E}_n)_i + R_i^{\mathcal{E}_n} = 0, \quad (3.20)$$

where $\mathcal{E}_n = n T_n$,

$$\begin{aligned} R_i^{\mathcal{E}_n} = & -h_i^* \left\{ -\mu_n[(T_n)_i] \varphi_{xx,i} - \mu_n[(T_n)_i] (\varphi_{x,i})^2 / (T_n)_i + \right. \\ & \left. + 1/\tau_\omega^n [(T_n)_i] - 1/(\tau_\omega^n [(T_n)_i] (T_n)_i) \right\} (\mathcal{E}_n)_i, \end{aligned} \quad (3.21)$$

and the coefficients of this difference scheme are defined as follows

$$(\tilde{A})_i^n = \frac{\beta_n D_n[(T_n)_{i-1}^{l+1}]}{h_i} f \left(\frac{1 + \beta_n}{\beta_n} \frac{\varphi_i^{l+1} - \varphi_{i-1}^{l+1}}{(T_n)_{i-1}^{l+1}} \right), \quad (3.22)$$

$$(\tilde{B})_i^n = \frac{\beta_n D_n[(T_n)_{i+1}^{l+1}]}{h_{i+1}} f \left(\frac{1 + \beta_n}{\beta_n} \frac{\varphi_{i+1}^{l+1} - \varphi_i^{l+1}}{(T_n)_{i+1}^{l+1}} \right), \quad \tilde{C}_i^n = \bar{A}_{i+1}^n + \bar{B}_{i-1}^n. \quad (3.23)$$

Here we use standard difference scheme notation [7, 8] such as

$$\varphi_{xx,i} = \frac{1}{h_i^*} \left[\frac{\varphi_{i+1} - \varphi_i}{h_{i+1}} - \frac{\varphi_i - \varphi_{i-1}}{h_i} \right], \quad h_i = x_i - x_{i-1},$$

$$\varphi_{\bar{x},i} = (\varphi_{i+1} - \varphi_{i-1})/(2h_i^*), \quad h_i^* = (h_i + h_{i+1})/2.$$

Similarly, the discretised equation for holes is (at the time layer $l+1$):

$$\mathcal{F}_i^{\mathcal{E}_p} = \bar{A}_i^p(\mathcal{E}_p)_{i-1} + \bar{B}_i^p(\mathcal{E}_p)_{i+1} - \bar{C}_i^p(\mathcal{E}_p)_i + R_i^{\mathcal{E}_p} = 0 \quad (3.24)$$

where $\mathcal{E}_p = pT_p$,

$$\begin{aligned} R_i^{\mathcal{E}_p} = & -h_i^* \{ \mu_p[(T_p)_i] \varphi_{\bar{x},i} - \mu_n[(T_p)_i] (\varphi_{\bar{x},i})^2 / (T_p)_i + \\ & + 1/\tau_\omega^p[(T_p)_i] - 1/(\tau_\omega^p[(T_p)_i](T_p)_i) \} (\mathcal{E}_p)_i, \end{aligned} \quad (3.25)$$

and the coefficients of this difference scheme are defined as follows

$$(\bar{A})_i^p = \frac{\beta_p D_p[(T_p)_{i-1}^{l+1}]}{h_i} f_1 \left(\frac{1 + \beta_p \varphi_i^{l+1} - \varphi_{i-1}^{l+1}}{\beta_p (T_p)_{i-1}^{l+1}} \right), \quad (3.26)$$

$$(\bar{B})_i^p = \frac{\beta_p D_p[(T_p)_{i+1}^{l+1}]}{h_{i+1}} f \left(\frac{1 + \beta_p \varphi_{i+1}^{l+1} - \varphi_i^{l+1}}{\beta_p (T_p)_{i+1}^{l+1}} \right), \quad \bar{C}_i^p = \bar{A}_{i+1}^p + \bar{B}_{i-1}^p. \quad (3.27)$$

In order to adequately describe physical processes in a number of semiconductor devices the splitting algorithm based upon an “independent” solution of each nonlinear equation of system (1.9) may not be appropriate. In such cases energy balance equations have to be solved simultaneously with continuity equations leading to new challenges in the computational implementation of the proposed discrete schemes. We address these issues in Section 5.

4 Drift Diffusion Approximation as an Initial Approximation for the Solution of the Quasi-Hydrodynamic Model

Assuming that carriers are in thermal equilibrium with the lattice and that carrier temperatures are equal to the lattice temperature, we consider the standard drift diffusion model in the stationary case:

$$\begin{cases} \partial_{xx}\varphi = n - p - N, \\ -\partial_x J_n = F, \\ \partial_x J_p = F. \end{cases} \quad (4.1)$$

This system provides an initial approximation for the solution of the more complicated system (1.9). We solve each equation in system (4.1) individually, so that the results from the previous equation are used for the next equation until the global convergence is reached. In order to accelerate the convergence, the Poisson equation is solved twice, in each “semi-global” iteration after solving each continuity equation. The flowchart of the algorithm is shown in Figure 2.

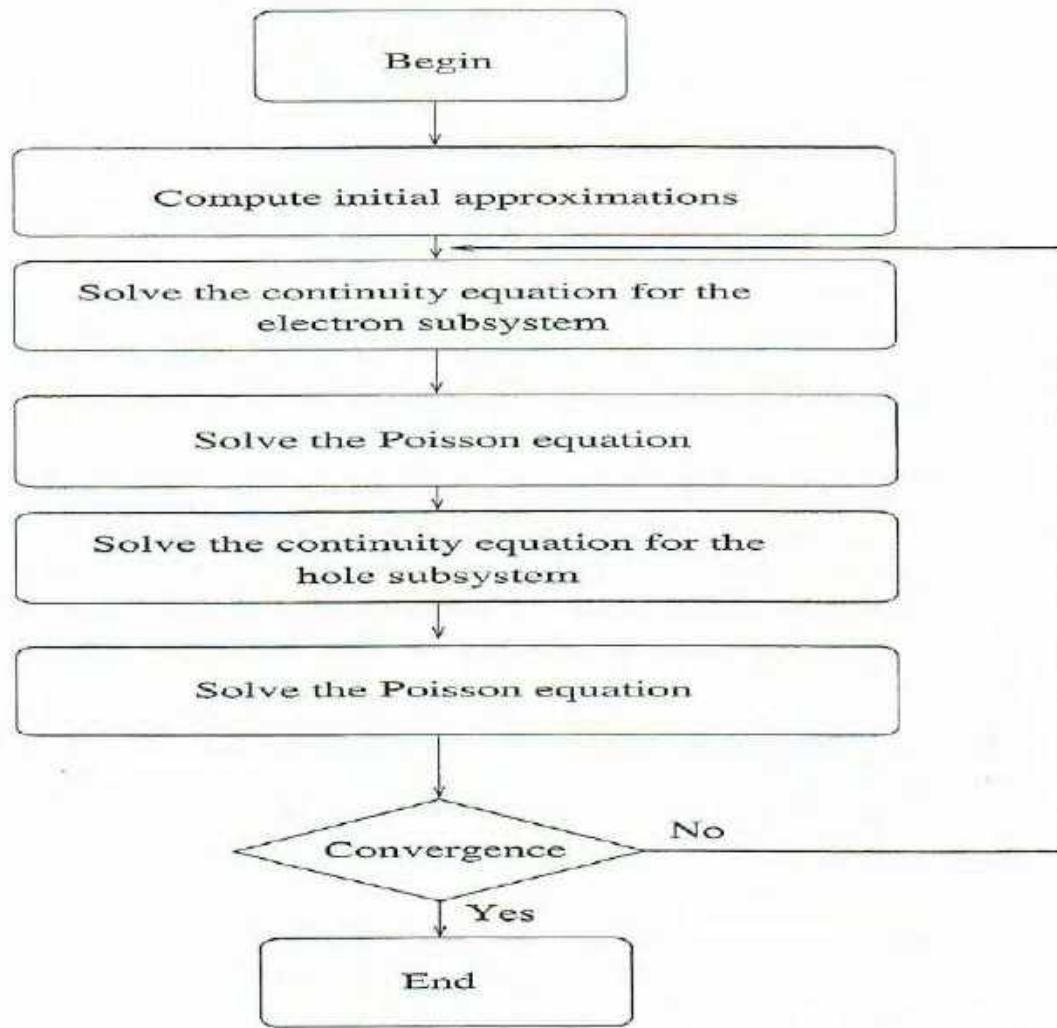


Figure 2: Flowchart of Program 1: the solution of the drift-diffusion model.

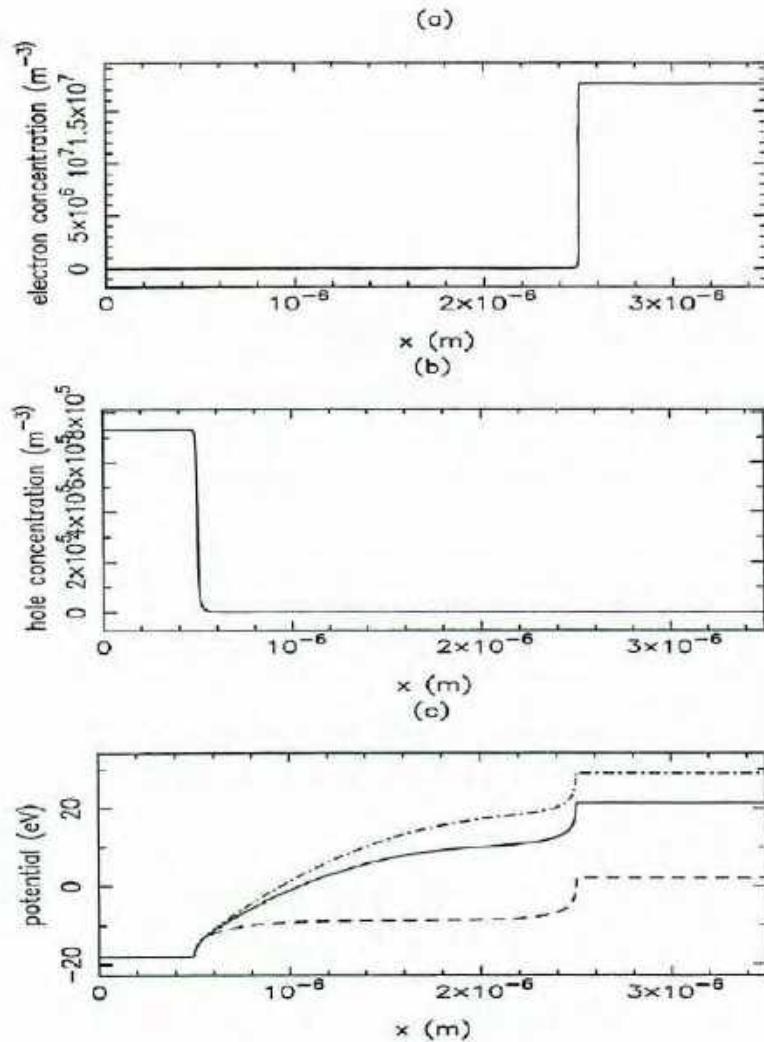


Figure 3: Results from Program 1: (a) concentration of electrons; (b) concentration of holes; (c) potential profile (the applied voltages are $V=-0.5, 0.0, 0.2$ (dashed, solid, and dot-dashed lines respectively).

The initial approximations for the drift-diffusion model are determined by the quasi-neutrality condition

$$n_{ie} \exp(V - \varphi) - n_{ie} \exp(\varphi - V) + N = 0, \quad (4.2)$$

where the initial approximation for the potential has the form

$$\varphi = V + \text{sign}(N) \ln(N/n_{ie}), \quad (4.3)$$

and V is the applied voltage. Then the concentrations can be determined using the Boltzmann statistics as follows

$$n = n_{ie} \exp(\varphi), \quad p = n_{ie} \exp(-\varphi), \quad (4.4)$$

where $n_{ie} = 1.4 \times 10^{10} \text{ cm}^{-3}$. As an example, we present in Figure 3 the results of computation of physical characteristics of the PIN device for different applied voltages $V=-0.5, 0.0, 0.2$.

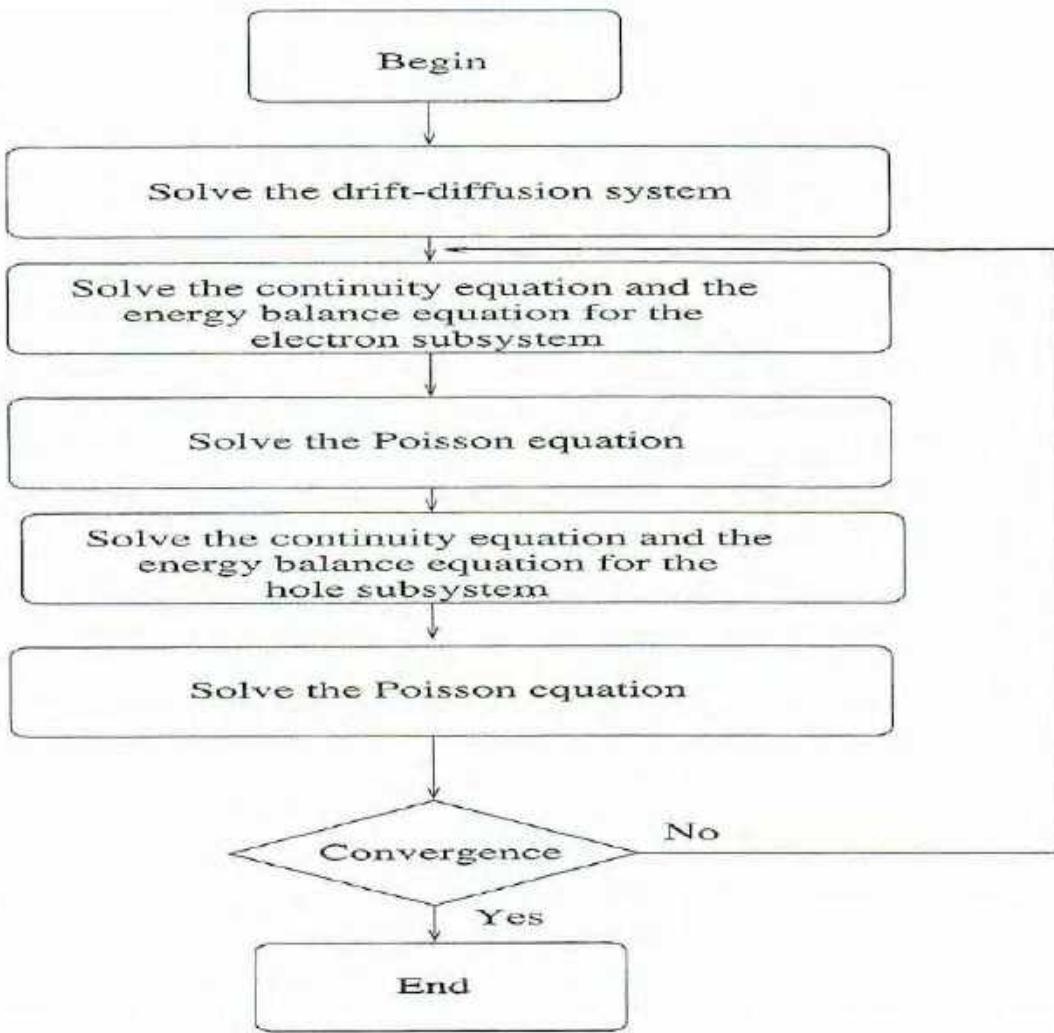


Figure 4: Flowchart of Program 2: the solution of the quasi-hydrodynamic model.

5 Computation with the Quasi-Hydrodynamic Model

Taking the solution of the drift-diffusion model as an initial approximation, we developed a program that accounts for the effect of carrier temperatures. This program allows us to solve the full system (1.9) including the energy balance equations. From the solution of the drift-diffusion system (4.1), in a single computational block we simultaneously solve the equation of continuity and the energy balance equation for electrons. After the new potential values are obtained from the Poisson equation, we solve (again, in a single computational block) the equation of continuity and the energy balance equation for holes. Finally, we correct the potential values by solving the Poisson equation with newly available values for the concentration and temperature of carriers. The whole procedure (illustrated in Figure 4) is repeated until the required accuracy is achieved. In implementing coupled computational blocks (between the continuity and energy balance equations), the Jacobian of the linearised system of non-linear equations requires a special attention. For the electron subsystem, the

Jacobian takes the form of the following matrix

$$\mathbf{J}_e = \begin{pmatrix} \frac{\partial \vec{\mathcal{F}}^n}{\partial \vec{n}} & \frac{\partial \vec{\mathcal{F}}^n}{\partial \vec{T}_n} \\ \frac{\partial \vec{\mathcal{F}}^{\mathcal{E}_n}}{\partial \vec{n}} & \frac{\partial \vec{\mathcal{F}}^{\mathcal{E}_n}}{\partial \vec{T}_n} \end{pmatrix}. \quad (5.1)$$

The first (upper left) block of the Jacobian $(\partial \vec{\mathcal{F}}^n / \partial \vec{n})$ is the same as in Section 3. The second (upper right) block $\partial \vec{\mathcal{F}}^n / \partial \vec{T}_n$ is determined by

$$\frac{\partial \mathcal{F}_i^n}{\partial (T_n)_{i-1}} = \frac{\partial A_i^n}{\partial (T_n)_{i-1}} n_{i-1}, \quad \frac{\partial \mathcal{F}_i^n}{\partial (T_n)_i} = \frac{\partial C_i^n}{\partial (T_n)_i} n_i, \quad \frac{\partial \mathcal{F}_i^n}{\partial (T_n)_{i+1}} = \frac{\partial B_i^n}{\partial (T_n)_{i+1}} n_{i+1}. \quad (5.2)$$

The third (lower left) block gives:

$$\frac{\partial \mathcal{F}_i^{\mathcal{E}_n}}{\partial n_{i-1}} = \tilde{A}_i^n(T_n)_{i-1}, \quad \frac{\partial \mathcal{F}_i^{\mathcal{E}_n}}{\partial n_i} = C_i^n(T_n)_i + \frac{\partial R_i^{\mathcal{E}_n}}{\partial n_i}, \quad \frac{\partial \mathcal{F}_i^{\mathcal{E}_n}}{\partial n_{i+1}} = \tilde{B}_i^n(T_n)_{i+1}, \quad (5.3)$$

and finally, the fourth (lower right) block of the matrix (5.1) is determined by

$$\frac{\partial \mathcal{F}_i^{\mathcal{E}_n}}{\partial (T_n)_{i-1}} = \frac{\partial \tilde{A}_i^n}{\partial (T_n)_{i-1}} (\mathcal{E}_n)_{i-1} + \tilde{A}_i^n n_{i-1}, \quad \frac{\partial \mathcal{F}_i^{\mathcal{E}_n}}{\partial (T_n)_i} = \frac{\partial \tilde{C}_i^n}{\partial (T_n)_i} (\mathcal{E}_n)_i + \tilde{C}_i^n n_i + \frac{\partial R_i^{\mathcal{E}_n}}{\partial (T_n)_i}, \quad (5.4)$$

$$\frac{\partial \mathcal{F}_i^{\mathcal{E}_n}}{\partial (T_n)_{i+1}} = \frac{\partial \tilde{B}_i^n}{\partial (T_n)_{i+1}} (\mathcal{E}_n)_{i+1} + \tilde{B}_i^n n_{i+1}. \quad (5.5)$$

The derivatives in relationships (5.2)–(5.5) are

$$\frac{\partial A_i^n}{\partial (T_n)_{i-1}} = \frac{1}{h_{i-1}} \left[\frac{\partial D_n}{\partial (T_n)_{i-1}} f(\lambda(\varphi, T_n)_i) - D_n f'(\lambda(\varphi, T_n)_i) \frac{\lambda(\varphi, T_n)_i}{((T_n)_{i-1})^{i+1}} \right] \quad (5.6)$$

$$\begin{aligned} \frac{\partial B_i^n}{\partial (T_n)_{i+1}} &= \frac{1}{h_{i+1}} \left[\frac{\partial D_n}{\partial (T_n)_{i+1}} f_1(\lambda(\varphi, T_n)_{i+1}) - D_n f'_1(\lambda(\varphi, T_n)_{i+1}) \times \right. \\ &\quad \left. \frac{\lambda(\varphi, T_n)_{i+1}}{(T_n)_{i+1}^{i+1}} \right], \quad \frac{\partial C_i^n}{\partial (T_n)_i} = \frac{\partial A_i^n}{\partial (T_n)_{i+1}} + \frac{\partial B_i^n}{\partial (T_n)_{i-1}} \end{aligned} \quad (5.7)$$

and

$$\frac{\partial \tilde{A}_i^n}{\partial (T_n)_{i-1}} = \frac{1}{h_{i-1}} \left[\beta_n \frac{\partial D_n}{\partial (T_n)_{i-1}} f(\lambda(\varphi, T_n)_i) - \tilde{\beta}_n D_n f'(\lambda(\varphi, T_n)_i) \frac{\lambda(\varphi, T_n)_i}{((T_n)_{i-1})^{i+1}} \right] \quad (5.8)$$

$$\begin{aligned} \frac{\partial \tilde{B}_i^n}{\partial (T_n)_{i+1}} &= \frac{1}{h_{i+1}} \left[\beta_n \frac{\partial D_n}{\partial (T_n)_{i+1}} f_1(\lambda(\varphi, T_n)_{i+1}) - \tilde{\beta}_n D_n f'_1(\lambda(\varphi, T_n)_{i+1}) \times \right. \\ &\quad \left. \frac{\lambda(\varphi, T_n)_{i+1}}{(T_n)_{i+1}^{i+1}} \right], \quad \frac{\partial \tilde{C}_i^n}{\partial (T_n)_i} = \frac{\partial \tilde{A}_i^n}{\partial (T_n)_{i+1}} + \frac{\partial \tilde{B}_i^n}{\partial (T_n)_{i-1}} \end{aligned} \quad (5.9)$$

where

$$\lambda(\varphi, T_n)_i = \frac{\varphi_i^{t+1} - \varphi_{i-1}^{t+1}}{(T_n)_{i-1}^{t+1}}.$$

The only term left to be determined is $R_i^{\mathcal{E}_n}$:

$$\begin{aligned} R_i^{\mathcal{E}_n} = & -h_i^* \left\{ -\mu_n[(T_n)_i] \varphi_{\bar{x}\bar{x},i} - \mu_n[(T_n)_i] (\varphi_{\bar{x},i})^2 / (T_n)_i + \right. \\ & \left. + 1/\tau_\omega^n[(T_n)_i] - 1/(\tau_\omega^n[(T_n)_i](T_n)_i) \right\} (\mathcal{E}_n)_i, \end{aligned} \quad (5.10)$$

where the normalized functions μ_n and τ_ω^n are given in the form: $\mu_n[(T_n)_i] = \mu_n^0 \sqrt{(T_n)_i}$ and $\tau_\omega^n = 1.5 \mu_n^0 \sqrt{(T_n)_i} / (v_s^n)^2$. Hence

$$R_i^{\mathcal{E}_n} = h_i^* \left(\mu_n^0 (T_n)_i^{1.5} \varphi_{\bar{x}\bar{x},i} + \mu_n^0 \sqrt{(T_n)_i} (\varphi_{\bar{x},i})^2 + \frac{2(1 - (T_n)_i)(v_s^n)^2}{3\mu_0^n \sqrt{(T_n)_i}} \right) n_i, \quad (5.11)$$

and

$$\frac{\partial R_i^{\mathcal{E}_n}}{\partial (T_n)_i} = h_i^* \left(1.5 \mu_n^0 \sqrt{(T_n)_i} \varphi_{\bar{x}\bar{x},i} + \frac{\mu_n^0}{2\sqrt{(T_n)_i}} (\varphi_{\bar{x},i})^2 - \frac{(1 + (T_n)_i)(v_s^n)^2}{3\mu_0^n (T_n)_i^{3/2}} \right) n_i. \quad (5.12)$$

Finally, in this section we provide results of computational experiments with the quasi-hydrodynamic model. All computations have been performed for 2^{10} grid points and this value was set as the default in the standard worksheet (see Section 6). The first group of experiments was designed for the ballistic diode (see Figure 1(b)). The distributions of concentrations, potential and temperature along the semiconductor structure for the two different applied voltages, $V=0.1V$ and $V=0.5V$, are given in Figures 5, 6 and 7 respectively. In Figure 8 we provide the profile of electron velocity, and in Figure 9 the Mach number (computed as the ratio $v_n / \sqrt{5T_n / (3m_n)}$, where m_n is the effective electron mass) is displayed.

The second group of experiments was conducted for a bipolar-type of devices, the PIN diode (see Figure 1(a)). In Figures 10–14 we present the computed distributions of the potential as well as concentrations and temperatures of carriers for this device. Plots of electron velocity and of the Mach number are given in Figures 15 and 16 respectively. A high, narrow velocity peak and a narrow area where the Mach number is greater than one, suggest that the quasi-hydrodynamic model may overestimate some physical characteristics of the devices. Similar effects were previously reported in the literature [2, 1].

We note that PIN diodes can work in both forward and reverse biased regimes. In the forward-bias case these devices exhibit a very low RF resistance, a higher conductivity and a larger breakdown than standard PN diodes. As can be seen from Figures 10–14, deviations of carrier temperature away from equilibrium values do not have a significant influence on the main characteristics of the device. This may not be the case in the reverse-bias condition [6].

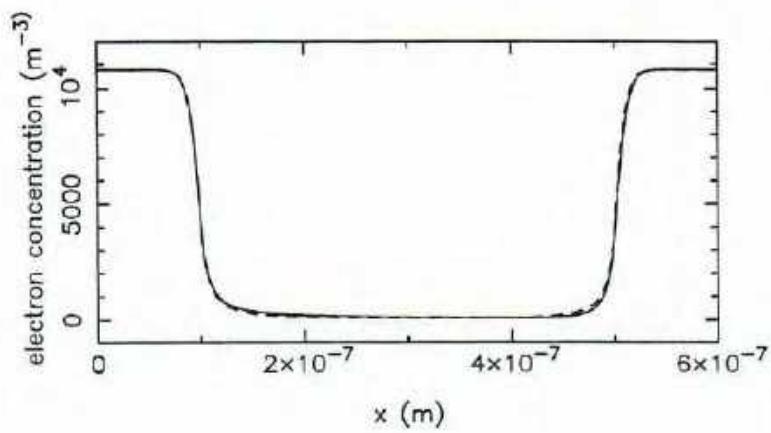


Figure 5: Concentration of electrons in normalized unit. Applied voltages are 0.1 V (dashed line) and 0.5 V (solid line).

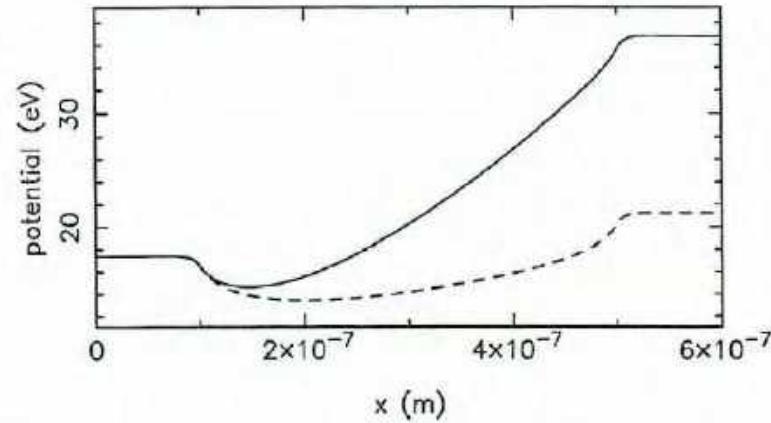


Figure 6: Potential profile in normalized unit. Applied voltages are 0.1 V (dashed line) and 0.5 V (solid line).

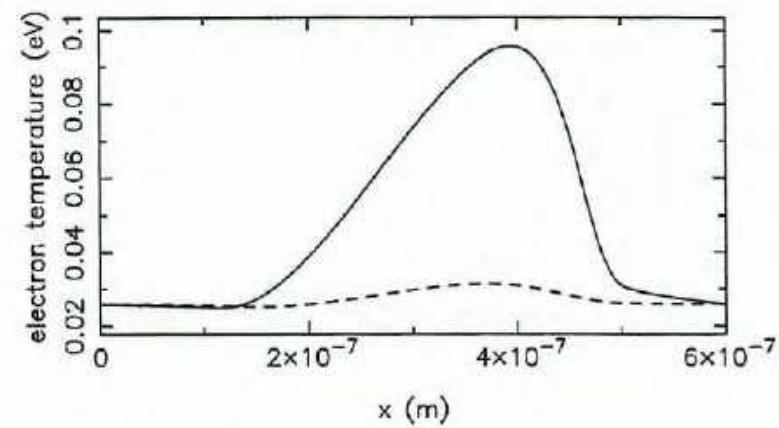


Figure 7: Temperature distributions of electrons in unit eV. Applied voltages are 0.1 V (dashed line) and 0.5 V (solid line).

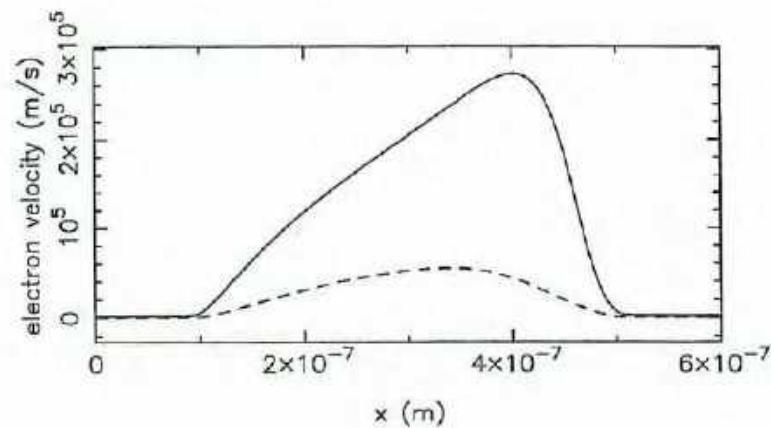


Figure 8: Velocity distributions of electrons in unit m/s . Applied voltages are 0.1 V (dashed line) and 0.5 V (solid line).

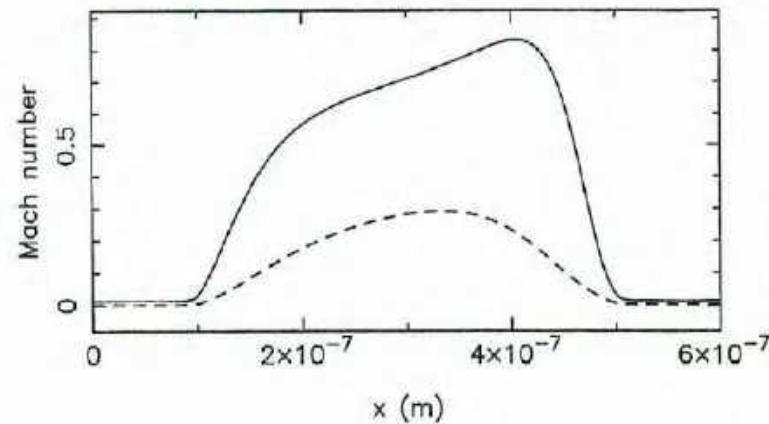


Figure 9: Mach number of electrons. Applied voltages are 0.1 V (dashed line) and 0.5 V (solid line).

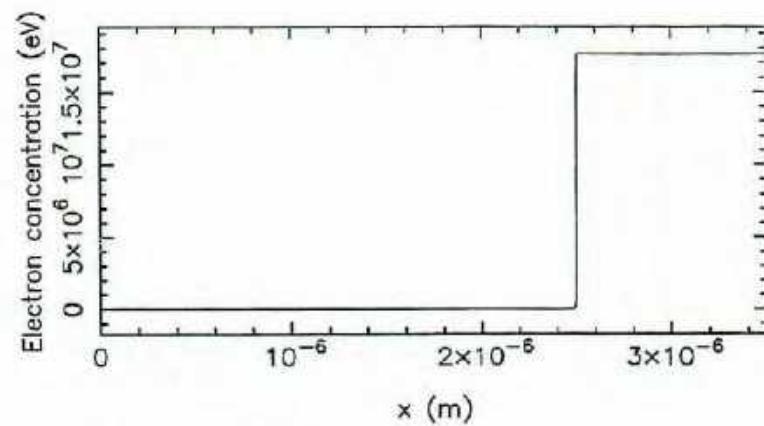


Figure 10: Concentration of electrons in normalized unit. Applied voltages are -0.5 V (dashed line), 0 V (solid line), and 0.05 V (dash-dotted line).

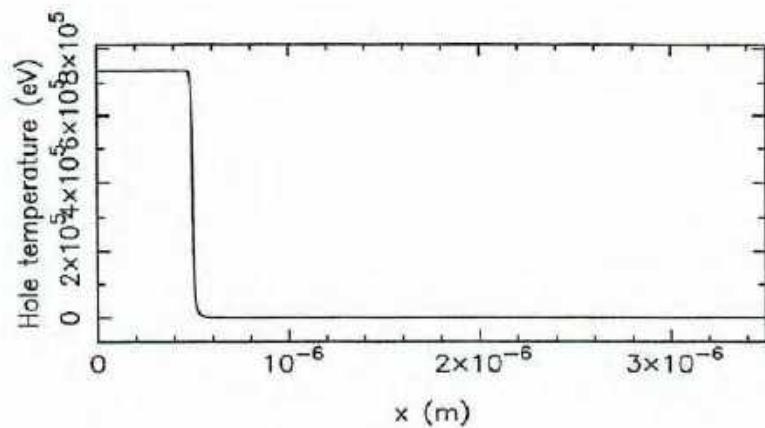


Figure 11: Concentration of holes in normalized unit. Applied voltages are -0.5 V (dashed line), 0. V (solid line), and 0.05 V (dash-dotted line).

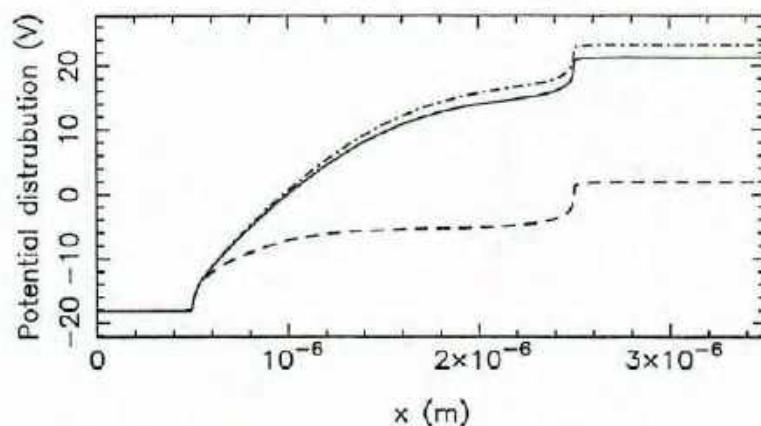


Figure 12: Potential profile in normalized unit. Applied voltages are -0.5 V (dashed line), 0. V (solid line), and 0.05 V (dash-dotted line).

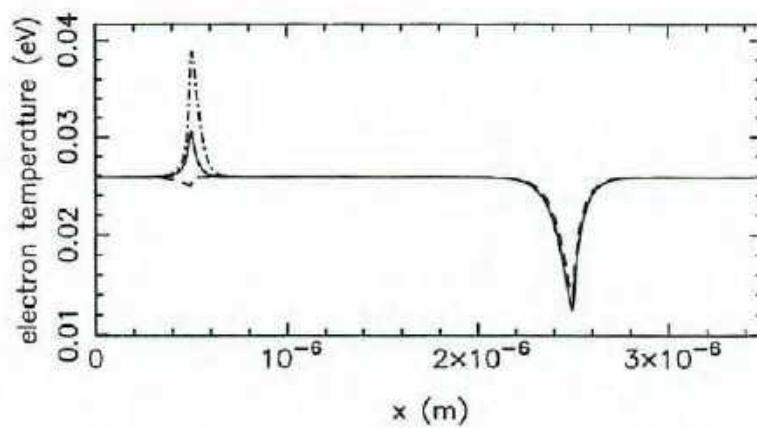


Figure 13: Temperature distributions of electrons in unit eV . Applied voltages are -0.5 V (dashed line), 0. V (solid line), and 0.05 V (dash-dotted line).

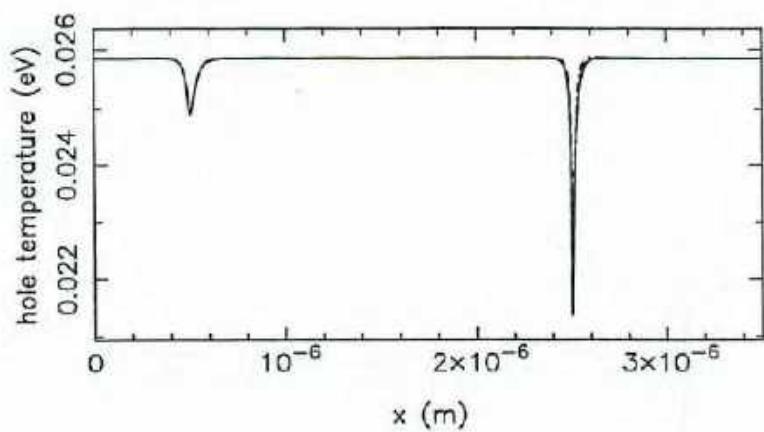


Figure 14: Temperature distributions of holes in unit eV. Applied voltages are -0.5 V (dashed line), 0. V (solid line), and 0.05 V (dash-dotted line).

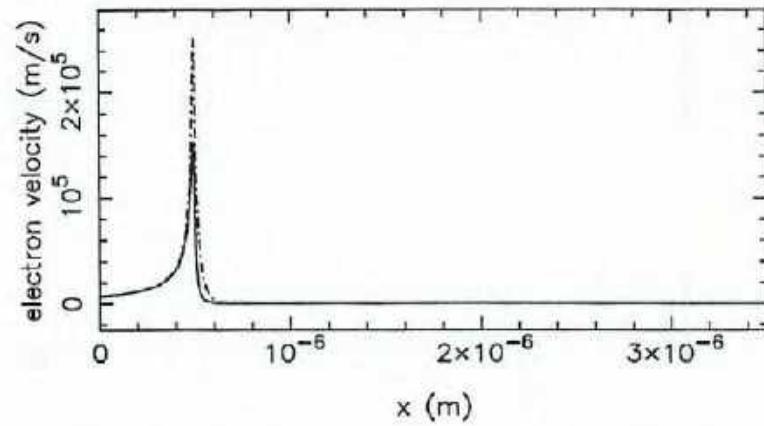


Figure 15: Velocity distributions of electrons in units m/s. Applied voltages are 0. V (solid line) and 0.05 V (dash-dotted line).

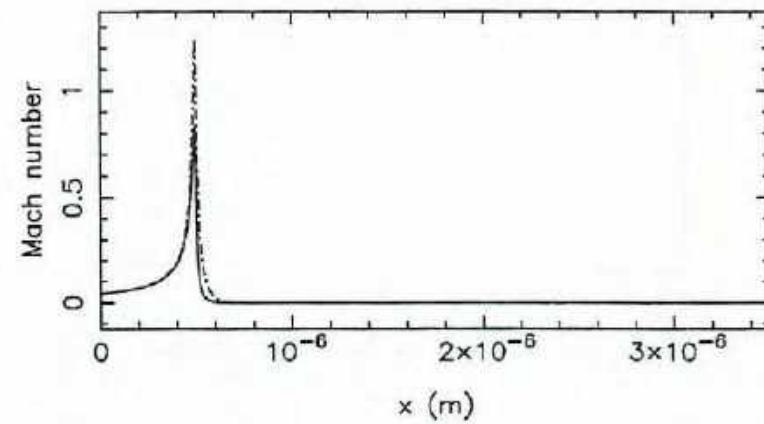


Figure 16: Mach number of electrons. Applied voltages are 0. V (solid line) and 0.05 V (dash-dotted line).

6 SCSIMU: A Package for Numerical Simulation of Semiconductor Devices with the Quasi-Hydrodynamic Model

In previous versions, the program was written in C using sparse solver routines that can be found in [9, 5] and in the Netlib library at <http://netlib.org/sparse/>. The latest version of the program was re-written in C++ without using any external packages. The key subroutines of the program can be found in Appendix A.

6.1 Developing platform, installation and usage

The program was developed on linux 2.0 systems with egcs C++ complier version 1.1 release, which can be downloaded from <http://egcs.cygnus.com/>.

The computation of initial approximations for the solution of system (1.1) (i.e. the solution of system (4.1)) typically takes just a few seconds on a Pentium II pro 233 MHz linux laptop. However, the time required to complete the second stage (i.e. the solution of the full system (1.9) varies dramatically, depending on the device the program simulates, the accuracy required and the applied voltage. For example, it takes about 50 seconds to achieve the convergence with the error smaller than 10^{-5} for the ballistic diode when the applied voltage is 1V. This time increases for the simulation of bipolar devices such as the PIN diode.

All subroutines, data files and TeX files are packed into one gzipped tar file scsimu-.xxx.tar.gz, where xxx stands for version number. The installation procedure is as follows:

- gunzip scsimu-.xxx.tar.gz
- tar -xvf scsimu-.xxx.tar
- edit configure.in and assign the name of your C++ compiler to the variable CXX
- change to the directory scsimu-xxxx and type command './configure'
- issue command 'make'.

If everything goes well, this creates an executable file scsimu in the current directory. The program supports a flexible worksheet environment which allows users to enter parameters in a simple way.

Executing scsimu -h prints a short help file about the package. The package supports three input methods:

1. user input;
2. worksheet input;
3. input only from one section of a worksheet.

For example, if we execute the program without supplying a worksheet, the program will ask for all required parameters interactively. The drawback of this method is that if you make a mistake in a previously entered parameter, you can not go back and correct it. A more user-friendly approach is the worksheet, which allows the user to create a data file for the simulation run of a specified device.

The first line of a new worksheet could be a comment line, like

```
# This worksheet is for the test device simulation with the SCSIMU.
```

Data plotting software typically ignore text after a # (if this is not the case you have to use another control symbol). Next, we have to specify parameters to be used by the SCSIMU in order to run a simulation. For this purpose we have proposed a "key-word" approach that requires the following format for the input

```
# <keyword> <value> <comments>
```

The program automatically assigns the value to the parameter represented by the keyword while ignoring any comment text placed after a numerical value. For example, if we wish to use 1000 discrete points in our simulation, we write:

```
#lattice_size 1000 You can add any comment here.
```

Other parameters are entered in a similar manner:

```
#window_size 10 This line is for the normalised length of the device  
#temperature 300 This line is for the lattice temperature  
#Comments after a number are acceptable  
# device maple  
# Don't put any comments after "maple" in the previous line  
#voltage 0.2 The applied voltage is 0.2 V  
#stage1_output test0.dat  
#stage2_output test1.dat  
#stage3_output test2.dat  
#end This is the last line of this worksheet.
```

As we have just seen we can add comments after the value if the parameter is an integer or a floating point number. This is not allowed if the parameter is a string (since the program can not identify which part of your string is needed). We demonstrated this situation in the above example when the type of the device was specified as "maple". Of course, we can add comments in between key-word lines. The order of keywords is not important, so you can list them in any order you like.

When the worksheet is created, we have to save it, for example, as example_maple.in (see Appendix A) and issue command scsimu example.in. The program will report what parameters were read and what values were assigned to them. We note that computation for different voltages can be easily implemented using the same worksheet file example_maple.in. What is required is to copy and paste the corresponding key-word lines between the newly created lines #begin maple_2 and #end maple_2 (see an example in Appendix A). Now, if you save the corrected version of the file and issue the command

```
scsimu example_maple.in maple_2,
```

the program will read all parameters enclosed within #begin maple_2 and #end maple_2.

The word “maple_2” in this example is referred to as the section name.

With the worksheet, we can keep all our simulation runs in one single file. Output file formats may vary from device to device. For example, for the PIN diode simulation we organise 3 output stages, where the output format of stages 1 and 2 includes $x, n, p, \varphi, T_n, T_p$, and the output format of stage 3 ($x, n, p, \varphi, T_n, T_p, v, M$) also includes the electron velocity and the Mach number. These outputs can be easily modified for the user’s specified device.

The package requires pre-defined keyword parameters that have to be entered by the user. All keywords must start with # and have the following format:

```
#keyword type{suggested value} [keyword notes]
```

By default we use 7 main parameters that are required to run the program:

1. #device string{maple,ballistic,PIN} [supported devices]
2. #temperature double{> 0} [temperature of the lattice]
3. #window_size double{> 0} [normalized length, usually set to 1]
4. #voltage double{} [applied voltage]
5. #stage1_output string
6. #stage2_output string
7. #stage3_output string

Other parameters are optional and can be used by changing our default values such as

1. #error_tolerance double{> 0, < 1} [absolute error tolerated in the program]
2. #lattice_size integer{> 100} [number of discrete points]

6.2 Core subroutines of the C++ program

In this section we provide details on core routines (including their functions and purposes) used in our C++ code.

1. int init(psi,n,p,Tn,Tp,F,h,x)
 - Purpose: to initialise various parameters and vectors including φ, n, p, T_n, T_p and to generate a numerical grid.
 - Arguments:
 - psi output (double *) <potential φ >;
 - n output (double *) <electron concentration>;
 - p output (double *) <hole concentration>;
 - Tn output (double *) <electron temperature T_n >;
 - Tp output (double *) <hole temperature T_p >;
 - F output (double *) <the recombination/generation/ionisation term>;
 - h output (double *) <grid step sizes>;
 - x output (double *) <grid point coordinates>.
 - Actions:

- $(Tn)_i$ and $(Tp)_i$ are set to 1.0 (in the normalised units);
- h_i are defined with respect to the device geometry;
- x_i are defined by the formula $x_{i+1} = x_i + h_i$ with $x_0 = 0$;
- n and p are set according to (4.4);
- psi is set according to (4.3).

2. void nextpsis(n,p,psi,x,h)

- Purpose: to solve the Poisson equation.
- Arguments
 - psi input/output (double *) <potential φ >;
 - n input/output (double *) <electron concentration>;
 - p input/output (double *) <hole concentration>;
 - h input (double *) <grid step sizes>;
 - x input (double *) <grid point coordinates>.
- Actions
 - Compute the Jacobian of the discretised Poisson equation (see (3.3), (3.4));
 - Solve the Poisson equation.

3. void nextnT_ns(n,p,psi,Tn,mun,h,x,t,tau,nn)

- Purpose: to solve the continuity equation for the electron subsystem.
- Arguments
 - psi input (double *) <potential φ >;
 - n input (double *) <electron concentration>;
 - nn output (double *) <electron concentration>;
 - p input (double *) <hole concentration>;
 - Tn input (double *) <electron temperature T_n >;
 - h input (double *) <grid step sizes>;
 - x input (double *) <grid point coordinates>;
 - t < is used only in the nonstationary case>;
 - tau < is used only in the nonstationary case>;
 - mun < is set to be a constant by default>.
- Actions
 - Compute the Jacobian of the discretised continuity equation for the electron subsystem (see (3.14));
 - Solve the continuity equation for the electron subsystem.

4. void nextpT_ps(n,p,psi,Tp,mup,h,x,t,tau,np)

- Purpose: to solve the continuity equation for the hole subsystem.
- Arguments
 - psi input (double *) <potential φ >;
 - n input (double *) <electron concentration>;
 - np output (double *) <hole concentration>;

- p input (double *) <hole concentration>;
- Tp input (double *) < electron temperature T_p >;
- h input (double *) <grid step sizes>;
- x input (double *) <grid point coordinates>.
- t < is used only in the nonstationary case>;
- tau < is used only in the nonstationary case>;
- mup < is set to be a constant by default>.
- Actions
 - Compute the Jacobian of the discretised continuity equation for the hole subsystem (see (3.18));
 - Solve the continuity equation for the hole subsystem.

5. double aci(x,t)

- Purpose: to define the doping profile function.
- Arguments
 - x (double *) <input> < the normalized x coordinate>;
 - t (double *) <input> <the normalised t coordinate>;
 - aci (double) <returns> <the normalized doping density>.
 - device (double, global)
- Actions
 - if the global variable device is 2, the ballistic diode profile is returned. Other values return the doping profile shown in Figure 1 (a). The user can easily define doping profiles of his/her choice.

6. double F(n,p)

- Purpose: to define the recombination function (see (1.8));
- Arguments
 - n (double *) <electron concentration>;
 - p (double *) <hole concentration>.

6.3 Utilities

1. void linpar(NN,A,B,C,in)

- Purpose: to solve the system of linear equations $\mathbf{X} \cdot \mathbf{out} = \mathbf{in}$, where \mathbf{X} is a band matrix of the form:

$$\begin{pmatrix} C & B & 0 & 0 & \dots \\ A & C & B & 0 & \dots \\ 0 & A & C & B & \dots \\ \dots & & & & \dots \end{pmatrix} \quad (6.1)$$

- Arguments
 - NN input (long) <the size of the matrix \mathbf{X} >;
 - A input (double*);
 - B input (double*);

- C input (double*);
- in input/output (double*) <computed results are stored in in>.

- Actions

- solve the equation $\mathbf{X} \cdot \mathbf{out} = \mathbf{in}$ and store the result out in in.

2. void haospsolve(NN,Jcob, out,in)

- Purpose: to solve the following equation $\mathbf{X} \cdot \mathbf{out} = \mathbf{in}$, where \mathbf{X} is a multi-band matrix of the form:

$$\begin{pmatrix} C_1 & B_1 & 0 & 0 & \dots & C_2 & B_2 & 0 & 0 & \dots \\ A_1 & C_1 & B_1 & 0 & \dots & A_2 & C_2 & B_2 & 0 & \dots \\ 0 & A_1 & C_1 & B_1 & \dots & 0 & A_2 & C_2 & B_2 & \dots \\ \dots & & & & & \dots & & & & \dots \\ C_3 & B_3 & 0 & 0 & \dots & C_4 & B_4 & 0 & 0 & \dots \\ A_3 & C_3 & B_3 & 0 & \dots & A_4 & C_4 & B_4 & 0 & \dots \\ 0 & A_3 & C_3 & B_3 & \dots & 0 & A_4 & C_4 & B_4 & \dots \\ \dots & & & & & \dots & & & & \dots \end{pmatrix} \quad (6.2)$$

- Arguments

- NN input (long) <the size of the matrix \mathbf{X} >.
- Jcob input (double**); the compact form of matrix (6.2);

$$\begin{pmatrix} A_1 & B_1 & C_1 & A_2 & B_2 & C_2 & A_3 & B_3 & C_3 & A_4 & B_4 & C_4 \\ A_1 & B_1 & C_1 & A_2 & B_2 & C_2 & A_3 & B_3 & C_3 & A_4 & B_4 & C_4 \\ A_1 & B_1 & C_1 & A_2 & B_2 & C_2 & A_3 & B_3 & C_3 & A_4 & B_4 & C_4 \\ A_1 & B_1 & C_1 & A_2 & B_2 & C_2 & A_3 & B_3 & C_3 & A_4 & B_4 & C_4 \\ \dots & & & & & & & & & & & \end{pmatrix} \quad (6.3)$$

- in input (double*) <the right-hand side vector>;
- computed results are stored into out.

- Actions

- solve the equation $\mathbf{X} \cdot \mathbf{out} = \mathbf{in}$.

6.4 Classes of the C++ program

1. QHDM.h <The main class>;
2. Bernoulli.h <Bernoulli functions>;
3. physics.h <Values of physical parameters used>;
4. device.h <“Library” class of different devices>;
5. ballistic.h <Ballistic diode>;
6. maple.h <Test (or “maple”) device>;
7. PIN.h <PIN device>;
8. linutil.h <Solver for sparse systems of linear equations>;
9. hutil.h <Some handy routines>;
10. worksheet.h <Dealing with worksheets>.

6.5 Core member functions of QHDH.h

1. `energy_electron()` <solving (3.20)>;
2. `energy_hole()` <solving (3.24)>;
3. `electron()` <solving (3.11)>;
4. `hole()` <solving (3.15)>;
5. `potential()` <solving (3.2)>;
6. `init()` <initial approximations>.

6.6 Graphical User Interface with pg2dplot

The output data files of simulation runs can be plotted using many interactive plotting programs including `gnuplot` and `xmgr`. We use the plotting program `pg2dplot`, which is a general purpose data visualization tool based on the PGPLT package.

The program reads a data file in which each line contains data in this format:

```
x1 x2 x3 x4 x5 x6 x7 ... .
```

Typing `pg2dplot -h` gives the following usage information:

```
pg2dplot: Plotting curves using pgplot.
Usage: pg2dplot -f data_file -xc m -yc n -x xlabel -y ylabel -t title
-xc m use the m-th column as x.
-yc n use the n-th column as y.
If no arguments are entered, the program uses the first column and the second column.
```

For example, after running program SCSIMU, we have a data file called “`test2.out`”. We wish to plot the temperature of electrons, label the y-axis with “`Tn`” and title “Temperature of electrons”. We therefore issue the command:

```
pg2dplot -f test2.out -yc 5 -y Tn -t "Temperature of electrons"
```

Options for the available Graphics devices/type can be checked by typing “?”. For example, typing “/xw” instructs the program to plot our curve in an X-window. For details of available options, refer to the manual of PGPLT at <http://astro.caltech.edu/~tjp/pgplot/>.

7 Testing Strategies and Examples

The package was extensively tested using a series of examples. This includes tests for each equation separately, as well as tests for the whole system. In this section we present results from two such test examples.

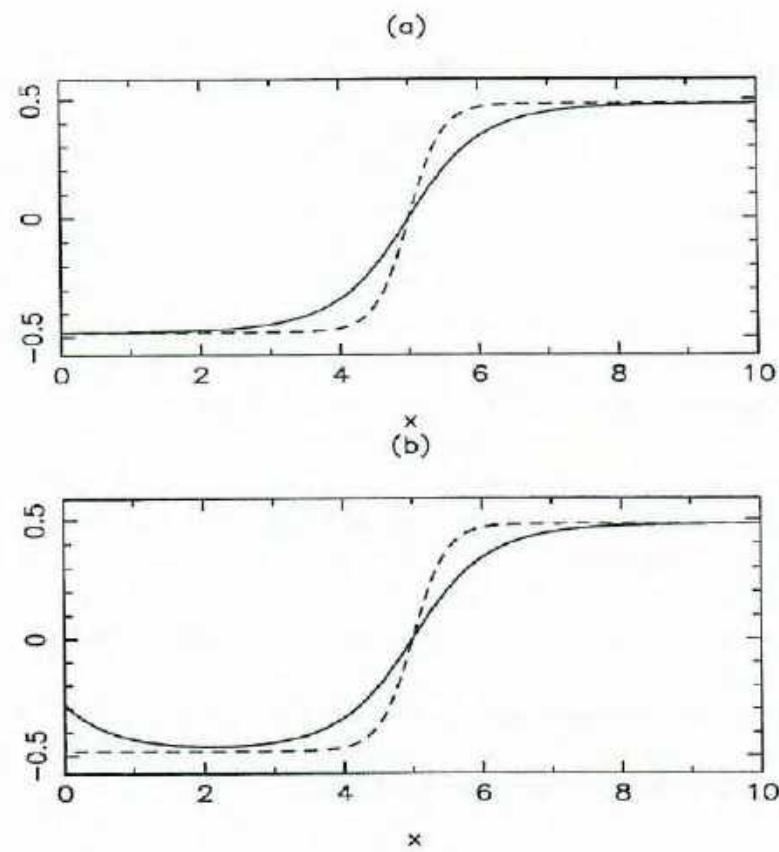


Figure 17: Test for the Poisson equation: applied voltages are (a) 0.0 V and (b) 0.2 V; the dashed lines are the initial approximations and the solid lines are the solutions of the Poisson equation.

7.1 Tests for each equation

We consider a semiconductor of length $L = 10$ (in the dimensionless length variable), which is doped with a concentration of electrically active impurities defined by [4]:

$$N(x) = \tanh\left(2L\left(\frac{x}{L} - \frac{1}{2}\right)\right). \quad (7.1)$$

We assume that the semiconductor is connected to an external potential $V(x)$ in such a way that $V(0) = V_0$ and $V(L) = 0$. Using the Boltzman statistics, the problem is reduced to the solution of the nonlinear Poisson equation

$$\frac{\partial^2 \varphi(x)}{\partial x^2} = e^{\varphi(x)} - e^{-\varphi(x)} - N(x). \quad (7.2)$$

with an external voltage V_0 given on the left-end and the grounded right-end. The results of computation for the initial approximations

$$\varphi(x) = \operatorname{asinh}(N(x)/2), \quad n(x) = 1, \quad p(x) = 1, \quad (7.3)$$

are presented in Figure 15. They are in a good agreement with the previously published results for this example [4]. Similar tests were developed for other equations of system (1.9).

7.2 Test of the whole system

Testing each equation of system (1.9) separately is an important component of our testing strategy. We also thoroughly tested system (1.9) as a whole, coupling all unknowns (φ, n, p, T_n, T_p). A series of test examples was developed for the stationary case:

$$\left\{ \begin{array}{l} \partial_{xx}\varphi = n - p - N, \\ -\partial_x J_n = F_n, \\ \partial_x J_p = F_p, \\ \partial_x Q_n = -J_n \partial_x \varphi + P_n, \\ \partial_x Q_p = -J_p \partial_x \varphi + P_p, \end{array} \right. \quad (7.4)$$

where

$$J_n = -n\mu_n \partial_x \varphi + \partial_x(T_n \mu_n n), \quad J_p = -p\mu_p \partial_x \varphi - \partial_x(T_p \mu_p p), \quad (7.5)$$

$$Q_n = \beta_n T_n n \mu_n \partial_x \varphi - \beta_n \partial_x[T_n D_n n], \quad Q_p = -\beta_p T_p p \mu_p \partial_x \varphi - \beta_p \partial_x[T_p D_p p]. \quad (7.6)$$

Assuming that

$$\left\{ \begin{array}{l} \bar{\varphi}(x) = 10(x^4/12 - x^3/6) + x - 5 + 1, \\ \bar{n}(x) = -45x^3 + 80x^2 - 35x + 2, \\ \bar{p}(x) = 35x^3 - 45x^2 + 10x + 2, \\ \bar{T}_n(x) = \sin(x^4) + 1, \\ \bar{T}_p(x) = \cos(x^4) + 1, \end{array} \right. \quad (7.7)$$

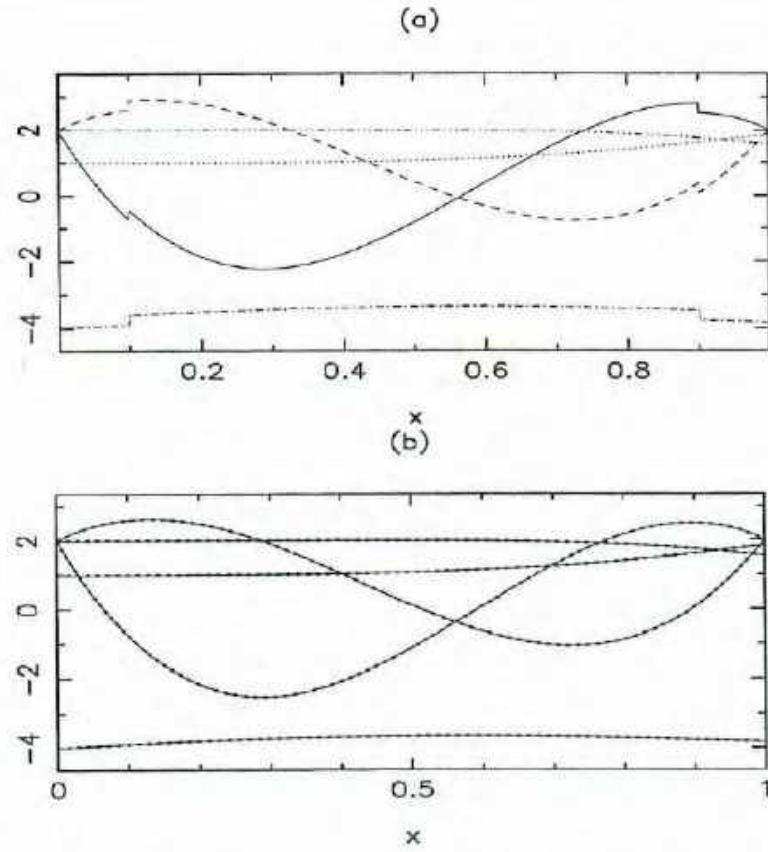


Figure 18: Test for system (1.9): (a) initial approximations ($\varphi'(x)$ - dash-dotted line; $n'(x)$ - solid line; $p'(x)$ - dashed line; $T'_n(x)$ - dotted line; $T'_p(x)$ - dash-dot-dotted line); (b) program outputs (solid lines) coincide with the solution (dotted lines).

we define F_n , F_p , P_n , and P_p by substituting (7.7) into (7.4). Then, we aim at the numerical solution of system (7.4) with the just defined functions F_n , F_p , P_n , and P_p . As the initial approximations we take the following functions

$$\begin{cases} \varphi'(x) = \bar{\varphi}(x) + A_1\delta(x), & n'(x) = \bar{n}(x) + A_2\delta(x), & p'(x) = \bar{p}(x) + A_3\delta(x), \\ T'_n(x) = \bar{T}_n(x) + A_4\delta(x), & T'_p(x) = \bar{T}_p(x) + A_5\delta(x), \end{cases} \quad (7.8)$$

where the perturbation function $\delta(x)$ and parameters A_i , $i = 1, \dots, 5$ varies from experiment to experiment. For example, when $A_i = 3$, $i = 1, 2, 3$, $A_j = 0.1$, $j = 4, 5$ and

$$\delta(x) = \begin{cases} 0.1, & 0.1 \leq x \leq 0.9, \\ 0, & \text{otherwise}, \end{cases} \quad (7.9)$$

the result of computation are shown in Fig. 18. These experiments, performed for a wide range of perturbation parameters and different spacial patterns for $\delta(x)$, demonstrated robustness and reliability of the program.

8 Acknowledgements.

Authors were supported by grant USQ-PTRP 17989 and by Australian Research Council Small Grant 17906. We thank Tim Passmore for his assistance at the final stage of preparation of this paper.

References

- [1] Apanovich, Y., Lyumkis, E., Polksy et al, Steady-State and Transient Analysis of Sub-micron Devices Using Energy Balance and Simplified Hydrodynamic Models, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 13, No. 6, 702–711, 1994.
- [2] C.L. Gardner, J.W. Jerome and D.J. Rose, Numerical Methods for the Hydrodynamic Device Model: Subsonic Flow, *IEEE Transactions on Computer-Aided Design*, Vol. 8, No. 5, 501–507, 1989.
- [3] Kakati, D., Ramanan, C. and Ramamurthy, V., Numerical Analysis of electrophysical characteristics of semiconductor devices accounting for the heat transfer, in *NEMA-CODE IV: Proceedings of the Fourth International Conference on the Numerical Analysis of Semiconductor Devices and Integrated Circuits: Trinity College, Dublin, Ireland*, Ed. J.J.H. Miller, Bool Press, 1985, 326–331.
- [4] Klvana, F., The Internal Field in Semiconductors, in *Solving Problems in Scientific Computing Using MAPLE and MATLAB*, Eds. W. Gander and J. Hrebicek, 1993, 59–67.
- [5] Kundert, K., *Sparse Matrix Techniques*, in *Circuit Analysis, Simulation and Design*, Ed. Albert Ruehli, North-Holland, 1986.
- [6] Melnik, R.V.N. & Melnik, K.N., Modelling of Nonlocal Physical Effects in Semiconductor Plasma Using Quasi-Hydrodynamic Models, *Computational Techniques and Applications: CTAC97*, Eds. J. Noye, M. Teubner, A. Gill, World Scientific, 1998, 441–448.
- [7] A.A. Samarskii, *Theorie der Differenzenverfahren*, (Series: Mathematik und ihre Anwendungen in Physik und Technik), Aufl. Leipzig: Akademische Verlagsgesellschaft Geest & Portig, 1984.
- [8] A.A. Samarskii and E.S. Nikolaev, *Numerical Methods for Grid Equations*, Basel, Boston: Birkhauser Verlag, 1989.
- [9] Vetterling, W.T. et al, *Numerical Recipes Example Book (C)*, Cambridge University Press, 1994.

A Appendix

The package is made available with no warranty whatsoever and readers are free to use and modify the package at their own risk. In addition to these programs you may need automake and autoconf as well as some installation programs available under the terms of the GNU General Public License as published by the Free Software Foundation, Inc.

A.1 Bernoulli.h

```
#ifndef __BERNOULLI_H__
#define __BERNOULLI_H__


/* The Bernoulli function is defined as
B(x)=x/(exp(x)-1)
*/

#include <math.h>
    const double _x1=-4.505456673639644e+01;
    const double _x2=-6.713653614580113e-07;
    const double _x3=6.713653614580113e-07;
    const double _x4=4.505456673639644e+01;
    const double _x5=1.139949853148886e+04;

class Bernoulli
{
//    const double _x1,_x2,_x3,_x4,_x5;

public:
    Bernoulli();
    ~Bernoulli();

    double B(double x);
    double dB(double x);
};

Bernoulli::Bernoulli()
{
}

double Bernoulli::B(double x)
{
    if (x<_x1) {return -x;}
    if (x<_x2) {return x/(exp(x)-1.0);}
    if (x<_x3) {return 1.0-x/2.0;}
    if (x<_x4) {return x*exp(-x)/(1.0-exp(-x));}
```

```
if (x<_x5) {return x*exp(-x);}
return 0.0;
}

double Bernoulli::dB(double x)
{
if (x<_x1) return -1.0;
if (x<-1.e-5) return (exp(x)-1.0-x*exp(x))/(exp(2.0*x)-2.0*exp(x)+1.0);
if (x<1.e-5) return -0.5+x/6.0-x*x*x/180.+x*x*x*x*x/5040.0;
if (x<_x3) return (exp(-x)-exp(-2.0*x)-x*exp(-x))/(1.0+exp(-2.0*x)-2.0*exp(-x));
if (x<_x4) return exp(-x)-x*exp(-x);
return 0.0;
}

#endif

A.2 PIN.h

#ifndef __PIN__DEVICE__
#define __PIN__DEVICE__

#include "physics.h"
#include "device.h"

class PIN: public device
{
public:
    PIN();
    ~PIN();

    double N(double );
};

PIN::PIN()
{
    _L=3.5 micro meter;
    _id=1;
}

PIN::~PIN()
{
}

double PIN::N(double x)
{
double c1=0.5 micro meter;
```

```
double c2=2.5 micro meter;
if (x<=c1)           return -1.0e18 per_cube_cent_meter;
if ((x>c1)&&(x<c2)) return 3.5e14 per_cube_cent_meter;
return 2.4e19 per_cube_cent_meter;
}

#endif

.A.3 QHDM.h

#ifndef _QHDM_H_
#define _QHDM_H_


#include <iostream>
#include <iomanip>
#include <algorithm>
#include <vector>
#include <string>
#include <worksheet.h>
#include <hutil.h>
#include <math.h>
#include <numeric>
#include "physics.h"
#include "device.h"
#include "PIN.h"
#include "ballistic.h"
#include "maple.h"
#include "linutil.h"
#include "Bernoulli.h"
#include "scsimu.h"

class QHDM
{
private:
    int _lattice_size,_max_iters,_prog1_iters,_prog2_iters;
    double _tolerance;

    double _window_size,_xb,_xe,_L;
    vector<double> _x,_phi,_n,_p,_Tn,_Tp,_h,_Jn,_Jp,_vn;
    vector<double> _A,_B,_C;
    double _t_cr,_phi_cr,_n_cr,_T_cr,_F_cr,_mu_cr,_D_cr,_Q_cr;
    double _v_cr;
    double _T,_L_cr;
```

```
double _voltage;

vector<double> _delta,_F;

device *_dev;
int _device;
void create();

Bernoulli *_Ber;

int _debug;

public:
QHDM();
QHDM(worksheet *ws);
~QHDM();

vector<double> & x() {return _x;}
vector<double> & phi() {return _phi;}
vector<double> & n() {return _n;}
vector<double> & p() {return _p;}
vector<double> & Tn() {return _Tn;}
vector<double> & Tp() {return _Tp;}
int ls() {return _lattice_size;}
double ws() {return _window_size;}
void init();
void save(const char *fn) {int i=_lattice_size/100;
    if (i==0) i=1;save(fn,i,10);}
void save(const char *fn, int n) {save(fn,n,10);}
void save(const char *, int , int );
double potential();
double electron();
double hole();
void debug() {_debug=1;}
double N(double & x);
double f(double x);
double f1(double x);
double df(double x);
double df1(double x);
double F(double n,double p);
double JFn(double n,double p);
double JFp(double n,double p);
double prog1();
void prog2();
double prog11();
double prog12();
```

```

double prog3();
// 
// double Dn(double & T) {return mun0*T*sqrt(T)/_mu_cr;}
// double Dp(double & T) {return mup0*T*sqrt(T)/_mu_cr;}
// double dDn(double & T) {return 1.5*sqrt(T)*mun0/_mu_cr;}
// double dDp(double & T) {return 1.5*sqrt(T)*mup0/_mu_cr;}
// double mun(double & T) {return (mun0/_mu_cr)*sqrt(T);}
// double dmun(double & T) {return 0.5*(mun0/_mu_cr)/sqrt(T);}
// double mup(double & T) {return (mup0/_mu_cr)*sqrt(T);}
// double dmup(double & T) {return 0.5*(mup0/_mu_cr)/sqrt(T);}
// double Dn(double & T) {return mun0*T/_mu_cr;}
// double Dp(double & T) {return mup0*T/_mu_cr;}
// double dDn(double & T) {return mun0/_mu_cr;}
// double dDp(double & T) {return mup0/_mu_cr;}
// double mun(double & T) {return (mun0/_mu_cr);}
// double dmun(double & T) {return 0.;}
// double mup(double & T) {return (mup0/_mu_cr);}
// double dmup(double & T) {return 0.;}
double fir(int i,vector<double> & Tn)
    {return 2.0*(_phi[i]-_phi[i-1])/(Tn[i]+Tn[i-1]);}
double fir(double beta, int i, vector<double> & Tn)
    {return 2.0*((1.0+beta)/beta)*
        (_phi[i]-_phi[i-1])/(Tn[i]+Tn[i-1]);}
double dfir(double beta, int i, vector<double> & Tn)
    {return -2.0*((1.0+beta)/beta)*
        (_phi[i]-_phi[i-1])/((Tn[i]+Tn[i-1])*(Tn[i]+Tn[i-1]));}
double energy_electron();
double energy_electron(vector<double> &, vector<double> &);
double energy_electron2();
double energy_hole();
void Jn();
void velocity();
};

QHDM::QHDM()
{
    _lattice_size=1000;
    _window_size=1.0;
    _tolerance=1.e-10;
    _device=1;
    _dev=new PIN();
    create();
}

QHDM::QHDM(worksheet *s1)
{
    _tolerance=1.e-5;
}

```

```
_lattice_size=1000;
_t_cr=1.e-9;
_max_iters=400;
_prog1_iters=80;
_prog2_iters=80;
_window_size=1.0;
string device_name;

s1->value("#lattice_size",_lattice_size,0);
s1->value("#window_size",_window_size,0);
s1->value("#temperature",_T);
s1->value("#t_cr",_t_cr,0);
s1->value("#error_tolerance",_tolerance,0);
s1->value("#voltage",_voltage);
s1->value("#device",device_name);
s1->value("#max_iters",_max_iters,0);
s1->value("#prog1_iters",_prog1_iters,0);
s1->value("#prog2_iters",_prog2_iters,0);

_dev=new device();
if (device_name=="maple") _dev=new maple();
if (device_name=="PIN") _dev=new PIN();
if (device_name=="ballistic") _dev=new ballistic();

_device=_dev->id();
if (_device==0) {cout<<"Sorry, that device is not supported yet."
    <<endl<<"Currently supported devices are: PIN, ballistic or maple."
    <<endl<<"For example:
#define PIN to simulate the PIN device. Bye"<<endl;exit(1);}

create();
}

QHDM::~QHDM()
{
}

void QHDM::create()
{
_x =vector<double>(_lattice_size);
_h =vector<double>(_lattice_size);
_p =vector<double>(_lattice_size);
_n = vector<double>(_lattice_size);
_Tn = vector<double>(_lattice_size);
_Tp = vector<double>(_lattice_size);
```

```
_Jn = vector<double>(_lattice_size);
_Jp = vector<double>(_lattice_size);
_phi= vector<double>(_lattice_size);
_vn = vector<double>(_lattice_size);
_A = vector<double>(_lattice_size);
_B = vector<double>(_lattice_size);
_C = vector<double>(_lattice_size);
_F=_delta=_C;

_debug=0;

_xb=0;_xe=_xb+_window_size;
coordinate(_x,_xb,_xe,_lattice_size);
for (int i=0;i<_lattice_size-1;i++) _h[i]=_x[i+1]-_x[i];
_h[_lattice_size-1]=_h[_lattice_size-2];

_T_cr=_T*Kb;
_L=_dev->length();
_L_cr=_L/_window_size;
_n_cr=epsilon0*Epsilon1*_T_cr/(_L_cr*_L_cr*charge_of_electron*charge_of_electron);
_phi_cr=_T_cr/charge_of_electron;
_t_cr=1.e-10;
_mu_cr=charge_of_electron*_L_cr*_L_cr/(_T_cr*_t_cr);
_v_cr=sqrt(_mu_cr*_T_cr/(charge_of_electron*_t_cr));
//double vtmpcr=_mu_cr*_phi_cr/_L_cr;
//cout <<vtmpcr << endl;exit(1);
cout <<"Here are the predefined scaling factors:\n"
     <<"L=<<_L<<"meters"<<endl
     <<"L_cr=<<_L_cr<<endl
     <<"n_cr=<<_n_cr<<"per cube meter"<<endl
     <<"T_cr=<<_T_cr<<"J"<<endl
     <<"phi_cr=<<_phi_cr<<"volt"<<endl
     <<"t_cr=<<_t_cr<<"second"<<endl
     <<"error_tolerance=<<_tolerance<<endl;

_Ber=new Bernoulli();

init();

}

void QHDM::save(const char *fn,int n,int p)
{
string space("    ");
ofstream outf(fn);
outf.setf(ios::scientific);
```

```

        outf.precision(p);
        Jn(); velocity();
        outf<<"### x_coordinate n p Tn Tp N #####"<><endl;
        for (int i=0;i<_lattice_size;i+=n)
            outf<<_x[i]*_L_cr<<space<<_n[i]<<space<<_p[i]<<space<<_phi[i]<<space
//             <<_Tn[i]*_T_cr/Kb<<space<<_Tp[i]<<space<<-_Jn[i]/_n[i]
            <<space<<_Jn[i]<<space<<N(_x[i])*_n_cr<<endl;
            <<_Tn[i]*_T_cr/charge_of_electron
            <<space<<_Tp[i]*_T_cr/charge_of_electron<<space<<-_Jn[i]*_v_cr/_n[i]
            <<space<<_vn[i]<<endl;
//             <<(_x[i]-0.5)*10.0<<space<<_F[i]<<space<<_A[i]<<space<<_B[i]
//             <<space<<_C[i]<<endl;
        }

void QHDM::init()
{
    _debug=0;
    double nieo=nie/_n_cr;
    if (_device==3) nieo=1.0;
    for (int i=0;i<_lattice_size;i++)
    {
        double tN=N(_x[i]);
//         cout <<tmp<<" "<<N<<" "<<_dev->N(tmp)<<endl;
        _phi[i]=sign(tN)*log(fabs(tN)/nieo);

        _n[i]=nieo*exp(_phi[i]);
        _p[i]=nieo*exp(-_phi[i]);
        _Tn[i]=1.0;
        _Tp[i]=1.0;
        if (_device==3) {_phi[i]=asinh(tN*_n_cr/2.0); _n[i]=1.0; _p[i]=1.0;}
        if (_debug)
        {
            cout <<"Debuging ..."\><endl;
            double tmp=(_x[i]-0.5)*10.0;
            _n[i]=exp(-tmp*tmp);
            _p[i]=exp(-tmp*tmp);
            _phi[i]=exp(-2.0*tmp*tmp);
            _Tn[i]=exp(-3.0 *tmp*tmp);
            _Tp[i]=exp(-3.0 *tmp*tmp);
        }
    }
}

if (_debug) return;
double nies=nieo*nies;

```

```
if (_device!=3) for (int n=1;n<_lattice_size;n++){
    if (_x[n]>.7)
    {
        double NN=N(_x[n]);
        _n[n]=sqrt(NN*NN/4.0+nies)+NN/2.0;
        _p[n]=sqrt(NN*NN/4.0+nies)-NN/2.0;
        _phi[n]=_voltage/_phi_cr+sign(NN)*log(fabs(NN)/
            (2.0*nies)+sqrt(NN*NN/(4.0*nies)+1.0));
    }
}

if (_device==3) {_phi[0]+=_voltage; return;}
// left boundary

int n=0;
double NN=N(_x[n]);
_n[n]=sqrt(NN*NN/4.0+nies)+NN/2.0;
_p[n]=sqrt(NN*NN/4.0+nies)-NN/2.0;
_phi[n]=sign(NN)*log(fabs(NN)/(2.0*nies)+sqrt(NN*NN/(4.0*nies)+1.0));

}

double QHDM::N(double & x)
{
    return _dev->N(x*_L_cr)/_n_cr;
}

double QHDM::energy_electron()
{
    if (_dev->id()==3) return 0;
    int n=_lattice_size;
    _B[0]=_B[n-1]=0.0;
    _A[0]=_A[n-1]=0.0;
    _C[0]=_C[n-1]=1.0;
    _delta[0]=_delta[n-1]=0.0;

    int i;
    vector<double> nn=_n,nTn=_Tn;
    vector<double>::const_iterator iter_max,iter_min;
    double init_error;
    double lambda=0.05;
    if (_dev->id()!=1) lambda=1.0;
    for (int iter=0;iter<400;iter++)

```

```

{
for(i=1;i<n-1;i++)
{
    double hp=(_h[i]+_h[i+1])/2.0;
    double Av= Betan* Dn(nTn[i-1])* f(fir(Betan,i,nTn))/_h[i ];
    double dAv=Betan*dDn(nTn[i-1])* f(fir(Betan,i,nTn))/_h[i ]+
        Betan* Dn(nTn[i-1])*df(fir(Betan,i,nTn))*dfir(Betan,i,nTn)/_h[i ];
    double dAvi=Betan*Dn(nTn[i-1])*df(fir(Betan,i,nTn))*dfir(Betan,i,nTn)/_h[i ];
    double Bv= Betan* Dn(nTn[i+1])*f1(fir(Betan,i+1,nTn))/_h[i+1];
    double dBv=Betan*dDn(nTn[i+1])*f1(fir(Betan,i+1,nTn))/_h[i+1] +
        Betan* Dn(nTn[i+1])*df1(fir(Betan,i+1,nTn))*dfir(Betan,i+1,nTn)/_h[i+1];
    double dBvi=Betan*Dn(nTn[i+1])*df1(fir(Betan,i+1,nTn))*dfir(Betan,i+1,nTn)/_h[i+1];
    double Cv= Betan* Dn(nTn[i])* f(fir(Betan,i+1,nTn))/_h[i+1] +
        Betan* Dn(nTn[i])*f1(fir(Betan,i ,nTn))/_h[i];
    double dCv=Betan*dDn(nTn[i])*f(fir(Betan,i+1,nTn))/_h[i+1] +
        Betan* Dn(nTn[i])*df(fir(Betan,i+1,nTn))*dfir(Betan,i+1,nTn)/_h[i+1] +
        Betan*dDn(nTn[i])*f1(fir(Betan,i,nTn))/_h[i] +
        Betan* Dn(nTn[i])*df1(fir(Betan,i,nTn))*dfir(Betan,i,nTn)/_h[i];
    double dCva=Betan* Dn(nTn[i])*df1(fir(Betan,i,nTn))*dfir(Betan,i,nTn)/_h[i];
    double dCvb=Betan* Dn(nTn[i])*df(fir(Betan,i+1,nTn))*dfir(Betan,i+1,nTn)/_h[i+1];

    double phixx=((_phi[i+1]-_phi[i])/_h[i+1]-
        (_phi[i]-_phi[i-1])/_h[i])/hp;
    double phix=(_phi[i+1]-_phi[i-1])/(2.0*hp);
    _delta[i]=0.0;

//    double tau_omega=0.5*mass_of_electron*mun0/
//        (charge_of_electron*_t_cr*_Tn[i])
//        +1.5*mun0*_T_cr/
//            (velocity_saturation*velocity_saturation*_t_cr*
//            charge_of_electron)*(_Tn[i]/(_Tn[i]+1.0));
//    double tau_omega=1.5*mun0*_T_cr/(_t_cr*charge_of_electron*
//        velocity_saturation*velocity_saturation)*sqrt(_Tn[i]);
    double tau_omega=0.4e-12/_t_cr;
    _F[i]= Av*nTn[i-1]*nn[i-1]
        +Bv*nTn[i+1]*nn[i+1]
        -Cv*nTn[i ]*nn[i]
}

```

```

+hp*( mun(nTn[i])*nTn[i]*nn[i]*phixx
      +nn[i]*mun(nTn[i])*phix*phix
      +(1.0-nTn[i])*nn[i]/tau_omega);
_F[i]=-_F[i]*lambda;
_A[i]=Av*nn[i-1]+dAv*nTn[i-1]*nn[i-1]-dCva*nTn[i]*nn[i];
_C[i]=-Cv*nn[i]-dCv*nTn[i]*nn[i]+dAvi*nTn[i-1]*nn[i-1]-
    dBvi*nTn[i+1]*nn[i+1]+hp*(
        (dmun(nTn[i])*nTn[i]+mun(nTn[i]))*nn[i]*phixx
        +dmun(nTn[i])*nn[i]*phix*phix-nn[i]/tau_omega);
_B[i]=Bv*nn[i+1]+dBv*nTn[i+1]*nn[i+1]-dCvb*nTn[i]*nn[i];
//if (_debug) {_A[i]=Av; _B[i]=Bv; _C[i]=Cv;}
}
if (_debug) return 0.0;
tridag(_A,_C,_B,_F,_delta);
for (i=1;i<n-1;i++)
{
//if (_delta[i]<1.e-1) nn[i]=nie/_n_cr*
    pow(_n_cr*nn[i]/nie,nTn[i]/(nTn[i]+_delta[i]));
    nTn[i]+=_delta[i];
}
//iter_max=max_element(_F.begin(),_F.end());
//iter_min=min_element(_F.begin(),_F.end());
//double error=*iter_max-*iter_min; if (error==0) error=*iter_max;
double error=max_error(_F);
if (iter==0) init_error=error;
//if (_debug)
//cout <<"energy electron Error=<<error<<" <<iter<<endl;
if (error<_tolerance) break;
if (max_error(_delta)<1.e-11) {error=0.0;cout<"can not improve much\n";break;}
}

_Tn=nTn;
return fabs(init_error);
}

double QHDM::energy_electron2()
{
vector<double> nn=_n,tmpTn,nTn=_Tn;
double olderror,error,lambda=1.0;

olderror=error=energy_electron(nTn,nn);

cout <<"Error=<< error <<endl;

tmpTn=nTn;

```

```

for (;;) {
    error=energy_electron(nTn,nn);
    cout << "olderror=" << olderror << " new error" << error << endl;
    if (error>olderror) {potential();energy_electron(nTn,nn);
        lambda*=0.95;nTn=tmpTn;cout << "No luck try lambda=" << lambda << endl;}
        else {lambda=1.0;tmpTn=nTn;olderror=error;}
    if (error<_tolerance) break;
    if (lambda<0.01) {lambda=1.0;tmpTn=nTn;olderror=error;} // give up
    tridag(_A,_C,_B,_F,_delta);

    for (int i=0;i<_lattice_size;i++) nTn[i]+=_delta[i]*lambda;

}
_Tn=nTn;
return error;
}

double QHDM::energy_electron(vector<double> & nTn, vector<double> & nn)
{
int n=_lattice_size;
_B[0]=_B[n-1]=0.0;
_A[0]=_A[n-1]=0.0;
_C[0]=_C[n-1]=1.0;
_delta[0]=_delta[n-1]=0.0;

int i;

for(i=1;i<n-1;i++)
{
    double hp=(_h[i]+_h[i+1])/2.0;
    double Av= Betan* Dn(nTn[i-1])* f(fir(Betan,i,nTn))/_h[i];
    double dAv=Betan*dDn(nTn[i-1])* f(fir(Betan,i,nTn))/_h[i] +
        Betan* Dn(nTn[i-1])*df(fir(Betan,i,nTn))* dfir(Betan,i,nTn)/_h[i];
    double dAvi=Betan*Dn(nTn[i-1])*df(fir(Betan,i,nTn))* dfir(Betan,i,nTn)/_h[i];
    double Bv= Betan* Dn(nTn[i+1])*f1(fir(Betan,i+1,nTn))/_h[i+1];
    double dBv=Betan*dDn(nTn[i+1])*f1(fir(Betan,i+1,nTn))/_h[i+1] +
        Betan* Dn(nTn[i+1])*df1(fir(Betan,i+1,nTn))* dfir(Betan,i+1,nTn)/_h[i+1];
    double dBvi=Betan*Dn(nTn[i+1])*df1(fir(Betan,i+1,nTn))* dfir(Betan,i+1,nTn)/_h[i+1];
    double Cv= Betan* Dn(nTn[i])* f(fir(Betan,i+1,nTn))/_h[i+1] +
        Betan* Dn(nTn[i])*f1(fir(Betan,i,nTn))/_h[i];
}

```

```

double dCv=Betan*dDn(nTn[i])*f(fir(Betan,i+1,nTn))/_h[i+1] +
    Betan* Dn(nTn[i])*df(fir(Betan,i+1,nTn))* dfir(Betan,i+1,nTn)/_h[i+1] +
    Betan*dDn(nTn[i])*f1(fir(Betan,i,nTn))/_h[i] +
    Betan* Dn(nTn[i])*df1(fir(Betan,i,nTn))* dfir(Betan,i,nTn)/_h[i];
double dCva=Betan* Dn(nTn[i])*df1(fir(Betan,i,nTn))* dfir(Betan,i,nTn)/_h[i];
double dCvb=Betan* Dn(nTn[i])*df(fir(Betan,i+1,nTn))* dfir(Betan,i+1,nTn)/_h[i+1];

double phixx=(_phi[i+1]-_phi[i])/_h[i+1]-
    (_phi[i]-_phi[i-1])/_h[i])/hp;
double phix=(_phi[i+1]-_phi[i-1])/(2.0*hp);
_delta[i]=0.0;

// double tau_omega=0.5*mass_of_electron*mun0/
//   (charge_of_electron*_t_cr*_Tn[i])
//   +1.5*mun0*_T_cr/(velocity_saturation*
//   velocity_saturation*_t_cr*charge_of_electron)*
//   (_Tn[i]/(_Tn[i]+1.0));
// double tau_omega=1.5*mun0*_T_cr/(_t_cr*charge_of_electron*
//   velocity_saturation*velocity_saturation)*sqrt(_Tn[i]);
double tau_omega=0.4e-12/_t_cr;
_F[i]= Av*nTn[i-1]*nn[i-1]
+Bv*nTn[i+1]*nn[i+1]
-Cv*nTn[i ]*nn[i]
+hp*( mun(nTn[i])*nTn[i]*nn[i]*phixx
+nn[i]*mun(nTn[i])*phix*phix
+(1.0-nTn[i])*nn[i]/tau_omega);

_F[i]=-_F[i];
_A[i]=Av*nn[i-1]+dAv*nTn[i-1]*nn[i-1]-dCva*nTn[i ]*nn[i];
_C[i]=-Cv*nn[i]-dCv*nTn[i ]*nn[i]+dAvi*nTn[i-1]*nn[i-1]+
dBvi*nTn[i+1]*nn[i+1]+hp*(
(dmun(nTn[i])*nTn[i]+mun(nTn[i]))*nn[i]*phixx
+dmun(nTn[i])*nn[i]*phix*phix-nn[i]/tau_omega);
_B[i]=Bv*nn[i+1]+dBv*nTn[i+1]*nn[i+1]-dCvb*nTn[i ]*nn[i];
//if (_debug) {_A[i]=Av; _B[i]=Bv; _C[i]=Cv; }
}

return inner_product(_F.begin(),_F.end(),_F.begin(),0.0);
}

double QHDM::energy_hole()
{
if (_dev->id()==3) return 0;
}

```

```
int n=_lattice_size;
_B[0]=_B[n-1]=0.0;
_A[0]=_A[n-1]=0.0;
_C[0]=_C[n-1]=1.0;
_delta[0]=_delta[n-1]=0.0;

int i;
vector<double> np=_p,nTp=_Tp;
vector<double>::const_iterator iter_max,iter_min;
double init_error;

for (int iter=0;iter<100;iter++)
{
    for(i=1;i<n-1;i++)
    {
        double hp=(_h[i]+_h[i+1])/2.0;
        double Av= Betap* Dp(nTp[i-1])* f1(fir(Betap,i,nTp))/_h[i];
        double dAv=Betap*dDp(nTp[i-1])* f1(fir(Betap,i,nTp))/_h[i] +
                    Betap* Dp(nTp[i-1])*df1(fir(Betap,i,nTp))* dfir(Betap,i,nTp)/_h[i];
        double dAvi=Betap*Dp(nTp[i-1])*df1(fir(Betap,i,nTp))* dfir(Betap,i,nTp)/_h[i];
        double Bv= Betap* Dp(nTp[i+1])*f(fir(Betap,i+1,nTp))/_h[i+1];
        double dBv=Betap*dDp(nTp[i+1])*f(fir(Betap,i+1,nTp))/_h[i+1] +
                    Betap* Dp(nTp[i+1])*df(fir(Betap,i+1,nTp))* dfir(Betap,i+1,nTp)/_h[i+1];
        double dBvi=Betap*Dp(nTp[i+1])*df(fir(Betap,i+1,nTp))* dfir(Betap,i+1,nTp)/_h[i+1];
        double Cv= Betap* Dp(nTp[i])* f1(fir(Betap,i+1,nTp))/_h[i+1] +
                    Betap* Dp(nTp[i])*f(fir(Betap,i,nTp))/_h[i];
        double dCv=Betap*dDp(nTp[i])*f1(fir(Betap,i+1,nTp))/_h[i+1] +
                    Betap* Dp(nTp[i])*df1(fir(Betap,i+1,nTp))* dfir(Betap,i+1,nTp)/_h[i+1];
        Betap*dDp(nTp[i])*f(fir(Betap,i,nTp))/_h[i] +
        Betap* Dp(nTp[i])*df(fir(Betap,i,nTp))* dfir(Betap,i,nTp)/_h[i];
        double dCva=Betap* Dp(nTp[i])*df(fir(Betap,i,nTp))* dfir(Betap,i,nTp)/_h[i];
        double dCvb=Betap* Dp(nTp[i])*df1(fir(Betap,i+1,nTp))* dfir(Betap,i+1,nTp)/_h[i+1];

        double phixx=((_phi[i+1]-_phi[i])/_h[i+1]-
                      (_phi[i]-_phi[i-1])/_h[i])/hp;
        double phix=(_phi[i+1]-_phi[i-1])/(2.0*hp);
        _delta[i]=0.0;
```

```

//      double tau_omega=0.5*mass_of_electron*mup0/
//          (charge_of_electron*_t_cr*_Tp[i])
//      +1.5*mup0*_T_cr/(velocity_saturation*
//          velocity_saturation*_t_cr*charge_of_electron)*
//          (_Tp[i]/(_Tp[i]+1.0));
double tau_omega=1.5*mup0*_T_cr/
    (_t_cr*charge_of_electron*velocity_saturation*
        velocity_saturation)*sqrt(_Tp[i]);
//      double tau_omega=1.0;
//      double tau_omega=0.4e-12/_t_cr;
_F[i]= Av*nTp[i-1]*np[i-1]
    +Bv*nTp[i+1]*np[i+1]
    -Cv*nTp[i ]*np[i]
    +hp*(-mup(nTp[i])*nTp[i]*np[i]*phixx
        +np[i]*mup(nTp[i])*phix*phix
        +(1.0-nTp[i])*np[i]/tau_omega);
_F[i]=-_F[i];
_A[i]=Av*np[i-1]+dAv*nTp[i-1]*np[i-1]-dCva*nTp[i ]*np[i];
_C[i]=-Cv*np[i]-dCv*nTp[i ]*np[i]+dAvi*nTp[i-1]*np[i-1]-
    dEvi*nTp[i+1]*np[i+1]
    +hp*(-dmup(nTp[i])*nTp[i]-mup(nTp[i]))*np[i]*phixx
    +dmup(nTp[i])*np[i]*phix*phix-np[i]/tau_omega;
_B[i]=Bv*np[i+1]+dBv*nTp[i+1]*np[i+1]-dCvb*nTp[i ]*np[i];
//if (_debug) {_A[i]=Av;_B[i]=Bv;_C[i]=Cv;}
}
if (_debug) return 0.0;
tridag(_A,_C,_B,_F,_delta);
for (i=1;i<n-1;i++)
{
//if (_delta[i]<1.e-1) np[i]=nie/_n_cr*pow(_n_cr*np[i]/nie,nTp[i]/
    (nTp[i]+_delta[i]));
    nTp[i]+=_delta[i];
}
//iter_max=max_element(_F.begin(),_F.end());
//iter_min=min_element(_F.begin(),_F.end());
//double error=*iter_max-*iter_min; if (error==0) error=*iter_max;
double error=max_error(_F);
if (iter==0) init_error=error;
if (_debug) cout <<"energy hole Error="<<error<<" " <<iter<<endl;
if (error<_tolerance) break;
if (max_error(_delta)<1.e-11) {error=0.0;cout<<"can not improve much\n";break;}
}

_Tp=nTp;

```

```
return fabs(init_error);
}

double QHDM::electron()
{
if (_dev->id()==3) return 0;
int n=_lattice_size;
_B[0]=_B[n-1]=0.0;
_A[0]=_A[n-1]=0.0;
_C[0]=_C[n-1]=1.0;
_delta[0]=_delta[n-1]=0.0;

int i;
vector<double> nn=_n;
vector<double>::const_iterator iter_max,iter_min;
double init_error;
for (int iter=0;iter<20;iter++)
{
    for(i=1;i<n-1;i++)
    {
        double Av= Dn(_Tn[i-1])*f (fir(i,_Tn))/_h[i];
        double Bv= Dn(_Tn[i+1])*f1(fir(i+1,_Tn))/_h[i+1];
        double Cv= -Dn(_Tn[i])*f(fir(i+1,_Tn))/_h[i+1]
                    -Dn(_Tn[i])*f1(fir(i,_Tn))/_h[i];
        _delta[i]=0.0;
        double hp=(_h[i]+_h[i+1])/2.0;
        _F[i]=Av*nn[i-1]+Bv*nn[i+1]+Cv*nn[i]+hp*F(nn[i],_p[i]);
        _F[i]=-_F[i];
        _A[i]=Av;
        _C[i]=Cv+hp*JFn(nn[i],_p[i]);
        _B[i]=Bv;
    }
    tridiag(_A,_C,_B,_F,_delta);
    for (i=1;i<n-1;i++) nn[i]+=_delta[i];
    iter_max=max_element(_F.begin(),_F.end());
    iter_min=min_element(_F.begin(),_F.end());
    double error=*iter_max-*iter_min;
    if (iter==0) init_error=error;
    if (_debug) cout <<"Error=<<error<<"    " <<iter<<endl;
    if (error<_tolerance) break;
}
_n=nn;
```

```
return fabs(init_error);
}

double QHDM::hole()
{

if (_dev->id()==3) return 0.0;
int n=_lattice_size;
_B[0]=_B[n-1]=0.0;
_A[0]=_A[n-1]=0.0;
_C[0]=_C[n-1]=1.0;
_delta[0]=_delta[n-1]=0.0;

int i;
vector<double> np=_p;
vector<double>::const_iterator iter_max,iter_min;
double init_error;
for (int iter=0;iter<20;iter++)
{
    for(i=1;i<n-1;i++)
    {
        double Av= Dp(_Tp[i-1])*f1(fir(i,_Tp))/_h[i];
        double Bv= Dp(_Tp[i+1])*f(fir(i+1,_Tp))/_h[i+1];
        double Cv= -Dp(_Tp[i])*f1(fir(i+1,_Tp))/_h[i+1]
                    -Dp(_Tp[i])*f(fir(i,_Tp))/_h[i];
        _delta[i]=0.0;
        double hp=(_h[i]+_h[i+1])/2.0;
        _F[i]=Av*np[i-1]+Bv*np[i+1]+Cv*np[i]+hp*_F[_n[i],np[i]];
        _F[i]=-_F[i];
        _A[i]=Av;
        _C[i]=Cv+hp*_JFp[_n[i],np[i]];
        _B[i]=Bv;
    }
    tridiag(_A,_C,_B,_F,_delta);
    for (i=1;i<n-1;i++) np[i]+=_delta[i];
    iter_max=max_element(_F.begin(),_F.end());
    iter_min=min_element(_F.begin(),_F.end());
    double error=*iter_max-*iter_min;
    if (iter==0) init_error=error;
    if (_debug) cout <<"Error="<
```

```
_p=np;
return fabs(init_error);
}

double QHDM::potential()
{
int i, n=_lattice_size;
vector<double> nphi=_phi;

_B[0]=_B[n-1]=1.0;
_A[0]=_A[n-1]=0.0;
_C[0]=_C[n-1]=0.0;
_delta[0]=_delta[n-1]=0.0;

vector<double>::const_iterator iter_max,iter_min;
double init_error;
for (int iter=0;iter<4000;iter++)
{
for(i=1;i<n-1;i++)
{
    double hp=(_h[i]+_h[i+1])/2.0;

    _F[i]=(nphi[i+1]-nphi[i])/_h[i+1]-(nphi[i]-nphi[i-1])/_h[i]-
        hp*(_n[i]*exp(nphi[i]-_phi[i])-_p[i]*exp(_phi[i]-nphi[i])-
        -N(_x[i]));
    _F[i]=-_F[i];
    _C[i]= 1.0/_h[i+1];
    _A[i]= 1.0/_h[i];
    _B[i]=-1.0/_h[i+1]-1.0/_h[i]-
        hp*(_n[i]*exp(nphi[i]-_phi[i])+_p[i]*exp(_phi[i]-nphi[i]));
}

tridag(_A,_B,_C,_F,_delta);
for (i=1;i<n-1;i++) nphi[i]+=_delta[i];
iter_max=max_element(_F.begin(),_F.end());
iter_min=min_element(_F.begin(),_F.end());
double error=*iter_max-*iter_min;
if (iter==0) init_error=error;
if (_debug) cout <<"Error="<
```

```
_phi[i]=nphi[i];
}
return fabs(init_error);
}

double QHDM::f1(double x)
{
return _Ber->B(x);
}

double QHDM::df1(double x)
{
return _Ber->dB(x);
}

double QHDM::f(double x)
{
return _Ber->B(-x);
}

double QHDM::df(double x)
{
return -_Ber->dB(-x);
}

void QHDM::Jn()
{
for (int i=1;i<_lattice_size-1;i++)
{
    _Jn[i]=(Dn(_Tn[i+1]) *
    _n[i+1]*f1(fir(i+1,_Tn))-_
    Dn(_Tn[i]) * _n[i]
    *f(fir(i+1,_Tn)))/_h[i+1];
//    _Jn[i]=(Dn(_Tn[i+1]) * _n[i+1]-Dn(_Tn[i]))/_h[i+1]-
}
_Jn[0]=_Jn[1];
_Jn[_lattice_size-1]=_Jn[_lattice_size-2];
}

void QHDM::velocity()
{
    Jn();
    for (int i=1;i<_lattice_size-1;i++)
```

```
{  
    const double gamma=5.0/3.0;  
    double c=sqrt(gamma * _Tn[i]*_T_cr/ (0.26*mass_of_electron))/_v_cr;  
    _vn[i]=-_Jn[i]/(_n[i]*c);  
}  
_vn[0]=_vn[1];  
_vn[_lattice_size-1]=_vn[_lattice_size-2];  
}  
  
double QHDM::F(double n, double p )  
{  
double nieo,nies,taun,taup;  
taun=TauN/_t_cr;  
taup=TauP/_t_cr;  
nieo=nie/_n_cr;  
nies=nieo*nieo;  
return (p*n-nies)/(taun*(p+nieo)+taup*(n+nieo));  
}  
  
double QHDM::JFn(double n, double p )  
{  
double nieo,nies,tmp1,taun,taup;  
nieo=nie/_n_cr;  
nies=nieo*nieo;  
taun=TauN/_t_cr;  
taup=TauP/_t_cr;  
tmp1=(taun*(p+nieo)+taup*(n+nieo));  
return -(n*p-nies)*taup/(tmp1*tmp1)+p/tmp1;  
}  
  
double QHDM::JFp(double n, double p )  
{  
double nieo,nies,tmp1,taun,taup;  
nieo=nie/_n_cr;  
nies=nieo*nieo;  
taun=TauN/_t_cr;  
taup=TauP/_t_cr;  
tmp1=(taun*(p+nieo)+taup*(n+nieo));  
return -(n*p-nies)*taun/(tmp1*tmp1)+n/tmp1;  
}  
  
double QHDM::prog1()  
{  
vector<double> error(4);  
vector<double>::const_iterator iter_max,iter_min;
```

```
double r_error;

for (int i=0;i<_prog1_iters;i++)
{
    error[1]=electron();
    error[2]=potential();
    error[3]=hole();
    error[0]=potential();
    iter_max=max_element(error.begin(),error.end());
    cout <<"Prog1 Error=<<*iter_max<<"    "<<i<<endl;
    if (*iter_max<_tolerance) break;
    if (i==0) r_error=*iter_max;
}

return r_error;
}

double QHDM::prog11()
{
vector<double> error(3);
vector<double>::const_iterator iter_max,iter_min;
double r_error;
for (int i=0;i<_max_iters;i++)
{
    error[1]=energy_electron();
    error[2]=potential();
    error[0]=electron();potential();
    iter_max=max_element(error.begin(),error.end());
    cout <<"Prog11 Error=<<*iter_max<<"    "<<i<<endl;
    for (int j=0;j<3;j++) cout<<error[j]<<endl;
    if (*iter_max<_tolerance) break;
    if (i==0) r_error=*iter_max;
}
return r_error;
}

double QHDM::prog12()
{
vector<double> error(3);
vector<double>::const_iterator iter_max,iter_min;
double r_error;

for (int i=0;i<_max_iters;i++)
{
    error[0]=hole();potential();
    error[1]=energy_hole();
```

```
error[2]=potential();
iter_max=max_element(error.begin(),error.end());
cout <<"Prog12 Error=<<*iter_max<<"    " <<i<<endl;
if (*iter_max<_tolerance) break;
if (i==0) r_error=*iter_max;
}
return r_error;
}

double QHDM::prog3()
{
vector<double> error(3);
vector<double>::const_iterator iter_max,iter_min;

for (int i=0;i<_max_iters;i++)
{
    error[0]=prog12();
    error[1]=prog11();
    error[2]=prog1();
    iter_max=max_element(error.begin(),error.end());
    cout <<"Prog3 Error=<<*iter_max<<"    " <<i<<endl;
    if (*iter_max<_tolerance) break;

}
return *iter_max;
}

void QHDM::prog2()
{
vector<double> error(5);
vector<double>::const_iterator iter_max,iter_min;

for (int i=0;i<_max_iters;i++)
{
    error[0]=electron();
    error[1]=energy_electron();  prog1();
    error[2]=hole();
    error[3]=energy_hole();
    error[4]=potential();
    iter_max=max_element(error.begin(),error.end());
    cout <<endl
        <<"*****" <<endl
    <<"Error summaries:"<<endl
    <<"Maximum Error=<<*iter_max<<" in iteration " <<i<<endl
    <<"Error in continuity equation:"<<endl
    <<"Electron   "<<error[0]<<endl
```

```
<<"Hole      "<<error[2]<<endl
<<"Error in equation for energy balance:"<<endl
<<"Electron   "<<error[1]<<endl
<<"Hole      "<<error[3]<<endl
<<"Error in Poisson equation: "<<error[4]<<endl
<<"*****"<<endl;
if (*iter_max<_tolerance) break;

}

}

#endif
```

A.4 device.h

```
#ifndef __DEVICE__H__
#define __DEVICE__H__

#include <worksheet.h>
#include <iostream.h>
#include <math.h>
#include <vector>
#include "physics.h"

class device
{
protected:
    double _L;
    int _id;
public:
    device();
    ~device();
    const double & length() {return _L;}
    const int & id() {return _id;}
    virtual double N(double x) {cout<<"Virtual device\n";return 0.0;}
};

device::device()
{
    _L=0;_id=0;
}

device::~device()
{
```

```
}
```

```
#endif
```

A.5 main.cc

```
#include <iostream.h>
#include "QHDM.h"
#include <worksheet.h>
#include "device.h"

int main(int argc,char **argv)
{
    cout <<"Modelling semiconductor devices with SCSIMU"<<endl
        <<"All rights reserved and no warranties of any kind."<<endl
        <<"Email to he@physics.usyd.edu.au for further details."<<endl;

    worksheet *s1;

    switch (argc)
    {
        case 2:
            if (strcmp(argv[1],"-h")==0)
            {
                cout <<"Usage: scsimu , scsimu worksheet,
                      scsimu worksheet section_name"<<endl;
                exit(1);
            }
            s1=new worksheet(argv[1]);
            break;
        case 3:
            s1=new worksheet(argv[2],argv[1]);
            break;
        default:
            s1=new worksheet();
            break;
    }

    string stage1_fn,stage2_fn,stage3_fn;

    QHDM sim(s1);
    s1->value("#stage1_output",stage1_fn);
    s1->value("#stage2_output",stage2_fn);
```

```
s1->value("#stage3_output",stage3_fn);

s1->save("qhdm.log");
sim.save(stage1_fn.c_str(),1,12);
sim.prog1();
sim.prog1();
sim.save(stage2_fn.c_str(),1,12);
sim.prog3();
sim.prog2();
sim.save(stage3_fn.c_str(),1,12);
}
```

A.6 maple.h

```
#ifndef __maple__DEVICE__
#define __maple__DEVICE__

#include "physics.h"
#include "device.h"

class maple: public device
{
public:
    maple();
    ~maple();

    double N(double );
};

maple::maple()
{
    _L=10.0;
    _id=3;
}

maple::~maple()
{
}

double maple::N(double x)
{
    return tanh(2.0*_L*(x/_L-0.5));
}

#endif
```

A.7 ballistic.h

```
#ifndef __ballistic__DEVICE__
#define __ballistic__DEVICE__

#include "physics.h"
#include "device.h"

class ballistic: public device
{
public:
    ballistic();
    ~ballistic();

    double N(double );
};

ballistic::ballistic()
{
    _L=0.6 micro meter;
    _id=2;
}

ballistic::~ballistic()
{
}

double ballistic::N(double x)
{
    double c1=0.1 micro meter;
    double c2=0.5 micro meter;
    if (x<c1) return 5.e17 per_cube_cent_meter;
    if (x<c2) return 2.e15 per_cube_cent_meter;
    return 5.e17 per_cube_cent_meter;
}

#endif
```

A.8 physics.h

```
/* prefix */
#ifndef __PHYSICS_H_
#define __PHYSICS_H_

// SI prefixes From "A Physicist's Desk Reference" by H. L. Anderson (2 ed).
```

```
#define exa      *1.e18
#define peta     *1.e15
#define tera     *1.e12
#define giga      *1.e9
#define mega     *1.e6
#define kilo      *1.e3
#define hecto    *1.e2
#define deka      *1.e1
#define deci      *1.e-1
#define centi    *1.e-2
#define cent     *1.e-2
#define milli    *1.e-3
#define micro    *1.e-6
#define nano     *1.e-9
#define pico     *1.e-12
#define femto    *1.e-15
#define atto     *1.e-18

#define meter     *1.0
#define second   *1.0
#define kilogram *1.0
#define joule    *1.0 //m^2 kg s^-2
#define Coulomb *1.0

#define per_second *1.0
#define per_meter  *1.0
#define per_kilogram *1.0

#define square_cent_meters *1.e-4;
#define cube_cent_meters  *1.e-6
#define per_cube_cent_meter *1.e6
#define per_V *1.0
#define per_second *1.0

/* physical constants all in SI units */
/*speed of light ms^-1 */
const double speed_of_light = 2.998e8 meter per_second;

/* charge of an electron C */
const double charge_of_electron = 1.60217733e-19 Coulomb;

/* mass of an electron kg */
const double mass_of_electron = 9.11e-31 kilogram;
```

```
/* Plank constant Js */
const double hbar = 1.05457266e-34 joule second;
const double PC   = 6.6260755e-34 joule second;

/* Boltzmann constant JK^-1 */
const double Kb  = 1.380568e-23 ;

/* permittivity of vacum Fm^-1 or JV^-2m^-1 */
const double epsilon0 = 8.854e-12;

// device related parameters

const double Eps1 = 11.7; // permittivity of semi-conductor material

const double nie = 1.4e10 per_cube_cent_meter;
//this value is from C. L. Gardner et. al Trans. Com. Desig. 8 501 (1989)

/* TauN, TauP seconds */
const double TauN= 1.7e-5 second;
const double TauP=3.95e-4 second;

/* Betan, Betap */
const double Betan=2.5;
const double Betap=2.5;

const double mun0=1400.0 square_cent_meters ; // per Velot per second
const double mup0=400.0 square_cent_meters;// per Velot per second

const double velocity_saturation=1.e7 cent_meter_per_second;
#endif
```

A.9 scsimu.h

```
#ifndef __SCSIMU_H_HAO__
#define __SCSIMU_H_HAO__

double Dn(double Tn, double mun);

double Dp(double Tp, double mup);

double dDn(double Tn, double mun);

double dDp(double Tp, double mup);

double Dn(double Tn);
```

```
double Dp(double Tp);

double dDn(double Tn);

double dDp(double Tp);

#endif
```

A.10 t.cc

```
#include <string>
#include <iostream>
#include <vector>
#include "linutil.h"
#include <worksheet.h>
#include "scsimu.h"
#include "device.h"
#include "PIN.h"

class t
{
private:
    double tt(double &);
    double (t::*p)(double &);
    vector<double> _a;
public:
    t();
    ~t();
    vector<double> & a() {return _a;}
    void dis(double &);
};

t::t()
{
    p=&tt;
}

t::~t()
{}

double t::tt(double & x)
{
    cout <<x<<endl;
}
```

```
void t::dis(double &x)
{
    (this->*p)(x);
}

int main(int argc,char **argv)
{
t ta;
double tmp=5.3;
ta.dis(tmp);

vector<double> a,b,c,r,x;
a=b=c=r=x=vector<double>(5);

a[0]=0;
a[1]=3;
a[2]=5;
a[3]=8;
a[4]=3.3;

b[0]=3;
b[1]=4;
b[2]=6;
b[3]=9;
b[4]=1;

c[0]=2;
c[1]=5;
c[2]=7;
c[3]=1;
c[4]=0;

r[0]=1;
r[1]=2;
r[2]=3;
r[3]=4;
r[4]=5;

tridag(a,b,c,r,x);

for (int i=0;i<5;i++) cout<<x[i]<<endl;
worksheet *s1;
```

```
switch (argc)
{
    case 2:
        if (strcmp(argv[1], "-h") == 0)
        {
            cout << "Usage: scsimu , scsimu worksheet,
                    bpm worksheet section_name" << endl;
            exit(1);
        }
        s1 = new worksheet(argv[1]);
        break;
    case 3:
        s1 = new worksheet(argv[2], argv[1]);
        break;
    default:
        s1 = new worksheet();
        break;
}

/*
double xb, xe;
int nx;
string file;
s1->value("#xb", xb);
s1->value("#xe", xe);
s1->value("#nx", nx);
s1->value("#file", file);

ofstream outf(file.c_str());
double dx=(xe-xb)/(nx-1.0);

outf.setf(ios::scientific);
outf.precision(20);

for (int i=0;i<nx;i++)
{
    double x=xb+dx*i;
    outf << x << "   << 1.0/(exp(x)-1.0)-exp(x)*x/
        ((exp(x)-1.0)*(exp(x)-1.0)) << " <<
        -0.5+x/6.0-x*x*x/180.0+x*x*x*x*x/5040.0 << endl;
}

cout << Dn(3.0) << endl;
*/
```

```
device *dp;
dp=new PIN();
cout <<dp->length()<<endl;
cout <<dp->id()<<endl;
}
```

A.11 worksheet.h

```
/*
This program was written by H. Hao (Copyright© 1998).
Please contact Dr. Hao He (he@physics.usyd.edu.au)
if you wish to use it.

*/
#ifndef _HAO_WORKSHEET_H_
#define _HAO_WORKSHEET_H_

#include <iostream.h>
#include <string>
#include <fstream.h>
#include <complex>
#include <stdlib.h>
#include <vector>
#include <assert.h>
#include <math.h>
#define _max_words 256

const string text("0123456789abcdefghijklmnopqrstuvwxyz"
                  "ABCDEFGHIJKLMNOPQRSTUVWXYZ#[]@$%^&*()--=;:<>,.?");

const string blank("\t");

class worksheet {
private:
    int _n_lines;
    vector < string > *_p;

public:
    worksheet();
    worksheet(const char *);
    worksheet(const char *,double);
    worksheet(const char *sec, const char * fn);
    ~worksheet();
}
```

```
    void list();
    void user_input(string key);
    int value(string , string &, int flag=1);
    int value(string , int &, int flag=1);
    int value(string , double &, int flag=1);
    int value(string , complex<double> &, int flag=1);
    void save(const char *);
    void getwords(string,string *,int & ns);
    void getwords(string tline, vector<string> & );
    void getcol(vector<double> &,int c);
    void getcol(double z,int c1, vector<double> &da, int c);
    void getcol(string &z, int c1, vector<double> &da, int c);
    void save(const char *, int c1, int n);
}

worksheet::worksheet()
{
_n_lines=0;
_p=new vector< string >;
}

worksheet::worksheet(const char * fn)
{
_n_lines=0;
_p=new vector< string >

string tline;
string::size_type pos=0;
string end("#end");
ifstream infile(fn,ios::in);
if (!infile) {cout<<fn<<" not found!"<<endl;exit(1);}
while (!infile.eof())
{
    getline(infile,tline,'\n');
    pos=0;
    if ((pos=tline.find(end))!=string::npos) break;
    _p->push_back(tline);
    _n_lines++;
    tline="";
}
cout <<_n_lines<<" lines of text have been read."<<endl;
}

worksheet::worksheet(const char * fn, double)
```

```
{  
_n_lines=0;  
_p=new vector< string >;  
  
string tline;  
string::size_type pos=0;  
ifstream infile(fn,ios::in);  
if (!infile) {cout<<fn<<" not found!"<<endl;exit(1);}  
while (!infile.eof())  
{  
    getline(infile,tline,'\n');  
    pos=0;  
    if (tline[0]!='#'){  
        _p->push_back(tline);  
        _n_lines++;}  
    tline="";  
}  
_n_lines--;cout <<_n_lines<<" lines of text have been read."<<endl;  
}  
  
worksheet::worksheet(const char *sec, const char * fn)  
{  
_n_lines=0;  
_p=new vector< string >;  
  
string tline, ss[_max_words];  
string begin_word="#begin",end_word="#end";  
int ns,flag=0;  
ifstream infile(fn,ios::in);  
if (!infile) {cout<<fn<<" not found!"<<endl;exit(1);}  
while (!infile.eof())  
{  
    getline(infile,tline,'\n');  
    getwords(tline,ss,ns);  
    if (ns>1) if ((ss[0]==begin_word)&&(ss[1]==sec)) flag=1;  
    if (flag==1) {  
        _p->push_back(tline);  
        _n_lines++;  
        tline="";}  
    if (ns>1) if ((ss[0]==end_word)&&(ss[1]==sec)) flag=0;  
}  
cout <<_n_lines<<" lines of text have been read."<<endl;  
if (!_n_lines) exit(1);  
if (flag) {cout <<"Warning:  
no "<<end_word<<" "<<sec<<endl<<" was found!"<<endl;}  
}
```

```
void worksheet::save(const char *fn)
{
ofstream outfile(fn);
for (int i=0;i<_n_lines;i++) outfile<< (*_p)[i]<<endl;
outfile<<"#end\n";
outfile.flush();
}

void worksheet::save(const char *fn,int c1, int n)
{
string ss[_max_words],tline; int ns;
ofstream outfile(fn);
int p1=0;
double previousvalue,oldvalue,newvalue;getwords((*_p)[0],ss,ns);
previousvalue=newvalue=atof(ss[c1].c_str());
for (int i=0;i<_n_lines;i++)
{

getwords((*_p)[i],ss,ns);
if (ns>c1+1)
{oldvalue=newvalue;newvalue=atof(ss[c1].c_str());}
if (oldvalue!=newvalue) p1++;
// cout <<oldvalue<<" " <<newvalue<<" " <<p1<<" " <<ns<<endl;
if (p1==n) {previousvalue=newvalue;p1=0;}
if ((newvalue==previousvalue)|| (ns<c1)) outfile<< (*_p)[i]<<endl;
}
outfile.flush();
}

worksheet::~worksheet()
{
```



```
void worksheet::user_input(string key)
{
string tline;

cout << key<<"=";
getline(cin,tline,'\'\n\'');
_p->push_back(key+' '+tline);
_n_lines++;

}
```

```
void worksheet::list()
{
for (int i=0;i<_n_lines;i++) cout<< (*_p)[i]<<endl;
}

int worksheet::value(string key,string &a, int flag=1)
{
string ss[_max_words],tline; int ns=0;
for(;;) {
    for (int i=0;i<_n_lines;i++)
    {
        getwords( (*_p)[i],ss,ns);
//    cout<<ss[0]<<" "<<key<<" "<<(ss[0]==key)<< " "<<ns<<endl;
        if ((ns>1)&&(ss[0]==key)) {for (int j=1;j<ns;j++)
if (j<ns-1) a+=ss[j]+'\'; else a+=ss[j];cout<<key<<"=<<a<<endl;return 1;}
    }
    if (flag==1) user_input(key); else return 0;
return 1;
}

int worksheet::value(string key,int &a, int flag=1)
{
string ss[_max_words],tline; int ns;
for(;;){for (int i=0;i<_n_lines;i++)
{
    getwords((*_p)[i],ss,ns);
//    cout<<ss[0]<<" "<<key<<" "<<(ss[0]==key)<< " "<<ns<<endl;
    if ((ns>1)&&(ss[0]==key)) {a=atoi(ss[1].c_str());
    cout<<key<<"=<<a<<endl;return 1;}
}
    if (flag) user_input(key); else return 0;
return 1;
}

int worksheet::value(string key,double &a, int flag=1)
{
string ss[_max_words],tline; int ns;
for(;;){for (int i=0;i<_n_lines;i++)
{
    getwords((*_p)[i],ss,ns);
//    cout<<ss[0]<<" "<<key<<" "<<(ss[0]==key)<< " "<<ns<<endl;
    if ((ns>1)&&(ss[0]==key)) {a=atof(ss[1].c_str());
    cout<<key<<"=<<a<<endl;return 1;}
}
    if (flag) user_input(key); else return 0;
return 1;
}
```

```
}

int worksheet::value(string key,complex<double> &a, int flag=1)
{
string ss[_max_words],tline; int ns;
for(;;){for (int i=0;i<_n_lines;i++)
{
    getwords((*_p)[i],ss,ns);
// cout<<ss[0]<<" "<<key<<" "<<(ss[0]==key)<< " "<<ns<<endl;
    if ((ns>2)&&(ss[0]==key)) {a=complex<double>
(atof(ss[1].c_str()),atof(ss[2].c_str()));cout<<key<<"="<<a<<endl;return 1;}
}
if (flag) user_input(key); else return 0;
return 1;
}

void worksheet::getcol(vector<double> &da, int c)
{
string ss[_max_words],tline; int ns;
for (int i=0;i<_n_lines;i++)
{
    getwords((*_p)[i],ss,ns);
    if (ns>c)
    {
        da.push_back(atof(ss[c].c_str()));
    }
}
}

void worksheet::getcol(string &z, int c1, vector<double> &da, int c)
{
string ss[_max_words],tline; int ns;
for (int i=0;i<_n_lines;i++)
{
    getwords((*_p)[i],ss,ns);
    if (ns>c) {
        if (ss[c1].find(z)!=string::npos)
        {
            da.push_back(atof(ss[c].c_str()));
        }
    }
}
}

void worksheet::getcol(double z, int c1, vector<double> &da, int c)
```

```
{  
string ss[_max_words],tline; int ns;  
for (int i=0;i<_n_lines;i++)  
{  
    getwords((*_p)[i],ss,ns);  
    if (ns>c) {  
        double tmp=atof(ss[c1].c_str());  
        if (fabs(tmp-z)<1.e-8)  
        {  
            da.push_back(atof(ss[c].c_str()));  
        }  
    }  
}  
  
void worksheet::getwords(string tline, string *ss, int &ns)  
{  
string::size_type pos=0,prev_pos=0;  
//cout<tline<<endl;  
ns=0;pos=tline.find_first_of(text);  
while ((pos=tline.find_first_of(blank,pos))!=string::npos)  
{  
if (prev_pos!=0) prev_pos++;  
if (pos>prev_pos) {ss[ns]=tline.substr(prev_pos,pos-prev_pos);ns++;}  
prev_pos=pos;pos++;  
}  
if (prev_pos!=0) prev_pos++;  
if (pos>prev_pos) {ss[ns]=tline.substr(prev_pos,pos-prev_pos);ns++;}  
}  
  
void worksheet::getwords(string tline, vector<string> &ss)  
{  
string::size_type pos=0,prev_pos=0;  
if (!ss.empty()) ss.erase(ss.begin(),ss.end());  
prev_pos=tline.find_first_of(text);  
for(;;){  
    pos=tline.find_first_of(blank,prev_pos);  
//    cout <<prev_pos<<" "<<pos<<endl;  
    if (pos>prev_pos) {ss.push_back(tline.substr(prev_pos,pos-prev_pos));}  
    if (pos==string::npos) break;  
    prev_pos=tline.find_first_of(text,pos);  
}  
}  
#undef _max_words  
#endif
```

A.12 Makefile.am

Note that using Makefile.am with automake you will automatically generate Makefile.in.

```
bin_PROGRAMS = scsimu hello
hello_SOURCES = t.cc device.h
EXTRA_DIST = *.tex *.ps *.eps example*.in comments.last scsimu

scsimu_SOURCES = main.cc scsimu.h QHDM.h worksheet.h hutil.h \
device.h linutil.h physics.h Bernoulli.h PIN.h ballistic.h maple.h
```

A.13 configure.in

Note that using this file with autoconf you will automatically produce configure script.

```
dnl Process this file with autoconf to produce a configure script.
AC_INIT(main.cc)
AM_INIT_AUTOMAKE(scsimu)
VERSION=0.89

dnl Checks for programs.

CXX=egc++
CXXFLAGS="-O "
CXX_OPTIMIZE_FLAGS=""
CXX_DEBUG_FLAGS="-g -DBZ_DEBUG"

AC_PROG_CXX
AC_ARG_PROGRAM
AC_PROG_INSTALL

AC_OUTPUT([Makefile])
```

A.14 example.in

```
#Your comments on this example should be sent to he@physics.usyd.edu.au
```

```
#This worksheet for the test device simulation with SCSIMU
```

```
#lattice_size 1000      number of points along the one dimensional device
#window_size 10         the value of normalised length
#temperature 300        I guess that you know this one
#device maple
#voltage 0.2           try to apply 0.2 V to the device and see what happens
```

```
#stage1_output maple0.0.2.dat
#stage2_output maple1.0.2.dat
#stage3_output maple2.0.2.dat
#end

#ok, let's give another try, this time with voltage=0

#begin maple_2
#lattice_size 1000      number of points along the one dimensional device
>window_size 10          the value of normalised length
#temperature 300         I guess that you know this one
#device maple
#voltage 0              try to apply 0 V to the device and see what happens
#stage1_output maple0.0.0.dat
#stage2_output maple1.0.0.dat
#stage3_output maple2.0.0.dat
#end maple_2

#begin PIN_1
#lattice_size 1000      number of points along the one dimensional device
>window_size 1            the value of normalised length
#max_iters 40
#temperature 300         I guess that you know this one
#device PIN
#voltage 0.1             try to apply 0 V to the device and see what happens
#stage1_output PIN0.p0.1.dat
#stage2_output PIN1.p0.1.dat
#stage3_output PIN2.p0.1.dat
#end PIN_1

#begin PIN_14
#lattice_size 1000      number of points along the one dimensional device
>window_size 1            the value of normalised length
#max_iters 40
#temperature 300         I guess that you know this one
#device PIN
#voltage 0.05            try to apply 0 V to the device and see what happens
#stage1_output PIN0.p0.05.dat
#stage2_output PIN1.p0.05.dat
#stage3_output PIN2.p0.05.dat
#end PIN_14

#begin PIN_11
#lattice_size 1000      number of points along the one dimensional device
>window_size 1            the value of normalised length
```

```
#max_iters 40
#temperature 300      I guess that you know this one
#device PIN
#voltage 0.5          try to apply 0 V to the device and see what happens
#stage1_output PIN0.p0.5.dat
#stage2_output PIN1.p0.5.dat
#stage3_output PIN2.p0.5.dat
#end PIN_11

#begin PIN_13
#lattice_size 1000    number of points along the one dimensional device
>window_size 1          the value of normalised length
#max_iters 40
#temperature 300      I guess that you know this one
#device PIN
#voltage 1.0          try to apply 0 V to the device and see what happens
#stage1_output PIN0.p1.0.dat
#stage2_output PIN1.p1.0.dat
#stage3_output PIN2.p1.0.dat
#end PIN_13

#begin PIN_2
#lattice_size 1000    number of points along the one dimensional device
>window_size 1          the value of normalised length
#temperature 300      I guess that you know this one
#max_iters 40
#device PIN
#voltage 0.            try to apply 0 V to the device and see what happens
#stage1_output PIN0.0.dat
#stage2_output PIN1.0.dat
#stage3_output PIN2.0.dat
#end PIN_2

#begin PIN_3
#lattice_size 1000    number of points along the one dimensional device
>window_size 1          the value of normalised length
#temperature 300      I guess that you know this one
#max_iters 40
#device PIN
#voltage -0.5         try to apply 0 V to the device and see what happens
#stage1_output PIN0.n0.5.dat
#stage2_output PIN1.n0.5.dat
#stage3_output PIN2.n0.5.dat
#end PIN_3

#begin PIN_4
```

```
#lattice_size 1000      number of points along the one dimensional device
>window_size 1           the value of normalised length
#temperature 300         I guess that you know this one
#max_iters 40
#device PIN
#voltage -1.5          try to apply 0 V to the device and see what happens
#stage1_output PIN0.n1.5.dat
#stage2_output PIN1.n1.5.dat
#stage3_output PIN2.n1.5.dat
#end PIN_4

#begin PIN_5
#lattice_size 1000      number of points along the one dimensional device
>window_size 1           the value of normalised length
#temperature 300         I guess that you know this one
#max_iters 80
#error_tolerance 1.e-5
#device PIN
#voltage -0.8          try to apply 0 V to the device and see what happens
#stage1_output PIN0.n0.8.dat
#stage2_output PIN1.n0.8.dat
#stage3_output PIN2.n0.8.dat
#end PIN_5

#begin ballistic_1
#lattice_size 1000      number of points along the one dimensional device
>window_size 1           the value of normalised length
#temperature 300         I guess that you know this one
#device ballistic
#voltage 0.1            try to apply 0 V to the device and see what happens
#stage1_output ballistic
#stage2_output ballistic1.0.1.dat
#stage3_output ballistic2.0.1.dat
#error_tolerance 1.e-5
#end ballistic_1

#begin ballistic_2
#lattice_size 1000      number of points along the one dimensional device
>window_size 1           the value of normalised length
#temperature 300 K       I guess that you know this one
#device ballistic
#voltage 1.0 V          try to apply 0 V to the device and see what happens
#stage1_output ballistic
#stage2_output test1.dat
#stage3_output test2.dat
#error_tolerance 1.e-5
```

```
#end ballistic_2

#begin ballistic_3
#lattice_size 1000      number of points along the one dimensional device
>window_size 1           the value of normalised length
#temperature 300 K      I guess that you know this one
#device ballistic
#voltage 0.5 V          try to apply 0 V to the device and see what happens
#stage1_output ballistic
#stage2_output ballistic1.0.5.dat
#stage3_output ballistic2.0.5.dat
#error_tolerance 1.e-5
#end ballistic_3

#begin ballistic_4
#lattice_size 1000      number of points along the one dimensional device
>window_size 1           the value of normalised length
#temperature 300 K      I guess that you know this one
#device ballistic
#voltage 2.0 V          try to apply 0 V to the device and see what happens
#stage1_output ballistic
#stage2_output ballistic1.2.0.dat
#stage3_output ballistic2.2.0.dat
#error_tolerance 1.e-5
#end ballistic_4

pg2dplot ballistic -yc 7 -d 'ballistic_profile.ps /
vps' -t '(b)' -x ' ' -y 'Log\|u(N) '
pg2dplot PIN0.0.dat -yc 7 -d 'PIN_profile.ps /
vps' -t '(a)' -x ' ' -y 'N '

set term postscript portrait
set output 'ballistic.n.ps'
plot 'ballistic2.1.0.dat', 'ballistic2.0.5.dat', 'ballistic2.2.0.dat'
set output 'ballistic.phi.ps'
plot 'ballistic2.1.0.dat' using 1:4, 'ballistic2.0.5.dat'
using 1:4, 'ballistic2.2.0.dat' using 1:4
set output 'ballistic.Tn.ps'
plot 'ballistic2.1.0.dat' using 1:5, 'ballistic2.0.5.dat'
using 1:5, 'ballistic2.2.0.dat' using 1:5

#begin PIN_6
#lattice_size 1000      number of points along the one dimensional device
>window_size 1           the value of normalised length
```

```
#max_iters 40
#temperature 300      I guess that you know this one
#device PIN
#voltage 0.3          try to apply 0 V to the device and see what happens
#stage1_output PIN0.p0.3.dat
#stage2_output PIN1.p0.3.dat
#stage3_output PIN2.p0.3.dat
#end PIN_6

#begin PIN_7
#lattice_size 1000    number of points along the one dimensional device
>window_size 1          the value of normalised length
#temperature 300      I guess that you know this one
#max_iters 40
#device PIN
#voltage -2.5         try to apply 0 V to the device and see what happens
#stage1_output PIN0.n2.5.dat
#stage2_output PIN1.n2.5.dat
#stage3_output PIN2.n2.5.dat
#end PIN_7

#begin PIN_8
#lattice_size 1000    number of points along the one dimensional device
>window_size 1          the value of normalised length
#temperature 300      I guess that you know this one
#max_iters 40
#device PIN
#voltage -1.0         try to apply 0 V to the device and see what happens
#stage1_output PIN0.n1.0.dat
#stage2_output PIN1.n1.0.dat
#stage3_output PIN2.n1.0.dat
#error_tolerance 1.e-6
#end PIN_8
```

A.15 hutil.h

```
#ifndef _HUTIL_H_
#define _HUTIL_H_

#include <complex>
#include <string>
#include <vector>
#include <math.h>

#define sp string("    ")
```

```
template <class T>
inline void coordinate(vector<T> &x, T xb, T xe, int n)
{
    T dx=(xe-xb)/(n-1);
    for (int i=0;i<n;i++) x[i]=dx*i;
}

template <class T>
inline int sign (T a)
{
    return (a>0) ? 1 : -1;
}

template <class T>
vector<T> conj( const vector<T> &B)
{
    unsigned int N = B.size();

    static vector<T> tmp(N);
    unsigned int i;

    for (i=0; i<N; i++)
        tmp[i] = conj( B[i]);

    return tmp;
}

template <class T>
vector<T> operator*(const vector<T> &A,
                     const vector<T> &B)
{
    unsigned int N = A.size();

    assert(N==B.size());

    static vector<T> tmp(N);
    unsigned int i;

    for (i=0; i<N; i++)
        tmp[i] = A[i] * B[i];

    return tmp;
}

template <class T>
```

```
vector<T> operator+(const vector<T> &A,
                      const vector<T> &B)
{
    unsigned int N = A.size();

    assert(N==B.size());

    static vector<T> tmp(N);
    unsigned int i;

    for (i=0; i<N; i++)
        tmp[i] = A[i] + B[i];

    return tmp;
}

template <class T>
vector<T> operator-(const vector<T> &A,
                      const vector<T> &B)
{
    unsigned int N = A.size();

    assert(N==B.size());

    static vector<T> tmp(N);
    unsigned int i;

    for (i=0; i<N; i++)
        tmp[i] = A[i] - B[i];

    return tmp;
}

template <class T>
vector<T> operator*(const vector<T> &A,
                      const double &B)
{
    unsigned int N = A.size();

    static vector<T> tmp(N);
    unsigned int i;

    for (i=0; i<N; i++)
        tmp[i] = A[i] * B;
```

```
    return tmp;
}

template <class T>
vector<T> operator*(const double &A,
                     const vector<T> &B)
{
    unsigned int N = B.size();

    static vector<T> tmp(N);
    unsigned int i;

    for (i=0; i<N; i++)
        tmp[i] = A * B[i];

    return tmp;
}

template <class T>
vector<T> operator*(const complex<double> &A,
                     const vector<T> &B)
{
    unsigned int N = B.size();

    static vector<T> tmp(N);
    unsigned int i;

    for (i=0; i<N; i++)
        tmp[i] = A * B[i];

    return tmp;
}

template <class T>
T max_error(const vector<T> &F)
{
    vector<T>::const_iterator iter_max,iter_min;
    iter_max=max_element(F.begin(),F.end());
    iter_min=min_element(F.begin(),F.end());
    double error=*iter_max-*iter_min; if (error==0) error=*iter_max;
    return error;
}

#endif
```

A.16 linutil.h

```
#define NRANSI
#ifndef __LINUTIL__H__
#define __LINUTIL__H__

#include <iostream>
#include <vector>

// solving a banded matrix equation
// a the lower band
// b the middle band
// c the upper band
// r right hand vector
// x the solution

template <class T>
void tridag(vector<T> &a, vector<T> & b, vector<T> & c,
             vector<T> & r, vector<T> & u)
{
    int n=a.size();
    long j;
    T bet;
    vector<T> gam(n);

    if (b[0] == 0.0) cout<<"Error 1 in tridag"<<endl;
    u[0]=r[0]/(bet=b[0]);
    for (j=1;j<n;j++) {
        gam[j]=c[j-1]/bet;
        bet=b[j]-a[j]*gam[j];
        if (bet == 0.0) cout<<"Error 2 in tridag"<<endl;
        u[j]=(r[j]-a[j]*u[j-1])/bet;
    }
    for (j=(n-1);j>=1;j--)
        u[j-1] -= gam[j]*u[j];
}

#endif
```

USQ



TOOWOOMBA

**MODELLING NONLOCAL PROCESSES
IN SEMICONDUCTOR DEVICES WITH
EXPONENTIAL DIFFERENCE
SCHEMES (Part 2)**

R V N Melnik

Mathematical Modelling & Numerical Analyses Group
Department of Mathematics & Computing, USQ

Hao He

THE UNIVERSITY OF
SOUTHERN
QUEENSLAND

Working Papers

FACULTY
OF
SCIENCES

**MODELLING NONLOCAL PROCESSES
IN SEMICONDUCTOR DEVICES WITH
EXPONENTIAL DIFFERENCE
SCHEMES (Part 2)**

R V N Melnik

Mathematical Modelling & Numerical Analyses Group
Department of Mathematics & Computing, USQ

Hao He

Department of Theoretical Physics
School of Physics, University of Sydney
Faculty of Sciences Working Paper Series
SC-MC-9831
30 November 1998

MODELLING NONLOCAL PROCESSES IN SEMICONDUCTOR DEVICES WITH EXPONENTIAL DIFFERENCE SCHEMES

Part 2: Numerical Methods and Computational Experiments

R. V. N. Melnik *

Mathematical Modelling & Numerical Analysis Group,
Department of Mathematics and Computing,
University of Southern Queensland, QLD 4350, Australia

Hao He

Department of Theoretical Physics,
School of Physics, University of Sydney, NSW 2006

Abstract

In a companion paper [11], based on the concept of relaxation time we considered the hierarchy of models for semiconductor devices. We focused at the quasi-hydrodynamic model as a reasonable compromise for the modelling of nonlocal and non-equilibrium processes in semiconductor plasma. The model was reduced to a form that is convenient for the numerical discretisation.

In this paper we construct efficient exponential difference schemes and apply them to modelling transport phenomena in semiconductors. Stability conditions, computational convergence and algorithmic realisations of the proposed schemes are discussed with numerical examples.

Key words: time relaxation, semiconductors, quasi-hydrodynamic models, exponential difference schemes.

*Corresponding author, E-mail: melnik@usq.edu.au

1 Introduction

In the space-time region $\bar{G} = \{(x, t) : 0 \leq x \leq 1, 0 \leq t \leq \bar{T}\}$ we consider the normalised form of the quasi-hydrodynamic model (see notation in Appendix and details in [11])

$$\left\{ \begin{array}{l} \partial_{xx}\varphi = n - p - N, \\ \partial_t n - \partial_x J_n = F, \\ 3/2\partial_t(nT_n) + \partial_x Q_n = -J_n\partial_x\varphi + P_n, \\ \partial_t p + \partial_x J_p = F, \\ 3/2\partial_t(pT_p) + \partial_x Q_p = -J_p\partial_x\varphi + P_p, \end{array} \right. \quad (1.1)$$

where

$$J_n = -n\mu_n\partial_x\varphi + \partial_x(T_n\mu_n n), \quad J_p = -p\mu_p\partial_x\varphi - \partial_x(T_p\mu_p p), \quad (1.2)$$

$$Q_n = \beta_n T_n n \mu_n \partial_x \varphi - \beta_n \partial_x [T_n D_n n], \quad Q_p = -\beta_p T_p p \mu_p \partial_x \varphi - \beta_p \partial_x [T_p D_p p]. \quad (1.3)$$

The system (1.1)–(1.3) is supplemented by the normalised initial

$$n(x, 0) = \bar{n}_0(x), \quad p(x, 0) = \bar{p}_0(x), \quad T_n(x, 0) = T_p(x, 0) = 1, \quad (1.4)$$

and boundary conditions

$$p - n + N = 0, \quad pn = n_{ie}^2, \quad T_n = T_p = 1, \quad x \in \partial\bar{G} = \{0, 1\}, \quad (1.5)$$

$$\varphi(0, t) = 0, \quad \varphi(1, t) = \tilde{U} + \tilde{\varphi}_{cont}. \quad (1.6)$$

It is also assumed that the condition $J_n(x, 0) = J_p(x, 0) = 0$ and the normalised conjugating conditions $\varphi(0, 0) = 0, \varphi(1, 0) = \tilde{U} + \tilde{\varphi}_{cont}$ are satisfied.

Model (1.1)–(1.6) allows us to adequately describe a number of non-stationary physical phenomena in semiconductor devices, including carrier heating and velocity overshoot. One of the main features of this model is accounting for a non-equilibrium and non-local character of electron-hole semiconductor plasma, a feature absent in the classical drift-diffusion model. Since technological advances lead to further reduction of device sizes, a higher density of configuration and power density of scattering, non-local and non-equilibrium phenomena are becoming increasingly important in device simulation.

In order to adequately describe these phenomena of semiconductor plasma it is often unnecessary to invoke the Boltzmann model, solution of which is known to be costly and “noisy” with a great deal of redundant information [1]. An important direction in engineering applications of semiconductor device theory is the analysis of “intermediate” (between the Boltzmann and drift-diffusion) models, such as (1.1)–(1.6). These models require efficient computational procedures, the development and justification of which is a challenging problem in applied mathematics (see [11] and references therein).

We organise this paper as follows.

- In Section 2 we discuss monotone exponential schemes constructed for the discretisation of continuity and energy balance equations in the quasi-hydrodynamic model. Stability issues for these schemes are also discussed in this section.
- In Section 3 we review known semi-implicit schemes that are used in semiconductor device modelling context and propose two algorithms for computational implementation of the schemes discussed in Section 2.
- In Section 4 we specify the choice of the initial approximation and stopping criteria used in our algorithms.
- In Section 5 we present results of computational experiments.
- Conclusions and future directions are discussed in Section 6.

2 Monotone Exponential Schemes for the Continuity and Energy Balance Equations

Major mathematical and computational challenges in the numerical solution of system (1.1)–(1.6) are connected with efficient approximation of the energy balance equation [1, 2, 3, 11, 23].

We start our analysis from the consideration of this equation in the stationary case. This case provides us with a clear picture of computational difficulties which can be “hidden” in the non-stationary case by an appropriate reduction of the time step.

We introduce a non-uniform grid in \tilde{G} [11]

$$\hat{\omega}_{hr} = \hat{\omega}_h \times \hat{\omega}_r, \quad (2.1)$$

where

$$\hat{\omega}_h = \{x_{i+1} = x_i + h_i, i = 0, \dots, N_0, x_0 = 0, x_{N_0+1} = 1, \sum_{i=0}^{N_0} h_i = 1\},$$

$$\hat{\omega}_r = \{t_j = t_{j-1} + \tau_j, j = 1, \dots, K-1, t_0 = 0, t_K = T_f, \sum_{j=1}^{K-1} \tau_j = T_f\},$$

and first we consider the difference scheme for the approximation of the stationary energy balance equation on the grid $\hat{\omega}_h = \bar{\omega}_h$ (i.e when $h_i \equiv h$) [23, 3]

$$\bar{A}_i(\mathcal{E}_n)_{i-1} - \bar{C}_i(\mathcal{E}_n)_i + \bar{B}_i(\mathcal{E}_n)_{i+1} = 0, \quad (2.2)$$

with $\mathcal{E}_n = nT_n$ and the coefficient of scheme (2.2) defined as

$$\bar{A}_i = \frac{\mu_n[(T_n)_{i-1}]}{h^2} f \left(\frac{\varphi_i - \varphi_{i-1}}{(T_n)_{i-1/2}} \right) [\beta_n(T_n)_{i-1} - hE_{i-1/2}/2], \quad (2.3)$$

$$\bar{B}_i = \frac{\mu_n[(T_n)_{i+1}]}{h^2} f_1 \left(\frac{\varphi_{i+1} - \varphi_i}{(T_n)_{i+1/2}} \right) [\beta_n(T_n)_{i+1} + hE_{i+1/2}/2], \quad (2.4)$$

$$\bar{C}_i = \bar{A}_{i+1} + \bar{B}_{i-1} + \bar{G}_i, \quad (2.5)$$

$$\begin{aligned} \bar{G}_i &= \frac{\mu_n[(T_n)_i]}{h^2} \left[f_1 \left(\frac{\varphi_i - \varphi_{i-1}}{(T_n)_{i-1/2}} \right) (\varphi_i - \varphi_{i-1}) - f \left(\frac{\varphi_{i+1} - \varphi_i}{(T_n)_{i+1/2}} \right) (\varphi_{i+1} - \varphi_i) \right] + \\ &\quad \frac{(T_n)_i - 1}{\tau_\omega^n[(T_n)_i](T_n)_i}. \end{aligned} \quad (2.6)$$

For scheme (2.2)–(2.6) the conditions of the Karetkina lemma [11] requiring that $\bar{A}_i > 0$, $\bar{B}_i > 0$, will be satisfied if [3]

$$\beta_n(T_n)_{i-1} - hE_{i-1/2}/2 > 0, \quad \beta_n(T_n)_{i+1} + hE_{i+1/2}/2 > 0. \quad (2.7)$$

Inequalities (2.7) lead to the restriction on the space discretisation step similar to (5.3) from [11],

$$h < 2\beta/E^*, \quad \text{where } E^* = \max_{i=1,\dots,N_0} |E_{i+1/2}|, \quad (2.8)$$

which may be burdensome for high electric fields. However, to overcome this condition is difficult. Indeed, if a splitting technique is used for the numerical solution of (1.1)–(1.6) and condition (2.8) is violated, then in the general case the positiveness of the solution cannot be guaranteed.

In order to relax the requirement (2.8), we apply exponential difference schemes [17]. In semiconductor device context such schemes were first used by Scharfetter and Gummel for the numerical solution of the drift-diffusion equations (see references, for example, in [10]).

2.1 Continuity equations

We recall the procedure for the construction of exponential difference schemes on the example of the 1D stationary continuity equation in the absence of recombination/generation/ionisation processes

$$\frac{\partial J_n}{\partial x} = 0. \quad (2.9)$$

The standard change of variables $n(x) \rightarrow n^{\text{new}}(x)$ in this case (see [17, 3, 10, 23] and references therein) is

$$n(x) = n^{\text{new}}(x) \exp \left[\int_{x_0}^x \left\{ \frac{1}{T_n} \frac{\partial \varphi(\xi)}{\partial \xi} - \frac{1}{D_n(T_n)} \frac{\partial D_n(\xi)}{\partial \xi} \right\} d\xi \right], \quad (2.10)$$

where x_0 is an arbitrary number such that $x_0 < x$. Substitution (2.10) into (1.2) leads to the following expression for the current density:

$$J_n(x) = D_n(T_n) \exp \left[\int_{x_0}^x \left\{ \frac{1}{T_n} \frac{\partial \varphi(\xi)}{\partial \xi} - \frac{\ln D_n(\xi)}{\partial \xi} \right\} d\xi \right] \frac{\partial n^{\text{new}}}{\partial x}. \quad (2.11)$$

If we now integrate this expression on the interval $[x_i, x_{i+1}]$ (assuming that the quantities $(J_n)_{i+1/2}$, $(D_n)_{i+1/2}$, $(\mu_n)_{i+1/2}$ are constants) and return to the old variable $n(x)$ we obtain

$$(J_n)_{i+1/2} = \frac{1}{\int_{x_i}^{x_{i+1}} J_n^*(x) dx} (D_n)_{i+1/2} [n_{i+1} J_n^*(x_{i+1}) - n_i], \quad (2.12)$$

where

$$\begin{aligned} J_n^*(x) &= \exp \left[- \int_{x_i}^x \left\{ \frac{1}{T_n(\xi)} \frac{\partial \varphi(\xi)}{\partial \xi} - \frac{\partial \ln D_n(\xi)}{\partial \xi} \right\} d\xi \right] = \\ &= \exp \left[- \frac{\varphi(x) - \varphi(x_i)}{T_n(x^*)} + \ln \frac{D_n(x)}{D_n(x_i)} \right]. \end{aligned} \quad (2.13)$$

Assuming, for example, that for $x^* \in [x_i, x_{i+1}]$

$$T_n(x^*) = \text{const} = \frac{1}{2} ((T_n)_i + (T_n)_{i+1}) = (T_n)_{i+1/2}, \quad (2.14)$$

expression (2.12) can be transformed to the following form:

$$\begin{aligned} (J_n)_{i+1/2} &= \frac{D_n((T_n)_i)}{h} \frac{\varphi_{i+1} - \varphi_i}{(T_n)_{i+1/2}} \left[\exp \frac{\varphi_{i+1} - \varphi_i}{(T_n)_{i+1/2}} - 1 \right]^{-1} \times \\ &\quad \left[n_{i+1} \frac{(D_n)_{i+1}}{(D_n)_i} - n_i \exp \left(\frac{\varphi_{i+1} - \varphi_i}{(T_n)_{i+1/2}} \right) \right], \end{aligned} \quad (2.15)$$

which coincides with the approximation (5.6) from [11]. Now, if we integrate the equation (2.9) on the interval $[x_{i-1/2}, x_{i+1/2}]$, we obtain that

$$[(J_n)_{i+1/2} - (J_n)_{i-1/2}]h = 0. \quad (2.16)$$

Substitution of the corresponding expressions for $(J_n)_{i\pm 1/2}$ (see (2.15)) into (2.16) leads to the following difference scheme:

$$[\Lambda_n(\varphi, T_n)n]_i = \frac{A_i}{h} n_{i-1} + \frac{B_i}{h} n_{i+1} - \frac{C_i}{h} n_i = 0, \quad (2.17)$$

where

$$A_i = \frac{(D_n)_{i-1}}{h} f \left(\frac{\varphi_i - \varphi_{i-1}}{(T_n)_{i-1/2}} \right), \quad B_i = \frac{(D_n)_{i+1}}{h} f_1 \left(\frac{\varphi_{i+1} - \varphi_i}{(T_n)_{i+1/2}} \right), \quad C_i = A_{i+1} + B_{i-1}. \quad (2.18)$$

Applying the above procedure to the non-stationary continuity equation on the non-uniform grid (2.1) we obtain

$$\frac{n_i^{l+1} - n_i^l}{\tau_{i+1}} = \frac{1}{h_i^*} [A_i^n n_{i-1}^{l+1} + B_i^n n_{i+1}^{l+1} - C_i^n n_i^{l+1}] + F_i, \quad (2.19)$$

where index l indicates the corresponding time-layer and the coefficients A_i^n , B_i^n and C_i^n are determined by the following formulae

$$A_i^n = \frac{D_n[(T_n)_{i-1}^{l+1}]}{h_i} f \left(\frac{\varphi_i^{l+1} - \varphi_{i-1}^{l+1}}{(T_n)_{i-1/2}^{l+1}} \right), \quad h_i = x_i - x_{i-1}, \quad h_i^* = \frac{h_i + h_{i+1}}{2}, \quad (2.20)$$

$$B_i^n = \frac{D_n[(T_n)_{i+1}^{l+1}]}{h_{i+1}} f_1 \left(\frac{\varphi_{i+1}^{l+1} - \varphi_i^{l+1}}{(T_n)_{i+1/2}^{l+1}} \right), \quad C_i^n = A_{i+1}^n + B_{i-1}^n. \quad (2.21)$$

Similarly, we derive the exponential difference scheme for the continuity equation for holes:

$$\frac{p_i^{l+1} - p_i^l}{\tau_{l+1}} = \frac{1}{h_i^*} [A_i^p p_{i-1}^{l+1} + B_i^p p_{i+1}^{l+1} - C_i^p p_i^{l+1}] + F_i, \quad (2.22)$$

where

$$A_i^p = \frac{D_p[(T_p)_{i-1}^{l+1}]}{h_i} f_1 \left(\frac{\varphi_i^{l+1} - \varphi_{i-1}^{l+1}}{(T_p)_{i-1/2}^{l+1}} \right), \quad (2.23)$$

$$B_i^p = \frac{D_p[(T_p)_{i+1}^{l+1}]}{h_{i+1}} f \left(\frac{\varphi_{i+1}^{l+1} - \varphi_i^{l+1}}{(T_p)_{i+1/2}^{l+1}} \right), \quad C_i^p = A_{i+1}^p + B_{i-1}^p \quad (2.24)$$

Remark 2.1 From the computational point of view, splitting algorithms are quite appealing in application to (1.1)–(1.6). However, in the application of such algorithms, a special care should be taken in approximating F in the RHS of (2.19) and (2.22). For example, for the approximation of F in the RHS of (2.19), the values of n^{l+1} can be found from the approximation of the Poisson equation. In this case the values of p have to be taken from the time layer l , which may significantly slow down the convergence. If convergence is satisfactory, then the computed value of n^{l+1} can be used for the approximation of F in the RHS of (2.22).

2.2 Energy balance equations

Now we are in the position to consider approximation procedures for the most difficult equations in system (1.1)–(1.6), for energy balance equations. Our approach is different from that proposed in [23]. We recall [11] that balance energy equations can be reduced to the forms amenable to computationally efficient schemes. For example, for the electron system we have

$$\begin{aligned} \frac{3}{2} \frac{\partial \mathcal{E}_n}{\partial t} &= \beta_n \frac{\partial^2 [D_n(T_n) \mathcal{E}_n]}{\partial x^2} - (1 + \beta_n) \frac{\partial}{\partial x} \left[\mu_n(T_n) \mathcal{E}_n \frac{\partial \varphi}{\partial x} \right] + \\ &\mu_n(T_n) \mathcal{E}_n \frac{\partial^2 \varphi}{\partial x^2} + \mathcal{E}_n \frac{\mu_n(T_n)}{T_n} \left(\frac{\partial \varphi}{\partial x} \right)^2 - \frac{\mathcal{E}_n (1 - 1/T_n)}{\tau_\omega^n(T_n)}, \end{aligned} \quad (2.25)$$

where $\mathcal{E}_n = n T_n$. As it was noted in [3] equation (2.25) can also be obtained from the third equation of system (1.1) by using the following identity

$$\frac{\partial}{\partial x} [\mu_n(T_n) n T_n] \frac{\partial \varphi}{\partial x} = \frac{\partial}{\partial x} \left[(\mu_n(T_n) n T_n) \frac{\partial \varphi}{\partial x} \right] - \mu_n(T_n) n T_n \frac{\partial^2 \varphi}{\partial x^2}. \quad (2.26)$$

In order to construct an exponential difference scheme for equation (2.25), we follow a procedure similar to that for the continuity equation. We introduce the change of variables $\mathcal{E}_n \rightarrow \mathcal{E}_n^{\text{new}}$ which is analogous to (2.10)

$$\mathcal{E}_n(x) = \mathcal{E}_n^{\text{new}}(x) \exp \left[\int_{x_0}^x \left\{ \frac{1 + \beta_n}{\beta_n} \frac{1}{T_n(\xi)} \frac{\partial \varphi(\xi)}{\partial \xi} - \frac{\partial \ln D_n(\xi)}{\partial \xi} \right\} d\xi \right], \quad (2.27)$$

where, as above, x_0 is an arbitrary number such that $x_0 < x$. As a result of transformations analogous to (2.11)–(2.16) we get the following difference scheme

$$\frac{3}{2} \frac{(\mathcal{E}_n)_i^{l+1} - (\mathcal{E}_n)_i^l}{\tau_{l+1}} = (\Lambda_{T_n}(\varphi^{l+1}, T_n^{l+1}) \mathcal{E}_n^{l+1})_i, \quad (2.28)$$

where

$$(\Lambda_{T_n}(\varphi, T_n) \mathcal{E}_n)_i = \frac{1}{h_i^*} [\tilde{A}_i^n (\mathcal{E}_n)_{i-1} + \tilde{B}_i^n (\mathcal{E}_n)_{i+1} - \tilde{C}_i^n (\mathcal{E}_n)_i] - \\ \left\{ -\mu_n [(T_n)_i] \varphi_{\bar{x}\bar{x},i} - \frac{\mu_n [(T_n)_i]}{(T_n)_i} (\varphi_{\bar{x},i})^2 + + \frac{1}{\tau_\omega^n [(T_n)_i]} - \frac{1}{\tau_\omega^n [(T_n)_i] (T_n)_i} \right\} (\mathcal{E}_n)_i, \quad (2.29)$$

and the coefficients of this difference scheme are defined as follows

$$(\tilde{A})_i^n = \frac{\beta_n D_n [(T_n)_{i-1}^{l+1}]}{h_i} f \left(\frac{1 + \beta_n}{\beta_n} \frac{\varphi_i^{l+1} - \varphi_{i-1}^{l+1}}{(T_n)_{i-1/2}^{l+1}} \right), \quad (2.30)$$

$$(\tilde{B})_i^n = \frac{\beta_n D_n [(T_n)_{i+1}^{l+1}]}{h_{i+1}} f_1 \left(\frac{1 + \beta_n}{\beta_n} \frac{\varphi_{i+1}^{l+1} - \varphi_i^{l+1}}{(T_n)_{i+1/2}^{l+1}} \right), \quad (2.31)$$

$$\tilde{C}_i^n = \tilde{A}_{i+1}^n + \tilde{B}_{i-1}^n. \quad (2.32)$$

We use standard difference scheme notation [17, 18] and denote the second and the first central difference derivatives on the non-uniform grid (2.1) by

$$\varphi_{\bar{x}\bar{x},i} = \frac{1}{h_i^*} \left[\frac{\varphi_{i+1} - \varphi_i}{h_{i+1}} - \frac{\varphi_i - \varphi_{i-1}}{h_i} \right] \quad \text{and} \quad \varphi_{\bar{x},i} = \frac{\varphi_{i+1} - \varphi_{i-1}}{2h_i^*}$$

respectively.

We construct the scheme analogous to (2.28)–(2.32) for the solution of the energy balance equation for the hole system:

$$\frac{3}{2} \frac{(\mathcal{E}_p)_i^{l+1} - (\mathcal{E}_p)_i^l}{\tau_{l+1}} = (\Lambda_{T_p}(\varphi^{l+1}, T_p^{l+1}) \mathcal{E}_p^{l+1})_i, \quad (2.33)$$

where

$$(\Lambda_{T_p}(\varphi, T_p) \mathcal{E}_p)_i = \frac{1}{h_i^*} [\tilde{A}_i^p (\mathcal{E}_p)_{i-1} + \tilde{B}_i^p (\mathcal{E}_p)_{i+1} - \tilde{C}_i^p (\mathcal{E}_p)_i] - \\ \left\{ \mu_p [(T_p)_i] \varphi_{\bar{x}\bar{x},i} - \frac{\mu_p [(T_p)_i]}{(T_p)_i} (\varphi_{\bar{x},i})^2 + + \frac{1}{\tau_\omega^p [(T_p)_i]} - \frac{1}{\tau_\omega^p [(T_p)_i] (T_p)_i} \right\} (\mathcal{E}_p)_i, \quad (2.34)$$

and the coefficients of this difference scheme are defined as follows

$$(\tilde{A})_i^p = \frac{\beta_p D_p[(T_p)_{i-1}^{l+1}]}{h_i} f_1 \left(\frac{1 + \beta_p}{\beta_p} \frac{\varphi_i^{l+1} - \varphi_{i-1}^{l+1}}{(T_p)_{i-1/2}^{l+1}} \right), \quad (2.35)$$

$$(\tilde{B})_i^p = \frac{\beta_p D_p[(T_p)_{i+1}^{l+1}]}{h_{i+1}} f \left(\frac{1 + \beta_p}{\beta_p} \frac{\varphi_{i+1}^{l+1} - \varphi_i^{l+1}}{(T_p)_{i+1/2}^{l+1}} \right), \quad (2.36)$$

$$\tilde{C}_i^p = \tilde{A}_i^p + \tilde{B}_i^p, \text{ with } f_1(x) = f(-x). \quad (2.37)$$

2.3 Monotonicity and stability

All coefficients of difference schemes (2.28)–(2.32), (2.33)–(2.37) preserve the positiveness property, namely it is easy to see that

$$\tilde{A}_i^n, \tilde{B}_i^n, \tilde{C}_i^n > 0, \quad \tilde{A}_i^p, \tilde{B}_i^p, \tilde{C}_i^p > 0. \quad (2.38)$$

Unfortunately, this fact cannot guarantee monotonicity of the constructed schemes. The sign of the functions near $(\mathcal{E}_n)_i$ and $(\mathcal{E}_p)_i$ in the expressions (2.29) and (2.34) cannot be defined *a priori*, and in the general case it may change. For example, we can not claim that on each time step and in each space point x_i the conditions $n_i \geq 0$, $T_n - 1 \geq 0$ (and respectively $p_i \geq 0$, $T_p - 1 \geq 0$) will always be satisfied. We only can claim that sufficient conditions for monotonicity given in the Karetkina lemma (see Theorem 5.1 in [11]) will be satisfied if in addition to (2.38) the conditions $(G_n)_i \geq 0$, $(G_p)_i \geq 0$ are also satisfied. In the context of the approximation of energy balance equations these conditions lead to the following inequalities (see also [3, 10])

$$(G_n)_i = -\mu_n[(T_n)_i]\varphi_{\bar{x}\bar{x},i} - \frac{\mu_n[(T_n)_i]}{(T_n)_i}(\varphi_{\bar{x},i})^2 + \frac{(T_n)_i - 1}{\tau_\omega^n[(T_n)_i](T_n)_i} + \frac{1.5}{\tau_{i+1}} \geq 0, \quad (2.39)$$

$$(G_p)_i = \mu_p[(T_p)_i]\varphi_{\bar{x}\bar{x},i} - \frac{\mu_p[(T_p)_i]}{(T_p)_i}(\varphi_{\bar{x},i})^2 + \frac{(T_p)_i - 1}{\tau_\omega^p[(T_p)_i](T_p)_i} + \frac{1.5}{\tau_{i+1}} \geq 0. \quad (2.40)$$

It is easy to verify [3] that inequalities (2.39) and (2.40) will be satisfied if

$$\tau < 1.5/(E^*)^2. \quad (2.41)$$

In the stationary case conditions (2.39) and (2.40) can be simplified to

$$(G_n)_i = -\mu_n[(T_n)_i]\varphi_{\bar{x}\bar{x},i} - \mu_n[(T_n)_i](\varphi_{\bar{x},i})^2/(T_n)_i + ((T_n)_i - 1)/(\tau_\omega^n[(T_n)_i](T_n)_i) \geq 0, \quad (2.42)$$

$$(G_p)_i = \mu_p[(T_p)_i]\varphi_{\bar{x}\bar{x},i} - \mu_p[(T_p)_i](\varphi_{\bar{x},i})^2/(T_p)_i + ((T_p)_i - 1)/(\tau_\omega^p[(T_p)_i](T_p)_i) \geq 0. \quad (2.43)$$

and will be satisfied when (2.8) holds.

It is clear that in the case of large gradients of the potential (i.e. in high electric fields) both conditions, (2.8) and (2.41) are very restrictive computationally. Hence, when modelling non-stationary processes in non-highly doped semiconductors, purely explicit schemes may become a competitive alternative to the proposed schemes due to their minimal computational cost per time-step. However, such explicit schemes typically require the time-step to be of order $1/\max_{0 \leq x \leq 1} N$ (i.e. $\tau = \mathcal{O}(1/\max_{0 \leq x \leq 1} N)$) (see, for example, [14]). This causes problems when in the RHS of the Poisson equation we have a large dopant concentration N [26]. The use of purely implicit schemes cannot resolve all difficulties; firstly, because in the general case such schemes cannot guarantee absolute stability of the numerical algorithm (subject to the approximation of F), and secondly, the computational cost for their numerical realization on each time-step integration substantially increases, especially for devices with two types of carriers. Therefore, one of the most promising directions in the development of efficient numerical schemes in semiconductor device theory lies with semi-implicit schemes.

3 Semi-Implicit Schemes and Their Algorithmic Realizations

Semi-implicit schemes have been extensively applied in modelling semiconductor devices with drift-diffusion types of models (see [13, 14, 3] and references therein). Basic ideas of their algorithmic implementation are typically connected either with Mock's scheme, Polsky-Rimshans' scheme, or a self-consistent scheme. We describe these ideas below.

1. In the Mock scheme [13] the potential is determined from the continuity equation for the *total current* rather than from the Poisson equation as in standard procedures. This scheme is not conservative and contains the disbalance term of the numerical nature of the order $\mathcal{O}(\tau)$ ($\tau = \max_{j=1,\dots,K-1} \tau_j$). The presence of this term deteriorates the scheme accuracy in practice when the time-step increases in spite of the absolute stability of this scheme for the linear case.
2. In the Polsky-Rimshans scheme [14] the potential is sought in two stages similar to prediction-correction procedures. First, from the continuity equation for the *total current*, we find a prediction $\varphi^{l+1/2}$. Then we correct it using the Poisson equation written for the time layer $l+1$. The Poisson equation on each step is solved with the accuracy $\mathcal{O}(\tau^3)$.
3. In the *self-consistent scheme* proposed in [3] we determine the potential on the $(l+1)$ -time layer through n^l and p^l . Then we compute n^{l+1} and p^{l+1} using φ^{l+1} . On the next step we determine the potential using the purely implicit scheme for the Poisson equation. In doing so, the values n^{l+1} and p^{l+1} (for example, $n^{l+1} = n^l + \tau n_t$) are to be found from the semi-implicit scheme for the continuity equation which in the

homogeneous case has the following form

$$(n^{l+1} - n^l)/\tau_{n+1} = [D_n((T_n)^l)n^l]_{\bar{x}\bar{x}} - (a^l \varphi_{\bar{x}}^{l+1})_{\bar{x}}, \quad (3.1)$$

where

$$a_i = (n_{i-1}\mu_n((T_n)_{i-1}) + n_i\mu_n((T_n)_i))/2. \quad (3.2)$$

From the computational point of view, the last scheme is very attractive if we use the central-difference approximation for the current density. Indeed, in this case we get a linear equation with respect to φ^{l+1} . Otherwise, if we use the exponential scheme, we have to solve a nonlinear equation in order to determine φ^{l+1} . Similar to Mock's scheme, in this self-consistent scheme we have a disbalance term $\mathcal{O}(\tau)$, which in the stationary-regime limit tends to zero.

3.1 Simplest semi-implicit schemes and deceleration of convergence

During recent years the interest to the application of semi-implicit schemes to non-local models has been increasing. One of the simplest algorithms of this type is what is known as the "relaxation-to-the-stationary-regime" method, which is widely used for the solution of drift-diffusion models (see references in [10]). In order to determine unknowns $(\varphi, n, T_n, p, T_p)$ of system (1.1)–(1.6) with this method, all equations are solved alternately, but for the solution of each equation an implicit scheme (with respect to the leading variable of that equation) is applied. Conceptually, this algorithm is a nonlinear Gummel type algorithm. Modelling devices with the DDM, this algorithm typically provides the user with good convergence when applied to devices with low and middle levels of doping. Unfortunately, for high-doping level devices convergence of this algorithm may seriously deteriorate. For example, modelling bipolar transistors with this method, we observe that for high forward voltages the concentration of majority carriers approaches the concentration of minority carriers in the vicinity of junctions. As a result we have a strong coupling between concentrations of both types of carriers through the potential function (partly induced by the requirement of the quasi-neutrality at the boundaries). This causes difficulties in numerical simulation of such devices one of which is slow convergence. It is possible that physical reasons for the deceleration of the convergence of simplest semi-implicit algorithms may be different from the described above. For example, it is well-known that the coupling between electrostatic potential and carrier concentrations in MOS-transistors that work in strong inversion regimes also increases. Other sources of coupling in modelling transient processes may be caused by the bias current. In all such situations we may expect a decrease in the rate of convergence of simplest semi-implicit algorithms.

Ultimately, the roots of the described computational difficulties lie with the quality of approximation of recombination/generation/ionisation terms. We recall that in the standard Gummel algorithm, all values of concentrations are taken from the previous time-layer, that may be unsatisfactory for many problems. However, from the computational viewpoint it is very attractive to apply a Gummel-type algorithm to the solution of the QHDM (1.1)–(1.6), where we have to solve a system of five, rather than three (as in the DDM), strongly

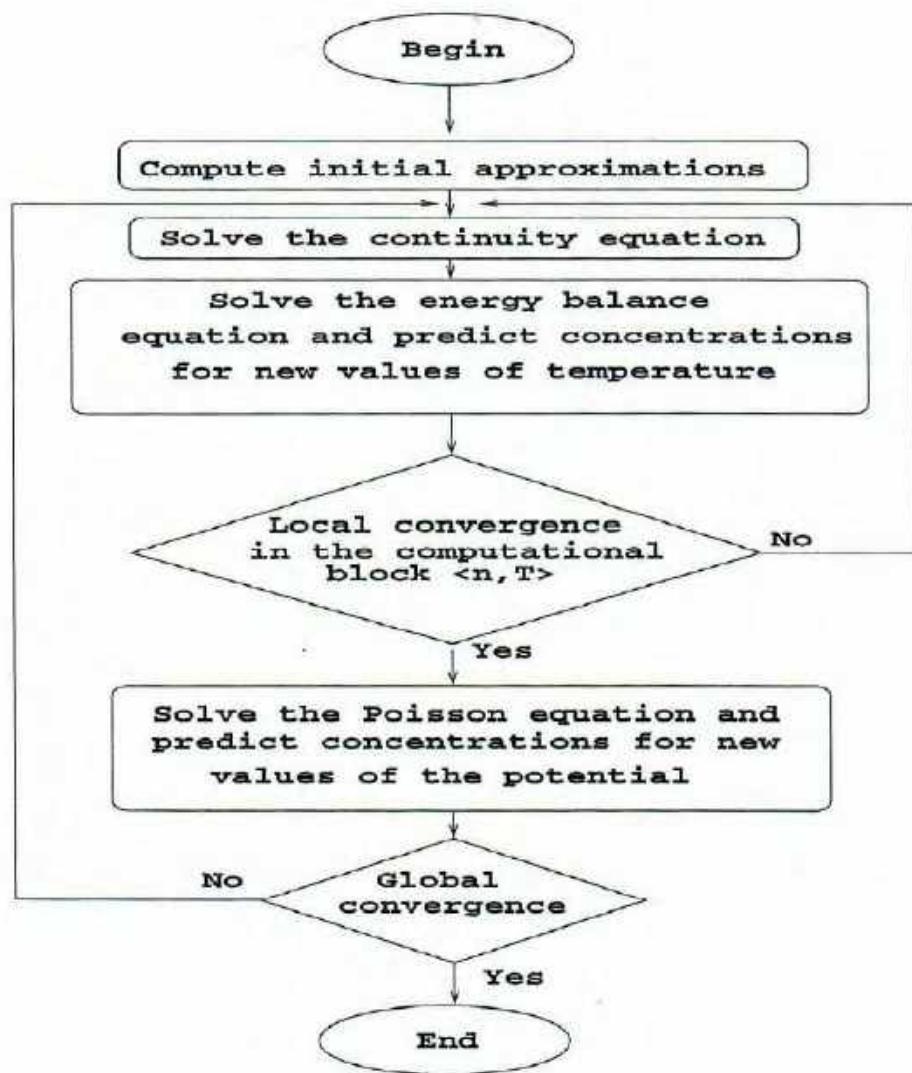


Figure 1: Conditionally coupled algorithm of the first order.

coupled nonlinear equations. Although for the last few decades attempts have been made to modify the Gummel algorithm in order to include a special treatment of the recombination/generation/ionisation terms, efficiency of the coupling of discretised equations in the Gummel-type algorithms critically depend on the type of modelling device and the strength of applied electric field. In this paper such a coupling is performed through the Boltzmann statistics and a Newton-type solver. In order to clarify the idea of such an algorithm we recall the connection between the *mixed basis* (n, p, φ), and the *hybrid basis*, (Φ_n, Φ_p, φ) in the case of the classical drift-diffusion model:

$$n = n_{\text{int}} \exp \left[\frac{\gamma \Delta E_G + \varphi - \varphi_n}{\varphi_T} \right], \quad p = n_{\text{int}} \exp \left[\frac{(1-\gamma) \Delta E_G + \varphi_p - \varphi}{\varphi_T} \right], \quad (3.3)$$

where ΔE_G is the effective bandgap narrowing [11], n_{int} is the intrinsic concentration, φ_T is the thermal potential, and γ is the experimentally measured parameter that takes into account the asymmetry factor. If we set $\gamma = 0.5$, then formulae (3.3) will simplify to (see formula (3.28) in [11])

$$n = n_{ie} \exp \left[\frac{\varphi - \varphi_n}{\varphi_T} \right], \quad p = n_{ie} \exp \left[\frac{\varphi_p - \varphi}{\varphi_T} \right] \quad (3.4)$$

or, finally, to

$$n = n_{ie} \exp \left(\frac{\varphi}{T} \right) \Phi_n, \quad p = n_{ie} \exp \left(-\frac{\varphi}{T} \right) \Phi_p, \quad (3.5)$$

where $\Phi_n = \exp(-\varphi_n/T)$ and $\Phi_p = \exp(\varphi_p/T)$ are the Fermi quasi-levels, and the temperature, T , is taken in energy units (multiplied by the factor k_b/q [11]).

3.2 Conditionally coupled algorithm of the first order

First, we explain the application of the Fermi-quasi-level representation (3.5) in the semi-implicit algorithm for the stationary drift-diffusion model:

Algorithm 3.0.

1. Given the initial approximation of φ, n, p we calculate the value of F of the recombination term using, for example, the Seidman-Choo scheme;
2. We sequentially solve the continuity equations with respect to n^{m+1} and p^{m+1} for computed values of φ^m and F^{m+1} (m is the index of "external" iterations with $m+1$ being current);
3. Using Newton's method we find φ^{m+1} in the following "internal" cycle:
 - (a) under the condition of steadiness of Fermi quasi-levels $(\Phi_n^{m+1}, \Phi_p^{m+1})$ we solve the discretized Poisson equation with respect to the correction $\delta\varphi_{k+1}^{m+1}$ (k is the index of internal iterations);
 - (b) we compute the value of the potential $\varphi_{k+1}^{m+1} = \varphi_k^{m+1} + \delta\varphi_{k+1}^{m+1}$ and (what is especially important under this approach!) we predict values of concentrations using (3.5) using the assumption of Fermi quasi-level steadiness:

$$n_{k+1}^{m+1} = n_k^{m+1} \exp \left(\frac{\delta\varphi_{k+1}^{m+1}}{T} \right), \quad p_{k+1}^{m+1} = p_k^{m+1} \exp \left(-\frac{\delta\varphi_{k+1}^{m+1}}{T} \right); \quad (3.6)$$

- (c) then we set

$$\varphi_k^{m+1} = \varphi_{k+1}^{m+1}, n_k^{m+1} = n_{k+1}^{m+1}, p_k^{m+1} = p_{k+1}^{m+1} \quad (3.7)$$

and go to the next *internal iteration* (i.e. set $k := k + 1$ and pass to the step (a)); internal iterations are conducted until the required accuracy is reached;

4. Steps 2 and 3 complete one external iteration; external iterations are conducted until the global convergence is reached.

This algorithm is conditionally coupled. Indeed, along with the calculation of the potential we calculate the prediction for n and p . In other words, quantities φ, n, p turn out to be coupled through the Boltzmann statistics (see (3.5)). We propose a generalization of Algorithm 3.0 to the quasi-hydrodynamic model and describe it on the example of the one-type carrier system. The flowchart of computations with this algorithm is shown on Fig. 1.

Algorithm 3.1.

1. We choose initial approximations for φ, n, T_n and calculate F ;
2. We sequentially solve the continuity and energy balance equations for computed values of φ^m and F^{m+1} (as before, m is the index of external iterations with $m + 1$ being current); for the solution of the energy balance equation we organise the following “internal” coupling procedure:

- (a) assuming the steadiness of Fermi quasi-levels we solve the discretized balance energy equation with respect to the corrections $\delta(T_n)_{k+1}^{m+1}$;
- (b) we compute the values of temperature on the *current internal iteration*, $(T_n)_{k+1}^{m+1} = (T_n)_k^{m+1} + \delta(T_n)_{k+1}^{m+1}$ and predict the values of concentration for just computed new values of temperature using the formula (assuming steadiness of Fermi quasi-levels):

$$n_{k+1}^{m+1} = n_{ie} \left(\frac{n_k^{m+1}}{n_{ie}} \right)^{\frac{(T_n)_k^{m+1}}{(T_n)_{k+1}^{m+1}}}; \quad (3.8)$$

- (c) we set $(T_n)_k^{m+1} = (T_n)_{k+1}^{m+1}, n_k^{m+1} = n_{k+1}^{m+1}$ and go to a new *internal iteration* by setting $k := k + 1$ and returning to (a); such internal iterations are performed until the given accuracy is achieved or the given number of times;
- 3. We perform a new “incomplete” external iteration for the computational block (n, T_n) (i.e. step 2); such “incomplete external” iterations are performed either up to the complete convergence or given number of times;
- 4. Then we solve the discretized Poisson equation

$$F_1(\varphi_{i-1}^{m+1}, \varphi_i^{m+1}, \varphi_{i+1}^{m+1}) = \frac{\varphi_{i+1}^{m+1} - \varphi_i^{m+1}}{h_{i+1}} - \frac{\varphi_i^{m+1} - \varphi_{i-1}^{m+1}}{h_i} - h_i^* \left[n_i^m \exp \left(\frac{\varphi_i^{m+1} - \varphi_i^m}{(T_n)_i^m} \right) - p_i^m \exp \left(\frac{-\varphi_i^{m+1} + \varphi_i^m}{(T_n)_i^m} \right) - N \right] = 0, \quad (3.9)$$

using the Newton method in a way similar to that described in step 3 of Algorithm 3.0 and assuming the relationships $n^{l+1} = n^l \exp((\varphi^{l+1} - \varphi^l)/(T_n)^l)$, and $p^{l+1} = p^l \exp((\varphi^l -$

$\varphi^{l+1})/(T_n)^l)$ (which are valid under constant temperatures); in other words, we organise a one more cycle of internal iterations:

$$\left(\frac{\partial F_1}{\partial \varphi_{i-1}} \right) \Big|_m \delta \varphi_{i-1}^{m+1} + \left(\frac{\partial F_1}{\partial \varphi_i} \right) \Big|_m \delta \varphi_i^{m+1} + \left(\frac{\partial F_1}{\partial \varphi_{i+1}} \right) \Big|_m \delta \varphi_{i+1}^{m+1} = -F_1 \Big|_m, \quad (3.10)$$

where $\delta \varphi_k^{m+1} = \varphi_k^{m+1} - \varphi_k^m$;

5. Steps 2–4 complete one external iteration; external iterations are performed until the convergence is reached.

Remark 3.1 “Incomplete” external iterations for the computational block (n, T_n) couples discretised versions of continuity and energy balance equations. The problem of the increase of the convergence rate for these iterations is addressed with the prognostic formula (3.8) (obtained under the steadiness of Fermi quasi-levels on the current iteration).

3.3 Coupling procedures using the Boltzmann statistics

“Incomplete” external iterations for the computational block (n, T_n) , employed in Algorithm 3.1, couple continuity and energy balance equations by the prognostic formula (3.8). This increases the rate of algorithm convergence. The prognostic formula (3.8) is formally obtainable from (3.5) under the assumption of steadiness of Fermi quasi-levels. This formula is applied only to the current iteration and is modified on the next iteration when new values of the temperature become available. In order to derive this formula we use the idea of (3.5) assuming that

$$n = n_{ie} \exp \left(\frac{\varphi - \varphi_n}{T_n} \right), \quad (3.11)$$

where T_n is the electron temperature in energy units. Hence, for the k^{th} iteration we have that

$$\ln \left(\frac{n_k}{n_{ie}} \right) = \frac{\varphi - \varphi_n}{(T_n)_k} \quad \text{or} \quad \varphi - \varphi_n = (T_n)_k \ln \left(\frac{n_k}{n_{ie}} \right). \quad (3.12)$$

Using (3.11) and (3.12) and assuming the constant value of the potential over two subsequent internal iterations, we obtain that

$$n_{k+1} = n_{ie} \exp \left[\frac{\varphi - \varphi_n}{(T_n)_{k+1}} \right] = n_{ie} \exp \left[\frac{(T_n)_k}{(T_n)_{k+1}} \ln \left(\frac{n_k}{n_{ie}} \right) \right] = n_{ie} \left(\frac{n_k}{n_{ie}} \right)^{\frac{(T_n)_k}{(T_n)_{k+1}}}. \quad (3.13)$$

Formula (3.13) allows us to couple the values of concentrations over two subsequent internal iterations through the values of temperature. Using (3.13) as a predictor, the discretized energy balance equation can be rewritten in the form

$$F_2 ((T_n)_{i-1}^{m+1}, (T_n)_i^{m+1}, (T_n)_{i+1}^{m+1}) = \left\{ \tilde{A}_i(n_{ie})_{i-1} \left(\frac{n_{i-1}^m}{(n_{ie})_{i-1}} \right)^{\frac{(T_n)_{i-1}^m}{(T_n)_{i-1}^{m+1}}} (T_n)_{i-1}^{m+1} + \right.$$

$$\tilde{B}_i(n_{ie})_{i+1} \left(\frac{n_{i+1}^m}{(n_{ie})_{i+1}} \right)^{\frac{(T_n)_{i+1}^m}{(T_n)_{i+1}^{m+1}}} (T_n)_{i+1}^{m+1} - \tilde{C}_i(n_{ie})_i \left(\frac{n_i^m}{(n_{ie})_i} \right)^{\frac{(T_n)_i^m}{(T_n)_i^{m+1}}} (T_n)_i^{m+1} \Big\} - \quad (3.14)$$

$$\left[-\mu \varphi_{zz} - \frac{\mu}{(T_n)_i^{m+1}} (\varphi_z)^2 + \frac{(T_n)_i^{m+1} - 1}{\tau_\omega^n (T_n)_i^{m+1}} \right] \times (n_{ie})_i \left(\frac{n_i^m}{(n_{ie})_i} \right)^{\frac{(T_n)_i^m}{(T_n)_i^{m+1}}} (T_n)_i^{m+1} = 0,$$

and the linearization procedure reduces this equation to the equation

$$\begin{aligned} & \left. \left(\frac{\partial F_2}{\partial (T_n)_{i-1}} \right) \right|_m \delta(T_n)_{i-1}^{m+1} + \left. \left(\frac{\partial F_2}{\partial (T_n)_i} \right) \right|_m \delta(T_n)_i^{m+1} + \\ & \left. \left(\frac{\partial F_2}{\partial (T_n)_{i+1}} \right) \right|_m \delta(T_n)_{i+1}^{m+1} = -F_2 \Big|_m, \end{aligned} \quad (3.15)$$

where $\delta(T_n)_j^{m+1} = (T_n)_j^{m+1} - (T_n)_j^m$, $j = i-1, i, i+1$.

3.4 Conditionally coupled algorithm of the second order

Algorithm 3.1 provides a computationally efficient tool for modelling a wide range of semiconductor devices. However, its main drawback lies with the assumption of constancy of the potential over two subsequent internal iterations. This assumption may not be fulfilled in the case when the coupling between the continuity and the energy balance equations is sufficiently strong. For example, difficulties may arise in the application of Algorithm 3.1 to the modelling of such devices as microwave PIN diodes that work in reverse-bias regimes [4, 8, 19, 12]. Clearly that in such cases Algorithm 3.1 has to be modified to account for an additional computational block for (p, T_p) . Moreover, if blocks (n, T_n) and (p, T_p) are to be treated sequentially, then a coupling between them has to be implemented (it can be done, for example, through the computational block solving the Poisson equation). When the strong coupling between continuity and energy balance equations is an intrinsic feature of the problem, we propose the *conditionally coupled algorithm of the second order*, which we refer to as Algorithm 3.2. Its most noticeable difference from Algorithm 3.1 is the absence of the iterative cycle inside of the block (n, T_n) (i.e. “incomplete” external iterations). The solution strategy with the second order conditionally coupled algorithm is presented on Fig. 2. A new QHDM computational block in Algorithm 3.2 contains the following steps:

1. Computation of values of carrier temperatures, T_n and T_p using exponential difference schemes (2.28)–(2.32) and (2.33)–(2.37) respectively;
2. Computation of concentrations, n and p , taken into account computed values of T_n and T_p using exponential difference schemes (2.19)–(2.21) and (2.22)–(2.24) respectively;
3. Recalculation of the potential φ taken into account the computed values of n , p , T_n , T_p .

The prediction stage for Algorithm 3.2 is realised by solving the classical drift-diffusion model.

Remark 3.2 Although the first order conditionally coupled algorithm described in Section 3.2 may meet serious computational challenges when applied to strongly coupled problems,

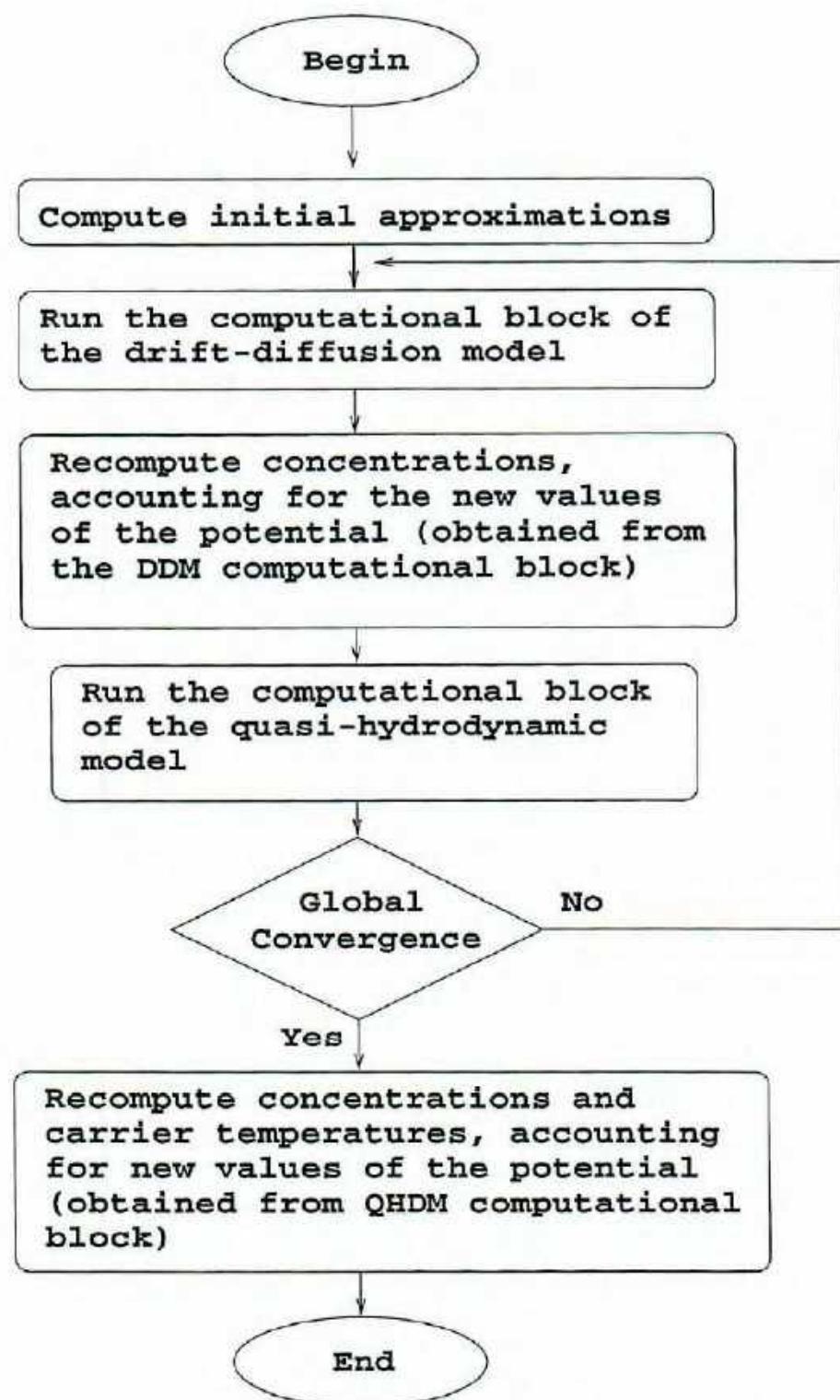


Figure 2: Conditionally coupled algorithm of the second order.

after sufficient number of iterations it will typically provides the user with a plausible qualitative picture of the main characteristics of devices. However, this simplified approach may not be adequate when the investigation is focused at the physical phenomena in semiconductor plasma, rather than at output characteristics of the device.

4 Initial Approximations, Stopping Criteria and the Solution of Linearized Problems

In order to guarantee the convergence of the algorithms described in Section 3, we have to take special efforts in constructing an appropriate initial approximation. As the initial approximation for the QHD computational block in Algorithm 3.2 we use the output from the solution of the DDM. Since the later model also requires an initial approximation, we use the assumption of quasi-neutrality

$$\rho = n - p - N = 0 \quad (4.1)$$

and thermal equilibrium

$$pn = n_{ie}^2 \quad (4.2)$$

in order to construct such an approximation. Using (4.1), (4.2) and assuming that $n = n_{ie} \exp((\tilde{U} - \varphi)/T_n)$, $p = n_{ie} \exp((\varphi - \tilde{U})/T_n)$ we determine the initial approximation for the potential as follows

$$\varphi = \tilde{U} + T_n \text{sign}(N) \ln \left[\frac{|N|}{2n_{ie}} + \sqrt{\left(\frac{|N|}{2n_{ie}} \right)^2 + 1} \right] \approx \tilde{U} + T_n \text{sign}(N) \ln \left(\frac{|N|}{n_{ie}} \right). \quad (4.3)$$

Then, the initial approximations for carrier concentrations can be found from the formulae

$$n = n_{ie} \exp \left(\frac{\varphi}{T_n} \right), \quad p = n_{ie} \exp \left(-\frac{\varphi}{T_p} \right), \quad (4.4)$$

that couples concentrations and the potential in the equilibrium case. As the initial approximations for carrier temperatures we assume their equality to the lattice temperature.

Remark 4.1 Strictly speaking the initial approximations for the carrier temperatures have to be computed, because the assumption of their equality to the lattice temperature may be dubious in simulation of some semiconductor devices such as reverse-bias PIN microwave diodes [8, 12]. However, in our numerical experiments we did not observe a deviation of the computed initial-temperature from the given equilibrium values for more than 8% (this was observed only in the neighbourhoods of p-n junctions).

Modelling semiconductor devices in high electric fields with the proposed schemes may lead to computational overflow due to the exponential character of these schemes. In order to avoid it, a special treatment of the Bernoulli functions $f(x)$, $f_1(x)$ and their derivatives

has to be implemented in the case when $x \rightarrow 0$. In our experiments we use the following formulae

$$f'(x) = \begin{cases} 1 - \frac{1}{2(1+x/2)^2}, & 0 \leq x \leq \tilde{\epsilon}, \\ \frac{1}{2(1-x/2)^2}, & -\tilde{\epsilon} < x < 0, \\ \frac{(\exp x + x \exp x)(\exp x - 1) - x(\exp x)^2}{(\exp x - 1)^2}, & \text{otherwise,} \end{cases} \quad (4.5)$$

$$f'_1(x) = \begin{cases} -\frac{1}{2(1+x/2)^2}, & 0 \leq x \leq \tilde{\epsilon}, \\ \frac{1}{2(1-x/2)^2} - 1, & -\tilde{\epsilon} < x < 0, \\ \frac{\exp x - 1 - x \exp x}{(\exp x - 1)^2}, & \text{otherwise,} \end{cases} \quad (4.6)$$

$$f(x) = \begin{cases} \frac{1}{1+|x|/2} + \frac{1}{2}(|x| + x), & |x| \leq \tilde{\epsilon}, \\ \frac{x \exp x}{\exp x - 1}, & \text{otherwise,} \end{cases} \quad (4.7)$$

$$f(x) = \begin{cases} \frac{1}{1+|x|/2} + \frac{1}{2}(|x| - x), & |x| \leq \tilde{\epsilon}, \\ \frac{x}{\exp x - 1}, & \text{otherwise} \end{cases} \quad (4.8)$$

with $\tilde{\epsilon}$ typically taken as 10^{-7} . Formulae (4.5)–(4.8) are easily obtainable from the expansion of exponential functions in power series ($\exp x \approx 1 + x + x^2/2$ and $\exp(-x) \approx 1 - x + x^2/2$).

The choice of stopping criteria for numerical algorithms is another important issue in modelling semiconductor devices. In our code we use the following criterion

$$\epsilon_{\varpi}^* = \begin{cases} \max_i |\varpi_i^{k+1} - \varpi_i^k|, & |\varpi_i^{k+1}| \leq 1, \\ \max_i \frac{|\varpi_i^{k+1} - \varpi_i^k|}{|\varpi_i^{k+1}|}, & |\varpi_i^{k+1}| > 1, \end{cases} \quad (4.9)$$

where ϖ is the corresponding function, for example, φ , n , T_n etc. Other criteria may also be chosen (see [3] and references therein). For example, in the one-dimensional case we may estimate the error of the conservative property of the total current that flows through the endpoints of the structure (i.e. endpoints of the interval $[0, 1]$). An inconvenience of this criterion becomes obvious for non-stationary problems where this quantity has to be checked at each moment of the transient process and the bias current has to be taken into account.

Finally, we consider technical issues of the implementation of computational blocks (n, T_n) and (p, T_p) connected with the solution of linearized systems of two coupled equations, continuity equation and the energy balance equation. For the sake of simplicity, we consider the

stationary electron system without the recombination term

$$\begin{cases} \Phi_{1i}(n_{i,i\pm 1}^{m+1}, (T_n)_{i,i\pm 1}^{m+1}) = 0, \\ \Phi_{2i}(n_{i,i\pm 1}^{m+1}, (T_n)_{i,i\pm 1}^{m+1}) = 0, \end{cases} \quad (4.10)$$

where

$$\Phi_{1i} = \frac{1}{h_i^*} (A_i n_{i-1}^{m+1} + B_i n_{i+1}^{m+1} - C_i n_i^{m+1}), \quad (4.11)$$

$$\begin{aligned} \Phi_{2i} = & \frac{1}{h_i^*} \left(\tilde{A}_i (\mathcal{E}_n)_{i-1}^{m+1} + \tilde{B}_i (\mathcal{E}_n)_{i+1}^{m+1} - \tilde{C}_i (\mathcal{E}_n)_i^{m+1} \right) - \\ & \left[-\mu_n \varphi_{\tilde{x}\tilde{x},i} - \frac{\mu_n}{(T_n)_i} (\varphi_{\tilde{x}})^2 + \frac{(T_n)_i^{m+1} - 1}{\tau_\omega^n (T_n)_i^{m+1}} \right] (\mathcal{E}_n)_i^{m+1}, \end{aligned} \quad (4.12)$$

$m+1$ denotes the current iteration of Newton's iterative process on which we determine corrections to the solution, $\delta\varpi_k^{m+1} = \varpi_i^{m+1} - \varpi_i^m$, and i is the index of the space grid point. Coefficients A_i, B_i, C_i and $\tilde{A}_i, \tilde{B}_i, \tilde{C}_i$ in (4.11)–(4.12) are determined by formulae (2.20)–(2.21) (or (2.23)–(2.24)) and (2.30)–(2.32) (or (2.35)–(2.37)) respectively.

The equations (4.10), linearised with respect to correction terms, have the following form:

$$\begin{cases} \sum_{k=i,i\pm 1} \left[\left(\frac{\partial \Phi_{1i}}{\partial n_k} \right)_m \delta n_k^{m+1} + \left(\frac{\partial \Phi_{1i}}{\partial (T_n)_k} \right)_m \delta (T_n)_k^{m+1} \right] = -\Phi_{1i}|_m, \\ \sum_{k=i,i\pm 1} \left[\left(\frac{\partial \Phi_{2i}}{\partial n_k} \right)_m \delta n_k^{m+1} + \left(\frac{\partial \Phi_{2i}}{\partial (T_n)_k} \right)_m \delta (T_n)_k^{m+1} \right] = -\Phi_{2i}|_m, \end{cases} \quad (4.13)$$

where derivatives in (4.13) are computed by the following formulae

$$\begin{aligned} \frac{\partial \Phi_{1i}}{\partial n_{i-1}} &= \frac{1}{h_i^*} A_i, \quad \frac{\partial \Phi_{1i}}{\partial n_{i+1}} = \frac{1}{h_i^*} B_i, \quad \frac{\partial \Phi_{1i}}{\partial n_i} = -\frac{1}{h_i^*} C_i, \quad \frac{\partial \Phi_{1i}}{\partial (T_n)_{i-1}} = \frac{1}{h_i^*} n_{i-1}^{m+1} (A_i)', \\ \frac{\partial \Phi_{1i}}{\partial (T_n)_{i+1}} &= \frac{1}{h_i^*} n_{i+1}^{m+1} (B_i)', \quad \frac{\partial \Phi_{1i}}{\partial (T_n)_i} = -\frac{1}{h_i^*} n_i^{m+1} (C_i)', \\ \frac{\partial \Phi_{2i}}{\partial n_{i-1}} &= \frac{1}{h_i^*} \tilde{A}_i (T_n)_{i-1}^{m+1}, \quad \frac{\partial \Phi_{2i}}{\partial n_{i+1}} = \frac{1}{h_i^*} \tilde{B}_i (T_n)_{i+1}^{m+1}, \\ \frac{\partial \Phi_{2i}}{\partial n_i} &= -\frac{1}{h_i^*} \tilde{C}_i (T_n)_i^{m+1} - \left[-\mu_n \varphi_{\tilde{x}\tilde{x},i}^{m+1} - \frac{\mu_n}{(T_n)_i^{m+1}} (\varphi_{\tilde{x}}^{m+1})^2 + \frac{(T_n)_i^{m+1} - 1}{\tau_\omega^n (T_n)_i^{m+1}} \right] (T_n)_i^{m+1}, \\ \frac{\partial \Phi_{2i}}{\partial (T_n)_{i-1}} &= \frac{1}{h_i^*} n_{i-1}^{m+1} [\tilde{A}_i + (T_n)_{i-1}^{m+1} (\tilde{A}_i)'], \quad \frac{\partial \Phi_{2i}}{\partial (T_n)_{i+1}} = \frac{1}{h_i^*} n_{i+1}^{m+1} [\tilde{B}_i + (T_n)_{i+1}^{m+1} (\tilde{B}_i)'], \\ \frac{\partial \Phi_{2i}}{\partial (T_n)_i} &= -\frac{1}{h_i^*} n_i^{m+1} [\tilde{C}_i + (T_n)_i^{m+1} (\tilde{C}_i)'] + \mu_n \varphi_{\tilde{x}\tilde{x},i}^{m+1} n_i^{m+1} - \frac{n_i^{m+1}}{\tau_\omega^n}. \end{aligned}$$

As a result, we have a large sparse system of linear equations with the matrix $2(N + 1) \times 2(N + 1)$ that has a block-tridiagonal structure. More precisely, it consists of 4 blocks each of which has $(N + 1) + 2N$ non-zero elements. Hence, in the most general case the total number of non-zero matrix entries cannot exceed $12N + 4$.

Such systems of linear equations may be effectively solved using direct methods that use the technology of sparse matrices (see [21] and references therein). In our experiments we used two packages for the solution of arising sparse systems. In the first package the data was packed into coupled lists. The program for the solution contains the algorithm for ordering and minimization of the number of non-zero elements, algorithms of symbolic and numerical factorization which are based on the representation of sparse matrices given by Singhal and Vlach (see references in [21]). The second package was based on the *sparse solver* presented in [25].

One of the main features of mathematical problems in semiconductor device theory is a large scattering of unknown quantities, the difficulty that has to be dealt with even for the dimensionalised systems of PDEs. Since classical iterative methods require at least estimates of spectrum boundaries for the guaranteed convergence, they may not be good candidates in the context of semiconductor device modelling. It is more appropriate to apply methods that do not require explicit knowledge of parameters that estimate the matrix spectrum. In this sense variational-type methods such as Kreig's method or methods based on biconjugate gradients seem to be very promising. However, when applying these methods the procedure for preconditioning requires special attention [6].

5 Numerical Experiments.

The constructed numerical schemes have been applied to modelling physical effects in electron-hole plasma of semiconductors.

As an example we present results on the modelling of a $n^+ - n - n^+$ ballistic diode. This device is often used to model the $n^+ - n - n^+$ channel in MEtal-Semiconductor Field-Effect Transistors (MESFET) and the modelling of this device is considered by a number of authors as a benchmark example [5, 2, 16]. The simulated diode is a unipolar device with $n^+ - n - n^+$ structure that has a central n region of length $0.4 \mu m$ bounded by two n^+ regions of length $0.1 \mu m$ each. The n^+ regions are doped at density $N = 5 \times 10^{17} \text{ cm}^{-3}$ while the n region is doped at $N = 2 \times 10^{15} \text{ cm}^{-3}$ (see Fig. 3). On Figure 4 we gives the electron concentration distribution calculated for the applied biases 0.1V , 0.5V and 1.0V . As one expects, for these applied voltages the concentration profile (see Figure 4) is similar to the shape of the doping distribution in the structure. When the bias increases we observe a drop in the concentration values in the right n^+ region. The electric potential as a function of the position at 0.1 , 0.5 and 1.0 V bias is given on Figure 5. As follow from (1.6) we applied the bias at the right contact while the left contact was grounded. This plot demonstrates electric field distribution over the semiconductor structure with the electron flow from left to right (note that this is opposite to the direction chosen in [16]). One can notice a slight drop in the electric field near the junction $n^+ - n$ (this drop leads to a slight "cooling" of electrons reported, for example, in [16] and assigned to a strong diffusion effect opposite to the carrier motion) and its maximum value near the junction $n - n^+$.

Figure 6 shows the electron temperature distribution (presented in the energy units for

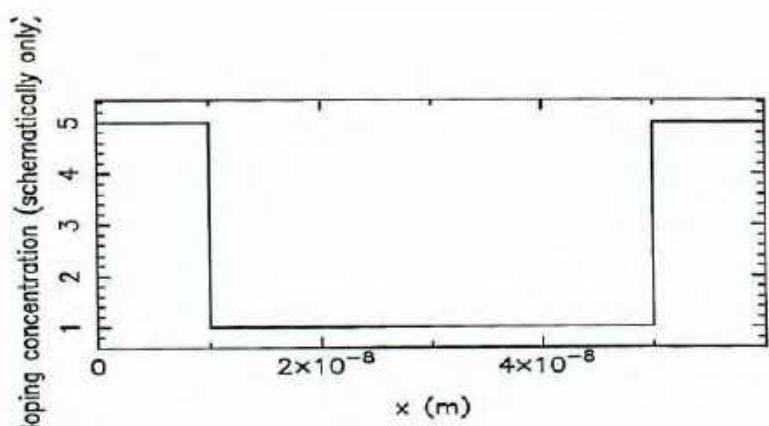


Figure 3: Dopant distribution in the ballistic diode.

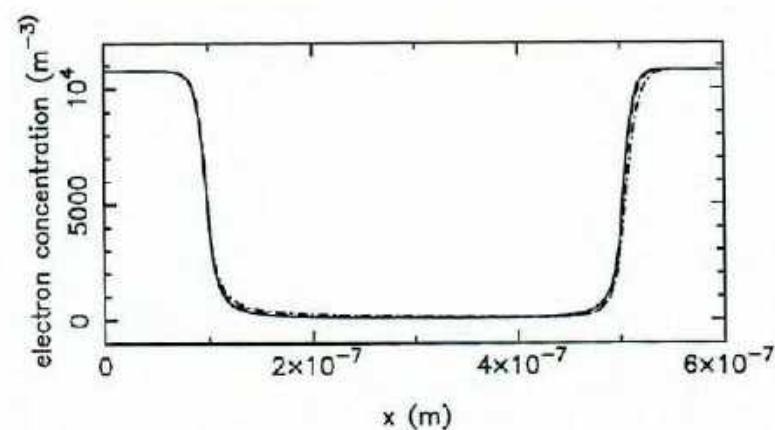


Figure 4: Concentration of electrons in normalised units.

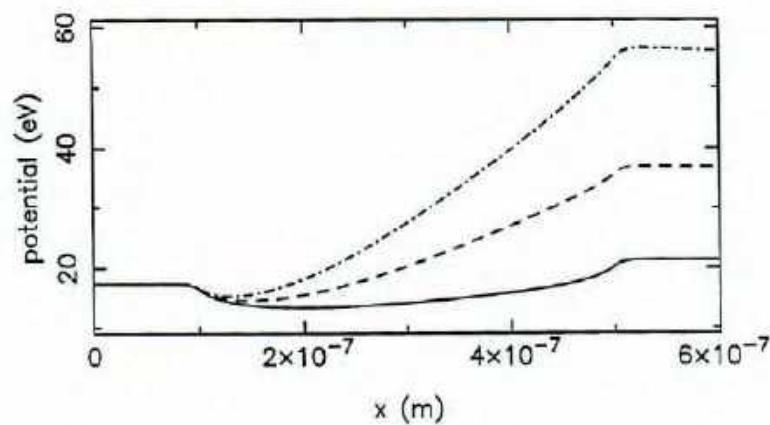


Figure 5: Electric potential.

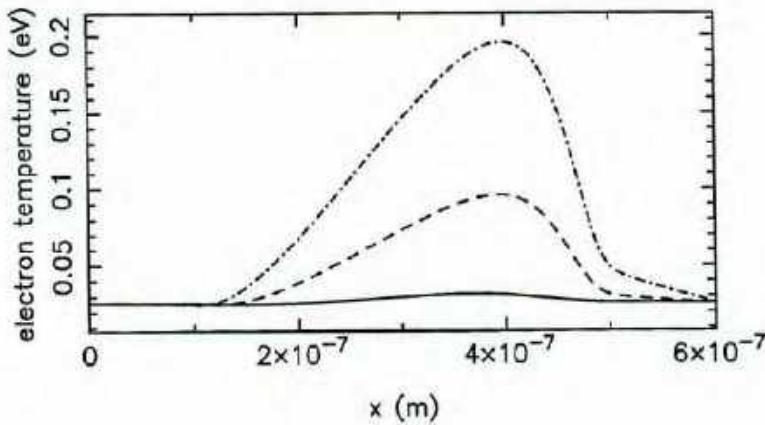


Figure 6: Electron temperature distribution.

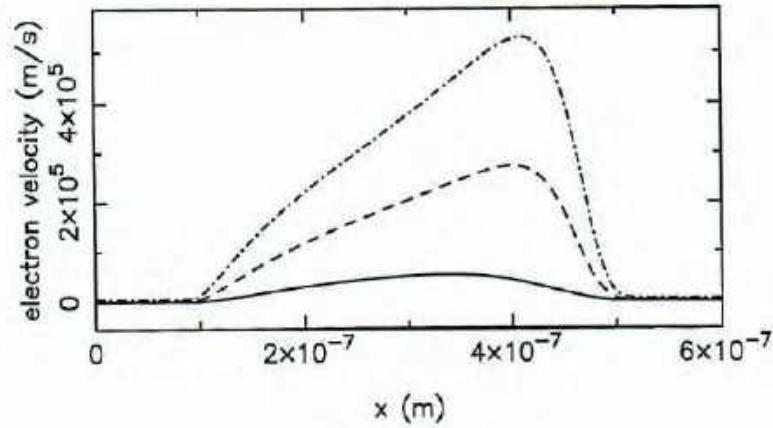


Figure 7: Electron velocity in the ballistic diode.

$T_l = 0.025\text{eV}$) calculated for the three applied biases. A shift of the temperature peak to the right as the applied bias increases is clearly visible on this plot. This is in agreement with computational results obtained by other authors [2, 16].

The velocity profile, computed according to formula (3.33) from [11], is presented on Figure 7. We note that such a quantitative velocity overshoot cannot be identified with the classical drift-diffusion model. Finally we display the ratio v_n/c where c is the sound speed computed by the formula

$$c = \sqrt{\gamma T_n/m_n} \quad \text{with} \quad \gamma = 5/3. \quad (5.1)$$

This ratio, presented on Figure 8, is sometimes referred to as the Mach number [5].

In [12] we reported some computational results on the simulation of physical processes in the PIN diodes used extensively for microwave control applications such as microwave switches and for electronically steered phased-array antennas [8]. Due to power-handling-capability requirements, the analysis of such devices has to include thermal effects. The results were obtained for a silicon $p^+ - i - n^+$ diode structure in the stationary case. Such diodes may be created by the molecular-beam epitaxy and are widely used as microwave switches. They may work in both direct and reverse biased regimes. Under forward-bias conditions, these devices exhibit a very low RF resistance, a higher conductivity and a larger

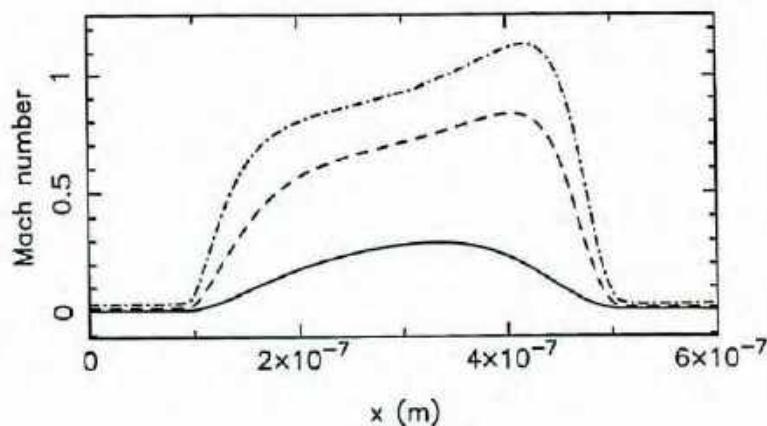


Figure 8: The Mach number in the ballistic diode.

breakdown compared to standard PN diodes; whereas under reverse-bias conditions they exhibit a very low constant capacitance. The latter case is also very interesting because the singular-perturbation-scaling technique can typically describe $p - n$ junctions under reverse biasing conditions only under small values of reverse biases. A singular perturbation analysis of reverse-biased semiconductor diodes for large applied biases is a difficult problem even in the case of the classical Van Roosbroek drift-diffusion model and is a topic of active research [4, 19]. These devices require further theoretical analysis and computational experiments using different models described in [11].

6 Conclusions and Future Directions.

In this paper we considered challenging problems of mathematical modelling in microelectronics. Based on the hierarchy of mathematical models for semiconductors we demonstrated that the quasi-hydrodynamic models provide a promising direction in the mathematical study of microstructures. These models belong to a wider class of nonlocal models which require the development of effective numerical procedures. For the investigation of non-equilibrium and non-local processes in semiconductors we proposed exponential monotone schemes and developed their algorithmic realisations. The results of theoretical analysis were demonstrated with computational experiments. We note that numerical schemes constructed in this work may be effectively applied to the investigation of EHP in the region of collector junction of bipolar transistors (BJT) as well as in the drain region of the Metal-Oxide Semiconductors (MOS). They can be used as a “building” block for modelling semiconductor superlattices and other layered structures in acousto- and opto-electronics. The technological progress in the design of optoelectronics devices, such as laser diodes (semiconductor lasers), LCD (Liquid Crystal Displays), light-emitting diodes (LEDs), and thin-film devices [9], requires further development of nonlocal mathematical models and efficient numerical methods for the investigation of physical processes in such devices.

One of the major challenges in the analysis of mathematical models arising in micro/optoelectronics consists of the investigation of a coupled system of equations with source

terms:

$$\begin{cases} \frac{\partial \mathbf{u}}{\partial t} + \frac{\partial \mathbf{F}_1(\mathbf{u}, \mathbf{v})}{\partial \mathbf{x}} \mathbf{v} = \mathbf{G}_1(\mathbf{u}, \mathbf{v}), \\ A \frac{\partial \mathbf{v}}{\partial \mathbf{x}} = \mathbf{G}_2(\mathbf{u}, \mathbf{v}), \end{cases} \quad (6.1)$$

where $\mathbf{u}(\mathbf{x}, t), \mathbf{v}(\mathbf{x}, t) \in \mathbb{R}^m$, A is a constant real matrix, $(\mathbf{x}, t) \in \mathbb{R}^n \times \mathbb{R}_+$, \mathbf{F}_1 is a given vector function and $\mathbf{G}_1, \mathbf{G}_2$ are source terms. We recall that the system (6.1) is a stiff system of PDEs if the time scales introduced by the source terms, \mathbf{G}_1 and \mathbf{G}_2 , are small compared to characteristic speeds and some appropriate length scale [24]. The mathematical analysis and the constructive solution of (6.1) in the class of piecewise-constant functions can be based on the approximation by Riemann problems which are simpler to solve than the standard Cauchy problem [15]. However, the solution of such a reduced problem may not exist. Alternatively, using the perturbation technique the system (6.1) can be reduced to a perturbed equation obtained by the substitution of \mathbf{v} , determined from the second equation of system (6.1), into the first equation. The perturbed equation is typically written with respect to a new (perturbed) variable \mathbf{u}_ϵ (see, for example, [15]). However, as a result of such a reduction, the definition of the parameter of perturbation, ϵ , in the reduced equation becomes coupled to the definition of the source terms and the natural space for perturbations becomes L_1 rather than L_2 . Immediate difficulties arising from this fact are that the flow map of the solution of the reduced equation might not be differentiable with respect to linear structure of L_1 and the contractivity of the flow for the perturbed equation with respect to L_1 -distance in the dimension higher than one cannot be guaranteed in general. These difficulties present a challenge for future work.

Acknowledgements.

Authors were supported by grant USQ-PTRP 17989 and by Australian Research Council Small Grant 17906. We thank Dr David Smith for his assistance at the final stage of preparation of this paper.

References

- [1] Aluru, N.R., Law, K.H., Pinsky, P.M. et al, Space-Time Galerkin/Least-Squares Finite Element Formulation for the Hydrodynamic Device Equations, *IEICE Trans. Electron.*, Vol. E77-C, No. 2, 227–235, 1994.
- [2] Apanovich, Y., Lyumkis, E., Polksky et al, Steady-State and Transient Analysis of Submicron Devices Using Energy Balance and Simplified Hydrodynamic Models, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 13, No. 6, 702–711, 1994.
- [3] Birukova, L.J. et al, Simulation algorithms for computing processes in electron plasma of submicron semiconductor devices, *Math. Modelling*, Vol. 1, No. 5, 11–22, 1989.
- [4] Brezzi, F., Capelo, A.C.S. and L. Gastaldi, A singular perturbation analysis of reverse-biased semiconductor diodes, *SIAM J. Math. Anal.*, Vol. 20, No. 2, 372–387, 1989.
- [5] C.L. Gardner, J.W. Jerome and D.J. Rose, Numerical Methods for the Hydrodynamic Device Model: Subsonic Flow, *IEEE Transactions on Computer-Aided Design*, Vol. 8, No. 5, 501–507, 1989.

- [6] Greenbaum, A., *Iterative Methods for Solving Linear Systems*, SIAM, 1997.
- [7] Jacobini, C. et al, A review of some charge transport properties of silicon, *Solid-State Electronics*, Vol. 20, No. 2, 77–89, 1977.
- [8] Kakati, D., Ramanan, C. and Ramamurthy, V., Numerical analysis of electrophysical characteristics of semiconductor devices accounting for the heat transfer, in *NEMACODE IV: Proceedings of the Fourth International Conference on the Numerical Analysis of Semiconductor Devices and Integrated Circuits: Trinity College, Dublin, Ireland*, Edited by J.J.H. Miller, Bool Press, 1985, 326–331.
- [9] Leigh, W.B. *Devices for Optoelectronics*, Marcel Dekker, 1996.
- [10] Lyumkis, E.D. et al Transient Semiconductor Device Simulation including energy balance equation, *COMPEL - The International Journal for Computation and Mathematics in Electrical and Electronic Engineering*, Vol. 11, No. 2, 311–325, 1992.
- [11] Melnik, R.V.N., He, H., Modelling nonlocal processes in semiconductor devices with exponential difference schemes. Part 1: Relaxation time approximations, *submitted, this journal*.
- [12] Melnik, R.V.N. and Melnik, K.N., Modelling of Nonlocal Physical Effects in Semiconductor Plasma Using Quasi-Hydrodynamic Models, *Computational Techniques and Applications: CTAC97*, Eds. J. Noye, M. Teubner, A. Gill, World Scientific, 1998, 441–448.
- [13] Mock, M.S., A time-dependent numerical model of the insulated-gate FET, *Solid-State Electronics*, Vol. 24, No. 10, 959–966, 1981.
- [14] Polsky, B.S. and Rimshans, J.S., Half-Implicit Difference Scheme for Numerical Simulation of Transient Processes in Semiconductor Devices, *Solid-State Electronics*, Vol. 29, No. 3, 321–328, 1986.
- [15] Raviart, P.-A. and Sainsaulieu, L., A nonconservative hyperbolic systems modelling spray dynamics. Part 1. Solution of the Riemann problem, *Mathematical Models and Methods in Applied Sciences*, Vol. 5, No. 3, 297–333, 1995.
- [16] Rudan, M., Odeh, F. and J. White, Numerical solution of the hydrodynamic model for a one-dimensional semiconductor device, *COMPEL - The International Journal for Computation and Mathematics in Electrical and Electronic Engineering*, Vol. 6, No. 3, 151–170, 1987.
- [17] A.A. Samarskii, *Theorie der Differenzenverfahren*, (Series: Mathematik und ihre Anwendungen in Physik und Technik), Aufl. Leipzig: Academische Verlagsgesellschaft Geest & Portig, 1984.
- [18] A.A. Samarskii and E.S. Nikolaev, *Numerical Methods for Grid Equations*, Basel, Boston: Birkhauser Verlag, 1989.
- [19] Schmeiser, C., On strongly reverse biased semiconductor diodes, *SIAM J. Appl. Math.*, Vol. 49, No. 6, 1734–1748, 1989.
- [20] Shur, M., *Physics of Semiconductor Devices*, Prentice Hall, 1990.
- [21] Singhal, K. and J. Vlach, *Computer Methods for Circuit Analysis and Design*, Van Nostrand Reinhold, New York, Toronto, 1994.
- [22] Sze, S. M., *Physics of Semiconductor Devices*, John Wiley & Sons, 1981.
- [23] Tang, T.-W., Extension of the Scharfetter-Gummel algorithm to the energy balance equation, *IEEE Transactions on Electronic Devices*, Vol. ED-31, No. 12, 1912–1914, 1984.
- [24] Tveito, A. and Winther, R., The solution of non-strictly hyperbolic conservation laws may be hard to compute, *SIAM J. Sci. Comput.*, Vol. 16, No. 2, 320–329, 1995.
- [25] Vetterling, W.T. et al, *Numerical Recipes Example Book (C)*, Cambridge University Press, 1994.
- [26] Widiger, D.J., Two-dimensional transient simulation of an idealized high electron mobility transistor, *IEEE Trans. Electron. Devices*, Vol. ED-32, No. 6, 1092–1103, 1985.

Appendix

The following notation for variables, constants and normalisation factors were used in this paper (the reader may consult [7, 22, 20, 11] for further details)

- n and p are concentrations of the majority (electrons) and minority (holes) respectively;
- φ and E are electrostatic potential and electric field strength respectively;
- T_n and T_p are carrier temperatures;
- F is the generation/recombination/ionisation term;
- P_n and P_p are rates of energy losses by scattering on the lattice for electrons and holes respectively;
- J_n and J_p are current densities;
- Q_n and Q_p are energy densities;
- D_n (D_p) and μ_n (μ_p) are diffusion and mobility coefficients;
- n_{ie} is the effective intrinsic concentration of carriers;
- U and φ_{cont} are applied voltage and the contact potential difference respectively;
- \bar{n}_0 and \bar{p}_0 are initial concentrations of carriers;
- f and f_1 are Bernoulli's functions;
- τ_ω^n and τ_ω^p are average energy relaxation times for electron and holes respectively;
- q is the electron charge ($q = |q|$ taken 1.6×10^{-19} Coulomb);
- ϵ is the relative dielectric permittivity of the semiconductor material (for Si it is 11.7 F/cm^{-1});
- ϵ_0 is the relative dielectric permittivity of vacuum (taken $8.85 \times 10^{-14} \text{ F/cm}^{-1}$);
- N is the doping density of a device (the summarised concentration of dopants);
- $\bar{\mathcal{E}}_n = 3nT_n/2$ and $\bar{\mathcal{E}}_p = 3pT_p/2$ are approximations of energy densities of carriers;
- τ_ω^n and τ_ω^p are characteristic times of energetic relaxation (taken $0.43 \times 10^{-12} \text{ s}$);
- T_l the lattice temperature (taken 300 K);
- β_n and β_p are Peltier coefficients (taken 2.5);
- v_s^n and v_s^p are saturation velocities of carriers;
- L is the length of the semiconductor structure;
- φ_{cont} is the contact potential;
- U is the applied voltage;
- $T_* = 0.0259$ is the normalisation factor for temperature;
- $\varphi_* = 0.0244$ is the normalisation factor for the potential;
- $\mu_* = 1$ and $D_* = T_*$ are normalisation factors for the mobility and diffusion coefficient respectively;
- $t_* = 5.0256 \times 10^{-5}$ is the time normalisation factor;
- $n_* = 1.2877 \times 10^{12}$ is the concentration normalisation factor;
- $J_* = 1.4349 \times 10^{-5}$ is the current density normalisation factor;
- $\alpha_* = 2.8571 \times 10^3$ and $c_* = 1.2 \times 10^{-19}$ are the normalisation factors for carrier ionisation and Auger recombination coefficients respectively;
- c_n and c_p are coefficients of the Auger recombination (taken 2.9×10^{-31} and 1.2×10^{-31} respectively);
- α_n and α_p are coefficients of collision ionization (taken 1×10^{-3} and 1×10^{-4} respectively);
- m_e is the electron mass;
- $m_n = 0.26m_e$ is the effective electron mass;
- $c_s = \sqrt{\gamma T_n/m_n}$ is the sound speed, where $\gamma = 5/3$ is the polytropic gas constant;
- k_b is the Boltzmann constant (taken 1.3×10^{-23});
- τ_n and τ_p are carriers life times (taken $1.7 \times 10^{-5} \text{ s}$ and $3.95 \times 10^{-4} \text{ s}$ respectively).

USQ



TOOWOOMBA

**APPROXIMATE MODELS OF DYNAMIC
THERMOVISCOELASTICITY
DESCRIBING SHAPE-MEMORY-ALLOY
PHASE TRANSITIONS**

R V N Melnik and A J Roberts
Department of Mathematics & Computing, USQ
SC-MC-9829
4 November 1998

THE UNIVERSITY OF
S O U T H E R N
Q U E E N S L A N D

Working Papers

FACULTY
OF
SCIENCES

**APPROXIMATE MODELS OF DYNAMIC
THERMOVISCOELASTICITY
DESCRIBING SHAPE-MEMORY-ALLOY
PHASE TRANSITIONS**

R V N Melnik and A J Roberts

Department of Mathematics & Computing, USQ

SC-MC-9829

4 November 1998

Approximate Models of Dynamic Thermoviscoelasticity Describing Shape-Memory-Alloy Phase Transitions

R.V.N. Melnik and A.J. Roberts

*Department of Mathematics and Computing, University of Southern Queensland, Toowoomba,
QLD 4305, Australia.*

1. Introduction

In this paper we consider two models for the approximate description of thermomechanical behaviour of viscoelastic materials. Accounting for thermal fields in such a description is important for all viscoelastic materials ranging from viscous fluids to elastic solids. The viscoelastic behaviour typically combines viscous and elastic properties and the relative proportion of this combination strongly depends on thermal characteristics of the material. Moreover, with changing thermal conditions, it is sometimes difficult to decide whether a particular material is a solid or a fluid. The key points in such decisions belong to the time of observation and to the choice of constitutive relations which couple stresses, deformation gradients, thermal fluxes and temperature.

Our analysis is based on the nonlocal theory of continuum mechanics which considers constitutive variables defined at a point as a function of their values over the whole spatial domain of interest rather than as a function at that point only [2]. This approach of rational mechanics allows us to derive a general model that is suitable for the description of thermomechanical behaviour of materials under a wide range of temperature and loading patterns. In our models we allow for the dependency of stresses not only on the deformation gradient and temperature but also on the rates of their changes. Such considerations put us closer to real situations where the time-dependent coupling between temperature and stresses have to include the velocity of the deformation gradient and the speed of thermal propagation. Another novelty of our paper is the accounting for finite speeds of thermal disturbances. We define the constitutive relationship for thermal fluxes using the Cattaneo-Vernotte equation which includes the classical Fourier law as a special limiting case (in the limit of zero relaxation time for heat fluxes). In particular this approach is critical in modelling short transient states in low temperature regimes.

During recent years a number of papers were devoted to the development of mathematical theory of thermomechanical phase transitions (see [19, 27, 12, 13, 1] and references therein). The majority of those papers dealt with important theoretical issues of models such as well-posedness and the global asymptotic behaviour of solutions. However, only a few papers have been devoted to the description of computational results using those models (see, for example, [20, 15] and references therein). Almost all developed models take into account neither the rate of thermal disturbances nor the relaxation time of thermal fluxes. However, the importance of these issues are well known in dynamic hyperbolic thermoelasticity where mathematical procedures and computational techniques have a longer history compared to that in thermoviscoelasticity [16, 23].

In dealing with the three main physical quantities of continuum mechanics (stresses, deformation gradients and displacements) it is important to take into account their coupling to the thermal field. This allows us to construct efficient mathematical models for the description of complicated phenomena, such as hysteresis, which are becoming increasingly important in a wide range of applications. In this paper we apply the de-

veloped models to the description of shape memory alloy effects in a large bar. It is well-known that for many types of shape memory materials the dependency of stresses on the deformation gradient upon loading and unloading is significantly different. Applying a large load at a low temperature, we may get a residual deformation gradient, which typically vanishes upon heating. The restoring of the original shape is referred to as the shape memory effect. This effect is discussed with two numerical examples.

The rest of the paper is organized as follows.

- Section 2 provides the reader with basic preliminaries and notation.
- The general formulation of the model is given in Section 3. In this section we specify the model for internal energy and derive restrictions on the model imposed by the second law of thermodynamics.
- In Section 4 we incorporate the Cattaneo-Vernotte equation for heat conduction into our model.
- Section 5 deals with the Landau-Devonshire model for the free energy function. The constitutive relation connecting stresses and the deformation gradient is also discussed in this section.
- In Section 6 we consider a one-dimensional model of thermoviscoelasticity and discuss the consequences of non-convexity of free energy function.
- Some numerical results are presented and discussed in Section 7.
- In Section 8 we use centre manifold theory to derive an approximate mathematical model for the description of thermomechanical behaviour of viscoelastic materials.

2. Preliminaries and Notation

Assume that an object of interest (a solid, fluid, gas or plasma) occupies the volume V in a fixed reference spatial configuration Ω at a certain time t_0 . This object in its spatio-temporal configuration will be referred to by the generic name "system". We aim to develop an efficient mathematical description of the dynamic thermomechanical behaviour of the system.

Let $\mathbf{x} = (x_1, x_2, x_3)$ be material (Lagrangian) coordinates of a material point of the system in the configuration Ω at time t_0 . Then the dynamics of the system is determined by the spatial displacements $\mathbf{u} = (u_1, u_2, u_3)$ of such material points as a function of the reference position, \mathbf{x} , and the time of interest, t . The partial derivative of displacement with respect to \mathbf{x} is identified with the symmetric strain tensor

$$\boldsymbol{\epsilon} = \text{sym} \left[\frac{\partial \mathbf{u}(\mathbf{x}, t)}{\partial \mathbf{x}} \right] \quad \text{or} \quad \epsilon_{ij}(\mathbf{x}, t) = \frac{1}{2} \left[\frac{\partial u_i(\mathbf{x}, t)}{\partial x_j} + \frac{\partial u_j(\mathbf{x}, t)}{\partial x_i} \right], \quad i, j = 1, 2, 3, \quad (2.1)$$

and the time derivative of the function \mathbf{u} is identified with the velocity of the system

$$\mathbf{v} = \frac{\partial \mathbf{u}}{\partial t} \quad \text{or} \quad v_i(\mathbf{x}, t) = \frac{\partial u_i(\mathbf{x}, t)}{\partial t}, \quad i = 1, 2, 3. \quad (2.2)$$

In (2.1) we require that $\det(I + \boldsymbol{\epsilon}) > 0$ which precludes a possibility of compression of the matter to zero and guarantees the local invertibility of $\mathbf{x} + \mathbf{u}(\mathbf{x}, t)$ [22]. Since the time derivatives are understood in the Lagrangian sense, \mathbf{x} is kept fixed in (2.2).

3. Nonlocal Models of Thermoviscoelasticity

The equation of motion requires information on forces acting per unit area of the matter and, hence, in a natural way, involves the concept of stresses. The stress is not a mere function of the deformation gradient, as it is often assumed. It also depends on temperature of the matter, its rate of change in time and the rate of change of deformation gradient ϵ . Let $\rho_0(\mathbf{x}, t_0) > 0$ be the density of the matter (the mass per unit volume) in the reference configuration Ω at time t_0 and $\rho(\mathbf{x}, t)$ be the density of the matter at time t where $t - t_0$ is sufficiently small. Then, in the Lagrangian system of coordinates (\mathbf{x}, t) , the equation for balance of mass is written in the form [22]

$$\rho(\mathbf{x}, t) \det(I + \epsilon(\mathbf{x}, t)) = \rho_0(\mathbf{x}, t_0). \quad (3.1)$$

The equation of motion has the following form

$$\rho \frac{\partial^2 \mathbf{u}}{\partial t^2} = \nabla_{\mathbf{x}} \cdot \mathbf{s} + \mathbf{F} \quad \text{with} \quad \mathbf{F} = \rho(\mathbf{f} + \hat{\mathbf{f}}) - \hat{\rho}\mathbf{v}, \quad (3.2)$$

where \mathbf{f} is a given body force per unit mass, $\hat{\rho}$ and $\hat{\mathbf{f}}$ are nonlocal mass and force residuals respectively, and \mathbf{s} is the stress tensor.

In Lagrangian coordinates the equation for energy balance has the form

$$\rho \frac{\partial}{\partial t} \left(e + \frac{\mathbf{v}^2}{2} \right) - \nabla_{\mathbf{x}} \cdot (\mathbf{s} \cdot \mathbf{v}) + \nabla \cdot \mathbf{q} = \rho \left(h + \hat{h} + \mathbf{f} \cdot \mathbf{v} - \frac{\hat{\rho}}{\rho} \left(e + \frac{\mathbf{v}^2}{2} \right) \right), \quad (3.3)$$

where e is the specific internal energy of the system, $\mathbf{v}^2 = \mathbf{v} \cdot \mathbf{v}$, h is the heat source density, \hat{h} is the nonlocal energy residual (see [2] for conditions on localised residuals) and \mathbf{q} is the heat flux. The scalar multiplication of (3.2) by \mathbf{v} gives

$$\rho \frac{\partial \mathbf{v}^2 / 2}{\partial t} - \mathbf{v} \cdot (\nabla \cdot \mathbf{s}) = (\mathbf{F}, \mathbf{v}) \equiv \rho(\mathbf{f} + \hat{\mathbf{f}}) \cdot \mathbf{v} - \hat{\rho}\mathbf{v}^2. \quad (3.4)$$

Taking into account normalisation, from (3.3) and (3.4) we get

$$\rho \frac{\partial e}{\partial t} - \mathbf{s}^T : (\nabla \mathbf{v}) + \nabla \cdot \mathbf{q} = g, \quad (3.5)$$

where $\mathbf{a}^T : \mathbf{b} = \sum_{i,j=1}^3 a_{ij} b_{ij}$ is the standard notation for the rank 2 tensors \mathbf{a} and \mathbf{b} and

$$g = \rho(h + \hat{h}) - \rho\hat{\mathbf{f}} \cdot \mathbf{v} - \hat{\rho} \left(e - \frac{\mathbf{v}^2}{2} \right). \quad (3.6)$$

The right-hand sides of equations (3.2) and (3.5) incorporate into the model nonlocal and dissipative effects of thermomechanical waves. As we shall see in the next section, under appropriate constitutive relations it is also possible to allow for a relaxation time for acceleration of the motion in response to applied gradients such as the deformation gradient and the temperature gradient.

We assume that there exists a one-to-one entropy function of the system state. We denote the density of such a function by η , and then the second law of thermodynamics is

$$\frac{\partial \eta}{\partial t} - \nabla \cdot \mathbf{r} \geq \xi + \hat{\xi} - \frac{\hat{\rho}}{\rho}, \quad (3.7)$$

where ξ is the entropy source density, r is the entropy flux density and $\hat{\xi}$ is the nonlocal entropy residual.

The system of equations (3.2), (3.5) combined with inequality (3.7) provides the general mathematical model for the description of thermomechanical behaviour of dynamic systems. The macroscopic modelling of such systems starts from the choice of constitutive relationships. We assume the existence of a functional Ψ invariant under a time shift and chose this functional in the form of the Helmholtz free energy

$$\Psi = e - \theta\eta, \quad (3.8)$$

where θ is the temperature of the system ($\theta > 0$, $\inf_{(x,t)} \theta = 0$). We also assume specific forms for the entropy flux and the entropy source density as

$$r = q/\theta, \quad \xi = h/\theta. \quad (3.9)$$

Using (3.8) in (3.5) and taking into account that

$$\nabla \cdot q = \theta \nabla \cdot (q/\theta) + (q \cdot \nabla \theta)/\theta, \quad (3.10)$$

from (3.7) and (3.9) we get the nonlocal formulation of the Clausius-Duhem inequality

$$-\frac{\hat{\rho}}{\rho} \left(\Psi - \frac{v^2}{2} \right) - \left(\frac{\partial \Psi}{\partial t} + \eta \frac{\partial \theta}{\partial t} \right) + s^T : \nabla v - \dot{f} \cdot v - \frac{q \cdot \nabla \theta}{\theta} - (\theta \hat{\xi} - \dot{h}) \geq 0. \quad (3.11)$$

The latter inequality together with requirements on localisation residuals (see [2] for details) impose restrictions on the choice of nonlocal residuals and the functions η , s and q . We assume that the entropy density is given in the form

$$\eta = -\frac{\partial \Psi}{\partial \theta}. \quad (3.12)$$

Finally, we have to specify the constitutive relationships that couple stresses, deformation gradients, temperature and heat fluxes

$$\Phi_1(s, \epsilon) = 0, \quad \Phi_2(q, \theta) = 0, \quad (3.13)$$

where it is implicitly assumed that these relations may involve spatial and temporal derivatives of the functions. In Section 4 and 5 we specify particular forms for Φ_1 and Φ_2 .

4. The Cattaneo-Vernotte Model for Heat Conduction

The choice of the function Φ_2 in (3.13) is made using the Cattaneo-Vernotte model

$$q + \tau_0 \frac{\partial q}{\partial t} = -k(\theta, \epsilon) \nabla \theta, \quad (4.1)$$

where τ_0 is the dimensionless thermal relaxation time and $k(\theta, \epsilon)$ is the thermal conductivity of the material (typically $k = 1 + \bar{\beta}\theta$ with the given dimensionless coefficient $\bar{\beta}$). Such a choice is made in order to account for the finite speeds of thermal wave propagation and thermally induced stress wave propagation coupled to the deformation gradient [11, 18].

In order to incorporate equation (4.1) into the general model of thermoviscoelasticity we use a consequence of (3.5)

$$\rho\tau_0 \frac{\partial^2 e}{\partial t^2} - \tau_0 \frac{\partial}{\partial t} [\mathbf{s}^T : (\nabla \mathbf{v})] + \tau_0 \nabla \cdot \left(\frac{\partial \mathbf{q}}{\partial t} \right) = \tau_0 \frac{\partial g}{\partial t}. \quad (4.2)$$

On the other hand, from (4.1) we get

$$\nabla \cdot \mathbf{q} + \tau_0 \nabla \cdot \left(\frac{\partial \mathbf{q}}{\partial t} \right) = -\nabla \cdot (k \nabla \theta). \quad (4.3)$$

Then from (3.5), (4.2), (4.3) we obtain the energy balance equation in the form

$$\rho \frac{\partial e}{\partial t} + \rho\tau_0 \frac{\partial^2 e}{\partial t^2} - \mathbf{s}^T : (\nabla \mathbf{v}) - \tau_0 \frac{\partial}{\partial t} [\mathbf{s}^T : (\nabla \mathbf{v})] - \nabla \cdot (k \nabla \theta) = G, \quad (4.4)$$

where

$$G = g + \tau_0 \frac{\partial g}{\partial t}. \quad (4.5)$$

During recent years, the interest in such a hyperbolic approach in the analysis of materials with memory has increased [6].

5. The Landau-Devonshire Model for the Helmholtz Free Energy and the Stress-Strain Relation

We start from the consideration of the one-dimensional case assuming the following approximation for the free energy of the system

$$\Psi(\theta, \epsilon) = \psi_0(\theta) + \psi_1(\theta)\psi_2(\epsilon) + \psi_3(\epsilon) \quad (5.1)$$

where $\psi_0(\theta)$ models thermal field contributions, $\psi_1(\theta)\psi_2(\epsilon)$ models shape-memory contributions and $\psi_3(\epsilon)$ models mechanical field contributions. These models are chosen in the following forms

$$\begin{cases} \psi_0(\theta) = \alpha_0 - \alpha_1 \theta \ln \theta, & \psi_1(\theta) = \frac{1}{2} \alpha_2 \theta, & \psi_2(\epsilon) = \epsilon^2, \\ \psi_3(\epsilon) = -\frac{1}{2} \alpha_2 \theta_1 \epsilon^2 - \frac{1}{4} \alpha_4 \epsilon^4 + \frac{1}{6} \alpha_6 \epsilon^6, \end{cases} \quad (5.2)$$

where all α_i and θ_1 are positive constants. The model (5.1)–(5.2), known as the Landau-Devonshire model for the Helmholtz free energy, covers a number of important practical cases. However, it belongs to the class of models which is difficult to investigate compared to the Landau-Devonshire-Ginzburg model. In the latter case an additional “smoothing” term in (5.1), known as the Ginsburg term γu_{xxxx} , allows us to obtain a bound of the deformation gradient (strain) using a well established technique [5].

Remark 5..1 A number of important characteristics of phase transformations (such as the size of hysteresis) may depend on the contributions of the interfacial energies. These contributions are often modeled with the Ginsburg correction term. However, the Ginsburg coefficient can only be determined in approximate order [28] and in the general case

this coefficient may not be temperature-independent. Another way to account for the contributions of interfacial energies is to take the free energy in the form [4, 17]

$$\Psi = (1-z)\bar{\psi}_1(\epsilon, \theta) + z\bar{\psi}_2(\epsilon, \theta) + z(1-z)\bar{\psi}_3, \quad (5.3)$$

where z is the volume fraction of martensite (i.e. the product phase), $(1-z)$ is the volume fraction of austenite (i.e. the parent phase), $\bar{\psi}_1, \bar{\psi}_2$ are the free energies of austenite and martensite respectively and $\bar{\psi}_3$ is the contribution from the interaction effect between austenite and martensite. We will not pursue these ideas in this paper.

Remark 5.2 Some authors include a linear term $\alpha^0\theta$ into $\psi_0(\theta)$. This term has no bearing on the final model and changes only the value of the coefficient of θ in the internal energy representation (see formula (6.2)), and thus is omitted.

In the general case for the choice of the function Φ_1 in (3.13) we allow the dependency of the stress on the rate of temperature and the deformation gradient

$$s = p \left[p(\theta, \epsilon) + \lambda \left(\frac{\partial \theta}{\partial t}, \frac{\partial \epsilon}{\partial t} \right) \right], \quad (5.4)$$

where

$$p(\theta, \epsilon) = \frac{\partial \Psi}{\partial \epsilon}, \quad \lambda \left(\frac{\partial \theta}{\partial t}, \frac{\partial \epsilon}{\partial t} \right) = \tilde{\mu}(\theta) \frac{\partial \epsilon}{\partial t} + \tilde{\nu}(\epsilon) \frac{\partial \theta}{\partial t}. \quad (5.5)$$

It is straightforward to deduce

$$p(\theta, \epsilon) = \alpha_2 \theta \epsilon + \frac{\partial \psi_3(\epsilon)}{\partial \epsilon} = [\alpha_2 \epsilon (\theta - \theta_1) - \alpha_4 \epsilon^3 + \alpha_6 \epsilon^5]. \quad (5.6)$$

6. One-Dimensional Hyperbolic Approximation of Shape-Memory-Alloy Dynamics

Using the model (5.1) and (5.2), from (3.12) we get

$$\eta = \alpha_1 (1 + \ln \theta) - \frac{1}{2} \alpha_2 \epsilon^2. \quad (6.1)$$

This enables us to find the internal energy of the system as a sum of thermal and mechanical fields contributions

$$e = \alpha_0 + \alpha_1 \theta - \frac{1}{2} \alpha_2 \theta_1 \epsilon^2 - \frac{1}{4} \alpha_4 \epsilon^4 + \frac{1}{6} \alpha_6 \epsilon^6 = \alpha_0 + \alpha_1 \theta + \psi_3(\epsilon). \quad (6.2)$$

The substitution of (6.2) into (4.4) leads to the final form of the energy balance equation. In particular, assuming symmetry of the deformation gradient tensor, we get

$$\rho \alpha_1 \left[\frac{\partial \theta}{\partial t} + \tau_0 \frac{\partial^2 \theta}{\partial t^2} \right] + A(\epsilon, \theta) - \nabla \cdot (k \nabla \theta) = G, \quad (6.3)$$

where the meaning of A is

$$A(\epsilon, \theta) = -\rho \alpha_2 \left\{ \theta \epsilon \frac{\partial \epsilon}{\partial t} + \tau_0 \frac{\partial}{\partial t} \left[\theta \epsilon \frac{\partial \epsilon}{\partial t} \right] \right\} - \rho \tilde{\mu}(\theta) \left\{ \left(\frac{\partial \epsilon}{\partial t} \right)^2 + \tau_0 \frac{\partial}{\partial t} \left[\left(\frac{\partial \epsilon}{\partial t} \right)^2 \right] \right\} - \rho \frac{\partial \theta}{\partial t} \left\{ \tilde{\nu}(\epsilon) \frac{\partial \epsilon}{\partial t} + \tau_0 \frac{\partial}{\partial t} \left[\tilde{\nu}(\epsilon) \frac{\partial \epsilon}{\partial t} \right] \right\}. \quad (6.4)$$

Equation (6.3) is solved together with the equation of motion (3.2) with respect to (u, θ) :

$$\begin{cases} C_v \left[\frac{\partial \theta}{\partial t} + \tau_0 \frac{\partial^2 \theta}{\partial t^2} \right] - k_1 \left[\theta \frac{\partial u}{\partial x} \frac{\partial^2 u}{\partial t \partial x} + \tau_0 \frac{\partial}{\partial t} \left(\theta \frac{\partial u}{\partial x} \frac{\partial^2 u}{\partial t \partial x} \right) \right] - \mu \left[\left(\frac{\partial^2 u}{\partial t \partial x} \right)^2 + \right. \\ \left. \tau_0 \frac{\partial}{\partial t} \left(\frac{\partial^2 u}{\partial t \partial x} \right)^2 \right] - \nu \left[\frac{\partial \theta}{\partial t} \frac{\partial^2 u}{\partial t \partial x} + \tau_0 \frac{\partial}{\partial t} \left(\frac{\partial \theta}{\partial t} \frac{\partial^2 u}{\partial t \partial x} \right) \right] - \frac{\partial}{\partial x} \left(k \frac{\partial \theta}{\partial x} \right) = G, \\ p \frac{\partial^2 u}{\partial t^2} - \frac{\partial}{\partial x} \left[k_1 \frac{\partial u}{\partial x} (\theta - \theta_1) - k_2 \left(\frac{\partial u}{\partial x} \right)^3 + k_3 \left(\frac{\partial u}{\partial x} \right)^5 \right] - \mu \frac{\partial^3 u}{\partial x^2 \partial t} - \nu \frac{\partial^2 \theta}{\partial x \partial t} = F, \end{cases} \quad (6.5)$$

where $C_v = \rho \alpha_1$, $k_1 = \rho \alpha_2$, $k_2 = \rho \alpha_4$, $k_3 = \rho \alpha_6$, $\mu = \rho \tilde{\mu}$, $\nu = \rho \tilde{\nu}$.

The initial conditions for the model (6.5) are chosen in the form

$$u(x, 0) = u^0(x), \quad \frac{\partial u}{\partial t}(x, 0) = u^1(x); \quad \theta(x, 0) = \theta^0(x), \quad \frac{\partial \theta}{\partial t}(x, 0) = \theta^1(x), \quad (6.6)$$

for given functions u^0 , u^1 , θ^0 , θ^1 . There are several distinct choices for boundary conditions to be used in our computational experiments. Mechanical boundary conditions are taken in one of the following forms (L is the length of the structure):

- “stress-free” boundary conditions: $s(0, t) = s(L, t) = 0$;
- “pinned end” boundary conditions: $u(0, t) = u(L, t) = 0$;
- or mixed mechanical boundary condition: $s(0, t) = 0$, $u(L, t) = 0$.

When displacements are given on boundaries, *a priori* bounds on strains are generally unknown which complicates the mathematical analysis of the problem. Computational results presented in Section 7 deal with this case. Thermal boundary conditions are chosen in one of the following form

- “thermal insulation” boundary conditions: $q(0, t) = q(L, t) = 0$, which reduce to $\frac{\partial \theta}{\partial x}(0, t) = \frac{\partial \theta}{\partial x}(L, t) = 0$ for the Fourier law (see (8.2));
- “controlled flux” boundary conditions: $\frac{\partial \theta}{\partial x}(0, t) = 0$, $-k \frac{\partial \theta}{\partial x}(L, t) = \beta[\theta - \theta^0(t)]$;
- or fixed temperature (“uncontrolled energy flow”) boundary conditions: $\theta(0, t) = \theta(L, t) = 0$.

In the last case additional assumptions are needed. By using the Leray-Schauder principle we have analysed the Cauchy problem for nonlinear hyperbolic model of thermoviscoelasticity (6.5). Our procedure makes use of the Lumer-Phillips theorem and the technique developed in [5]. We shall address details of this technique elsewhere.

Our final remark in this section goes to the choice of the function Ψ in the form (5.1) and (5.2) that brings major difficulties in the investigation of the model (6.5). Strictly speaking, the free energy function strongly depends upon the statistics of the phenomenon and has to be derived from a statistical model. Since van der Waals work on statistical mechanics it is a common practice to choose this function as a non-convex function of ϵ [14]. When dealing with shape memory alloys, minima of this function are known to correspond different phases of the material. For example, in the case of three minima, we expect one austenitic and two martensitic phase (see, for example, [8, 19, 28, 12]). Temperature plays a crucial role in the phase transition. Depending on the value of temperature, the material may alternate between a single thermodynamically unstable nonmonotone branch and multiple unstable branches. The character of this instability depends not only on the deformation gradient and temperature, but also on the rates of their changes.

7. Computational Experiments

In this section we present some numerical results on the thermal and mechanical control of a rod ($L = 1\text{cm}$) with a shape-memory-alloy core. The parameters of the Cu-based core are taken as follows

$$k = 1.9 \times 10^{-2} \text{cmg}/(\text{ms}^3\text{K}), \quad \rho = 11.1 \text{g}/\text{cm}^3, \quad C_v = 29 \text{g}/(\text{ms}^2\text{cmK}), \quad \theta_1 = 208\text{K}, \\ k_1 = 480 \text{g}/(\text{ms}^2\text{cmK}), \quad k_2 = 6 \times 10^6 \text{g}/(\text{ms}^2\text{cmK}), \quad k_3 = 4.5 \times 10^8 \text{g}/(\text{ms}^2\text{cmK}).$$

We use model (6.5) with $\tau_0 = 0 = \mu = \nu = 0$, initial conditions (6.6) and "pinned end & controlled flux" boundary conditions. This model was straightforwardly reduced to a differential-algebraic system in $\pi = (u, v, \theta)^T$ and stress s using second-order accurate spatial differences on staggered grids:

$$D \frac{\partial \pi}{\partial t} = \mathbf{f}, \quad s = k_1(\theta - \theta_1) \frac{\partial u}{\partial x} - k_2 \left(\frac{\partial u}{\partial x} \right)^3 + k_3 \left(\frac{\partial u}{\partial x} \right)^5, \quad (7.1)$$

where D is the diagonal matrix with $\text{diag}(D) = (1, \rho, C_V)$, $\mathbf{f} = (f_1, f_2, f_3)^T$ and

$$f_1 = v, \quad f_2 = \frac{\partial s}{\partial x} + F, \quad f_3 = k \frac{\partial^2 \theta}{\partial x^2} + k_1 \theta \frac{\partial u}{\partial x} \frac{\partial v}{\partial x} + G. \quad (7.2)$$

The developed code is robust and much simpler compared to computational procedures previously reported in the literature for shape-memory alloys [20, 15, 13].

Experiment 1 (thermal control of phase transformations). In this experiment we set uniform forcing $F = 500 \text{g}/(\text{ms}^2\text{cm}^2)$ and vary heating conditions given by $G = 375\pi \sin^3(t\pi/6) \text{g}/(\text{ms}^3\text{cm})$. We assume that the initial displacements are given in the form

$$u^0(x) = \begin{cases} -0.11809x, & 0 \leq x \leq 1/6, \\ 0.11809(x - 1/3), & 1/6 \leq x \leq 1/2, \\ 0.11809(2/3 - x), & 1/2 \leq x \leq 5/6 \\ 0.11809(x - 1), & 5/6 \leq x \leq 1 \end{cases} \quad (7.3)$$

and take the initial temperature as $\theta^0 = 200\text{K}$. Figure 1 (obtained with time step $7 \times 10^{-4}\text{ms}$ and space step $1/24\text{cm}$) demonstrates the transformation of $2M^+ + 2M^-$ martensites into an austenite (visible in the region of zero strain and displacements with superposed elastic vibrations as seen most clearly in the velocity field) after sufficient temperature has reached. Then upon cooling we observe a first order (martensitic) transition from the high temperature phase (austenite) to the low temperature phase (martensite). Upon the return to the low temperature regime the stable attractor with this applied thermomechanical forcing is not the original configuration but only two distinct martensite phases. The transformation $[2M^+ + 2M^-] \rightarrow A$ is accompanied by a decrease in temperature whereas the transformation $A \rightarrow [M^+ + M^-]$ is accompanied by an increase in temperature.

Experiment 2 (mechanical control of phase transformations). In this experiment we set $G = 0$, but vary the mechanical loading according to $F = 7000 \sin^3(\pi t/2) \text{g}/(\text{cm}^2\text{ms}^2)$. Starting from the austenite configuration ($u^0 = 0$) at intermediate temperature $\theta^0 = 255\text{K}$ we observe (see Figure 2 where the time step was $8 \times 10^{-4}\text{ms}$ and the space step was $1/16\text{cm}$) the transformation $A \rightarrow [M^+ + M^-] \rightarrow A \rightarrow [M^- + M^+] \rightarrow A$.

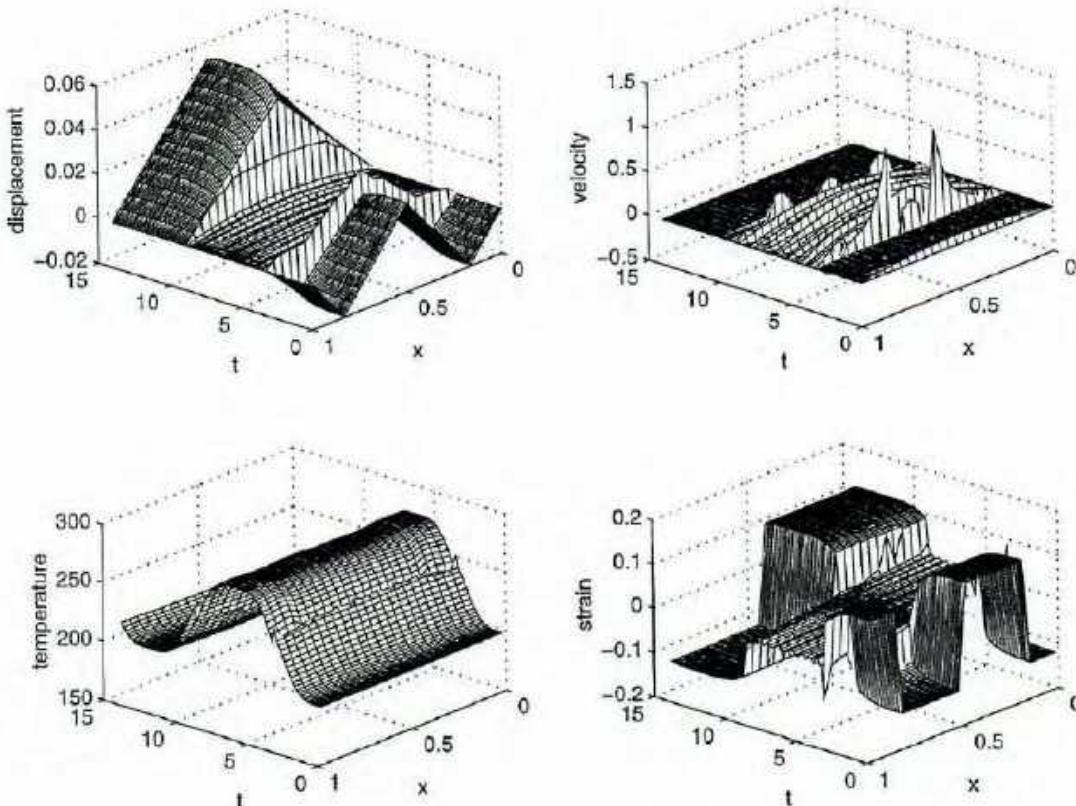


FIGURE 1. Thermally induced phase transformations.

In this experiment we observe the almost immediate transformations of austenite phases into two martensites upon the increase/decrease in loading. Note the relatively large heating/cooling associated with the transition into/out of martensite phase. A similar behaviour under different thermomechanical conditions was also observed in [20, 15]. In our code we have also incorporated the Ginsburg term by adding γu_{xxxx} to f_2 in (7.2). With reported values of the Ginsburg coefficient ($\gamma \sim 10^{-10} - 10^{-12}$) the Ginsburg term has a negligible effect on the thermomechanical behaviour of shape-memory alloys in the group of experiments described here. Accounting for interfacial energy contributions and the influence of mechanical and thermal dissipations on the dynamics of memory material require further investigation.

8. Construction of Approximate Models for Dynamic Thermoviscoelasticity Using Centre Manifold Theory

The model described in Section 6 will provide a good approximation of thermomechanical behaviour of a large shape memory alloy bar (see applications in [3]) only in the case the bar can be modelled by a thin rod with a shape memory alloy core. As an alternative to that model, in this section we construct a new model which is derived directly from the 3D model for shape memory alloy evolution (see (3.2), (6.3)) using centre manifold techniques (see, for example, [24]).

We assume that the shear stress in equation (3.2) is determined by its three components, the quasi-conservative component, s^q , the stress component due to mechanical dissipation, s^m , and the stress component due to thermal dissipations, s^t , (the latter is

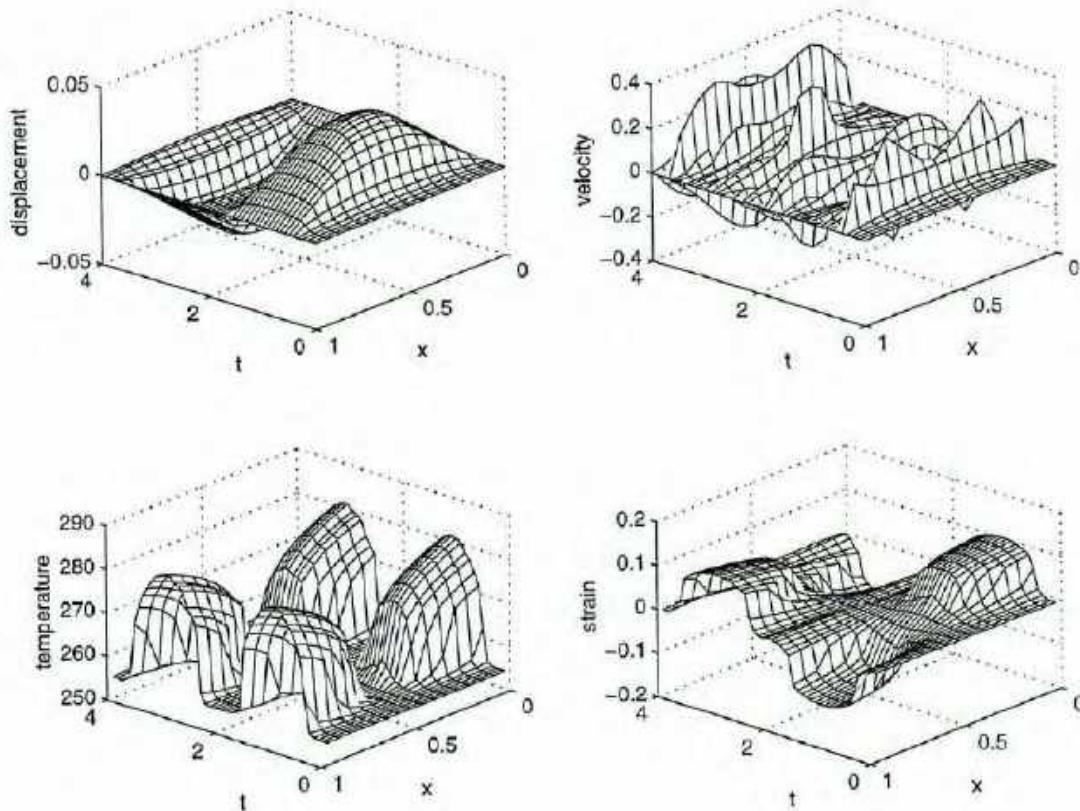


FIGURE 2. Mechanically induced phase transitions.

assumed to be negligible at this stage)

$$\mathbf{s} = \mathbf{s}^q + \mathbf{s}^m + \mathbf{s}^t, \quad \text{with} \quad \mathbf{s}^q = \rho \frac{\partial \Psi}{\partial \epsilon}, \quad \mathbf{s}^m = \rho \mu \frac{\partial \epsilon}{\partial t}, \quad \mathbf{s}^t = 0. \quad (8.1)$$

In the general case the heat flux is determined as the solution of equation (4.1). An approximation to this solution is provided by the following generalised form (see [19] and references therein)

$$\mathbf{q} = -k \nabla \theta - \alpha \frac{\partial k \nabla \theta}{\partial t}, \quad \alpha \geq 0, \quad (8.2)$$

which we will use with $\alpha = 0$ when (8.2) turns into the classical Fourier law.

The internal energy function e is defined from (3.8) by

$$e = \Psi - \theta \frac{\partial \Psi}{\partial \theta}. \quad (8.3)$$

In order to complete the formulation of the problem we specify a model for the free energy function Ψ . However, in the general 3D case one cannot use the shear strain as the order parameter as we usually do for the 1D case. One of the first approaches to deal with the 3D challenge was the Frémond model. This model uses different expressions for free energy functions for different phases (see, for example, [10]). All these expressions are essentially of the Landau-Ginsburg-type and contain the term $\gamma/2 \nabla \text{tr}(\epsilon)$ with $\gamma > 0$ introduced in order to smooth possibly very sharp spatial phase separation. In this

paper we use a different approach proposed in [9]. This approach generalises the classical Landau-Devonshire-Falk theory for shape memory alloys to the 3D case. The free energy function, based on the expansion up to sixth order in a single shear strain component [8], was extended to the three-dimensional case using the group theoretical approach proposed in [21]. In contrast to some other models (see, for example, [10]) strain-gradient terms are not involved in his expansion. We make use of this expansion and apply the following general representation of the free energy function

$$\Psi(\epsilon, \theta) = \psi^0(\theta) + \sum_{n=1}^{\infty} \psi^n(\epsilon, \theta), \quad (8.4)$$

where independent material parameters of the n -th order for $n = 1, 2, \dots$ are determined through strain invariants, \mathcal{I}_j^n , as follows

$$\psi^n = \sum_{j=1}^{j^n} \psi_j^n \mathcal{I}_j^n \quad \text{and} \quad \psi^0(\theta) = \psi_0(\theta). \quad (8.5)$$

The upper limit of the sum in (8.5), j^n , is the number of all invariant directions associated with a representation of the 48th order cubic symmetry group of the parent phase (see details in [9]). In order to adequately describe thermomechanical behaviour of shape-memory alloys we need to account for 6 terms in the sum of the expansion (8.4). In this case we have to determine 32 material parameters that make the application of formulae (8.4)–(8.5) fairly complicated. Using physically justified assumptions it is possible to reduce the number of required parameters. To achieve this, we make the same assumptions as in [9]. They conclude that odd degree invariants can be neglected in the expansion. Taking invariants up to the sixth order results in a representation with only 10 material constants which may depend on temperature

$$\Psi = \psi^0(\theta) + \sum_{j=1}^3 \psi_j^2 \mathcal{I}_j^2 + \sum_{j=1}^5 \psi_j^4 \mathcal{I}_j^4 + \sum_{j=1}^2 \psi_j^6 \mathcal{I}_j^6 \quad (8.6)$$

(we do not neglect the contribution of $\psi^0(\theta)$ as was done in [9]). The strain invariants \mathcal{I}_i^n of second, forth and sixth orders of the 48th order cubic symmetry group of the parent phase are

$$\begin{aligned} \mathcal{I}_1^2 &= \frac{1}{9}(\text{tr}(\epsilon_{ij}))^2, \quad \mathcal{I}_2^2 = \frac{1}{12}(2\epsilon_{33} - \epsilon_{11} - \epsilon_{22})^2 + \frac{1}{4}(\epsilon_{11} - \epsilon_{22})^2, \\ \mathcal{I}_3^2 &= \epsilon_{23}^2 + \epsilon_{13}^2 + \epsilon_{12}^2, \quad \mathcal{I}_1^4 = (\mathcal{I}_2^2)^2, \quad \mathcal{I}_2^4 = \epsilon_{23}^4 + \epsilon_{13}^4 + \epsilon_{12}^4, \quad \mathcal{I}_1^6 = (\mathcal{I}_2^2)^3 \\ \mathcal{I}_3^4 &= (\mathcal{I}_3^2)^2, \quad \mathcal{I}_4^4 = \mathcal{I}_2^2 \mathcal{I}_3^2, \quad \mathcal{I}_5^4 = \epsilon_{23}^2 \left[\frac{1}{6}(2\epsilon_{33} - \epsilon_{11} - \epsilon_{22}) - \frac{1}{2}(\epsilon_{11} - \epsilon_{22}) \right]^2 + \\ &\quad \epsilon_{13}^2 \left[\frac{1}{6}(2\epsilon_{33} - \epsilon_{11} - \epsilon_{22}) + \frac{1}{2}(\epsilon_{11} - \epsilon_{22}) \right]^2 + \frac{1}{9}\epsilon_{12}^2(2\epsilon_{33} - \epsilon_{11} - \epsilon_{22})^2, \\ \mathcal{I}_2^6 &= \frac{1}{36}(2\epsilon_{33} - \epsilon_{11} - \epsilon_{22})^2 \left(\frac{1}{36}(2\epsilon_{33} - \epsilon_{11} - \epsilon_{22})^2 - \frac{1}{4}(\epsilon_{11} - \epsilon_{22})^2 \right)^2. \end{aligned} \quad (8.7)$$

The ten material constants ψ_j^n in (8.6) differ from alloy to alloy and we use coefficients derived for Cu-based alloys [9] (units used here are consistent with those used in Section

7 for our numerical results on the dynamics of shape-memory alloys):

$$\begin{aligned}\psi_1^2 &= 5.92 \times 10^6 \text{ g}/(\text{ms}^2\text{cm}), \quad \psi_2^2 = (1.41 \times 10^5 + 46(\theta - 300)) \text{ g}/(\text{ms}^2\text{cm}), \\ \psi_3^2 &= (1.48 \times 10^6 - 940(\theta - 300)) \text{ g}/(\text{ms}^2\text{cm}), \quad \psi^0 = -\alpha_1 \theta \ln[(\theta - \theta_0)/\theta_0] \text{ g}/(\text{ms}^2\text{cm}), \\ \psi_1^4 &= (-1.182 \times 10^8 + 3.55 \times 10^5(\theta - 300)) \text{ g}/(\text{ms}^2\text{cm}), \\ \psi_2^4 &= 3.13 \times 10^9 \text{ g}/(\text{ms}^2\text{cm}), \quad \psi_3^4 = 1.64 \times 10^9 \text{ g}/(\text{ms}^2\text{cm}), \\ \psi_4^4 &= -5.53 \times 10^8 \text{ g}/(\text{ms}^2\text{cm}), \quad \psi_5^4 = -4.27 \times 10^8 \text{ g}/(\text{ms}^2\text{cm}), \\ \psi_1^6 &= 3.35 \times 10^{10} \text{ g}/(\text{ms}^2\text{cm}), \quad \psi_2^6 = 3.71 \times 10^{11} \text{ g}/(\text{ms}^2\text{cm}).\end{aligned}\tag{8.8}$$

Other material parameters are taken to be the same as those given in Section 7. We are interested in the construction of an adequate model for the description of thermo-mechanical behaviour of thin slabs in shape memory alloy materials. Starting from the 3D Falk-Konopka model and using centre manifold techniques (see, for example, [24]) we derive systematically an accurate low-dimensional model for the dynamics of the slab. The shape memory alloy is assumed to be of very large extent in the $x = x_1$ direction compared to its thickness of $2b$ in the $y = x_2$ direction ($-b < y < b$). For the sake of convenience we use a new temperature variable $\theta' = \theta - \theta_0$ where here $\theta_0 = 300$. For simplicity of the analysis we assume zero dissipation, $\alpha = \mu = 0$, and that there is no motion nor dependence in the x_3 direction.

A model for the dynamics of modes which vary slowly along the slab is derived for the unforced dynamics, $F = 0$, $G = 0$, and when “zero-stress & thermal-insulation” boundary conditions are specified on $y = \pm b$. The derivation of boundary conditions in the “long” direction x requires a quite delicate analysis [25] and these issues will not be addressed here. We only note that “pinned & insulating ends” boundary conditions may be used as a leading approximation. Modelling the long-wavelength, small-wavenumber modes along the slab, we neglect all longitudinal variations and look for eigenvalues of the cross-slab modes. It can be shown that generally there is a zero eigenvalue of multiplicity five and all the rest are pure imaginary (as dissipation has been omitted). Thus there exists a sub-centre manifold based upon these five modes (see [26] for an existence theorem), called a slow manifold as these five modes evolve slowly. Note that being on a sub-centre manifold the models we construct only have a weak assurance of asymptotic completeness, see the discussion in [7]. The zero eigenvalue of multiplicity five corresponds to longitudinal waves, large-scale bending, and one heat mode. The leading order structure of the critical eigenmodes are constant across the slab. Thus letting an overbar denote the y average, the amplitudes of the critical modes are chosen in the form

$$U_i(x, t) = \overline{u_i}, \quad V_i(x, t) = \overline{v_i}, \quad \Theta'(x, t) = \overline{\theta'}. \tag{8.9}$$

The low-dimensional model below is written in terms of these parameters.

The construction of the low-dimensional model is based upon the ansatz that there exists a low-dimensional invariant manifold upon which the amplitudes evolve slowly:

$$u_i = \mathcal{U}_i(\mathbf{U}, \mathbf{V}, \Theta'), \quad v_i = \mathcal{V}_i(\mathbf{U}, \mathbf{V}, \Theta'), \quad \theta = \mathcal{T}(\mathbf{U}, \mathbf{V}, \Theta'), \tag{8.10}$$

$$\text{where } \frac{\partial U_i}{\partial t} = V_i, \quad \frac{\partial V_i}{\partial t} = g_i(\mathbf{U}, \mathbf{V}, \Theta'), \quad \frac{\partial \Theta'}{\partial t} = g_\theta(\mathbf{U}, \mathbf{V}, \Theta'). \tag{8.11}$$

This ansatz is substituted into the differential-algebraic equations of 3D thermo-viscoelasticity (3.2), (3.5) and is solved to some order in the small parameters ∂_x , $E = \|U_x\| + \|V_x\|$ and $\vartheta = \|\Theta'\|$ with the computer algebra package Reduce. Thus, here we treat the strains as small, as measured by E , while permitting asymptotically large displacements and velocities. The displacement and temperature fields of the slow manifold, in terms of the amplitudes and the scaled transverse coordinate $Y = y/b$, are approximately

$$u_1 \approx U_1 - YbU_{2x} + 0.15(3Y^2 - 1)b^2U_{1xx}, \quad (8.12)$$

$$\begin{aligned} u_2 \approx & U_2 - (0.9 - 3.05e-5\Theta')YbU_{1x} + 0.15(3Y^2 - 1)b^2U_{2xx} \\ & - 141YbU_{1x}^3 + 1.00e-4(3Y - Y^3)b^3V_{1x}^2U_{1x}, \end{aligned} \quad (8.13)$$

$$\begin{aligned} \theta \approx & 300 + \Theta' - 2.43e6(3Y - Y^3)b^3(V_{1x}U_{2xx} + U_{1x}V_{2xx}) \\ & - 25.1(7 - 30Y^2 + 15Y^4)V_{1x}^3U_{1x}. \end{aligned} \quad (8.14)$$

These expressions have errors $\mathcal{O}(E^5 + \partial_x^{5/2} + \vartheta^{5/2})$ where the notation $\mathcal{O}(E^p + \partial_x^q + \vartheta^r)$ is used to denote terms involving $\partial_x^a E^b \vartheta^c$ such that $a/p + b/q + c/r \geq 1$. The mechanical and thermal field approximations represented by (8.12)–(8.14) have cross-slab structure. In particular, the sideways deformation u_2 (which is a nonlinear function of the longitudinal strains) of the shape memory alloy feed back at higher order to contribute to and complicate the longitudinal and thermal dynamics.

The model for the longitudinal dynamics on this slow manifold is

$$\begin{aligned} \rho \frac{\partial V_1}{\partial t} = & 2.97e6 U_{1xx} + 8.03e5 b^2 U_{1xxxx} \\ & + \partial_x [(922\Theta' - 0.0145\Theta'^2)U_{1x} - (4.28e9 - 1.31e7\Theta')U_{1x}^3 + 7.12e11 U_{1x}^5 \\ & + (2820 - 8.80\Theta')b^2V_{1x}^2U_{1x} + 1.24b^4V_{1x}^4U_{1x} - 5.42e4b^2V_{1x}^2U_{1x}^3] \\ & + \mathcal{O}(E^8 + \partial_x^4 + \vartheta^4). \end{aligned} \quad (8.15)$$

The first line in the right-hand side of (8.15) describes linear dispersive elastic waves along the slab, whereas the second line gives the temperature dependent quintic stress-strain relation of the shape memory alloy. Since $V_{1x} = U_{1xt}$, the remaining lines show effects upon this stress-strain relation due to rates of change of the strain.

Note that to this order of truncation there is no coupling to the bending modes of the slab which to the same error is simply the beam equation

$$\rho \frac{\partial V_2}{\partial t} = -9.91e5 b^2 U_{2xxxx} + \mathcal{O}(E^8 + \partial_x^4 + \vartheta^4). \quad (8.16)$$

There exists nonlinear coupling between the modes at higher order.

The corresponding energy equation for the temperature is

$$\begin{aligned} C_v \frac{\partial \Theta'}{\partial t} = & \kappa \Theta'_{xx} + (2.77e5 + 914\Theta' - 9.25\Theta'^2)U_{1x}V_{1x} \\ & + (3.94e9 + 1.26e7\Theta')V_{1x}U_{1x}^3 - (57.3 + 0.0117\Theta')b^2V_{1x}^3U_{1x} \\ & + 1.68e12V_{1x}U_{1x}^5 - 1.58e6b^2V_{1x}^3U_{1x}^3 - 0.0203b^4V_{1x}^5U_{1x} \\ & + 1.63e4b^2U_{1xx}V_{1xx} + 9.22e4b^2U_{2xx}V_{2xx} + \partial_x^2[-8151b^2U_{1x}V_{1x}] \\ & + \mathcal{O}(E^8 + \partial_x^4 + \vartheta^4). \end{aligned} \quad (8.17)$$

The first line in (8.17) describes the diffusion of heat generated or absorbed by mechanical strains, $\Theta U_{1x} V_{1x}$. However, in the thin slab the internal pattern of strains causes a much more complicated distribution of heating and cooling as summarised by the remaining lines. It is expected that virtually all of these should be retained in order to be consistent with the quintic stress-strain of the longitudinal wave equation. Computational experiments with the model derived in this section will be presented elsewhere.

9. Acknowledgements

The authors were supported by ARC Small Grant 179406. Special thanks go to Kerryn Thomas who contributed the results of computational experiments.

References

- [1] Anderssen, R.S., Götz, I. G. and K.-H. Hoffmann, The Global Behaviour of Elasto-Plastic and Visco-Elastic Materials with Hysteresis-Type State Equation, *SIAM J. Appl. Math.*, to appear.
- [2] Balta, F. and Suhubi, E. S., Theory of Nonlocal Generalised Thermoelasticity, *Int. J. Engng. Sci.*, **15**, 1977, 579–588.
- [3] Besselnik, P.A., Recent Development on Shape Memory Applications, *J. Phys. IV France, Colloque C5*, **7**, 1997, 581–590.
- [4] Bornert, M. & Muller, I., Temperature Dependence of Hysteresis in Pseudoelasticity, in *Free Boundary Value Problems*, Eds. K.-H. Hoffmann & J. Sprekels, Birkhauser Verlag, Basel, 1990, 27–35.
- [5] Chen, Z. and K.-H. Hoffmann, On a One-Dimensional Nonlinear Thermoviscoelastic Model for Structural Phase Transitions in Shape Memory Alloys, *Journal of Differential Equations*, **112**, 1994, 325–350.
- [6] Colli, P. and Grasselli, M., Justification of a Hyperbolic Approach to Phase Changes in Materials with Memory, *Asymptotic Analysis*, **10**, 1995, 303–334.
- [7] Cox, S.M. and Roberts, A.J., Initial conditions for models of dynamical systems, *Physica D*, **85**, 1995, 126–141.
- [8] Falk, F., Model Free Energy, Mechanics, and Thermodynamics of SMA, *Acta Metallurgica*, **28**, 1980, 1773–1780.
- [9] Falk, F. & Konopka, P., Three-Dimensional Landau Theory Describing the Martensitic Phase Transformation of SMA, *J. Phys.: Condens. Matter*, **2**, 1990, 61–77.
- [10] Fremond, M., Shape Memory Alloys. A Thermomechanical Model, in *Free Boundary Problems: Theory and Applications*, Eds.: K.-H. Hoffmann & J. Sprekels, Longman Scientific & Technical, 1990, 295–306.
- [11] Glass, D.E. and Tamme, K.K., Non-Fourier Dynamic Thermoelasticity with Temperature-Dependent Thermal Properties, *Journal of Thermophysics and Heat Transfer*, **8**, No. 1, 1994, 145–151.
- [12] Hoffmann, K.-H., Niezgodka, M. & Songmu, Z., Existence and Uniqueness of Global Solutions to an Extended Model of the Dynamic Developments in SMA, *Nonlinear Analysis: TMA*, **15**, No. 10, 1990, 977–990.
- [13] Hoffmann, K.-H. and Zou, J., Finite Element Approximations of Landau-Ginzburg's Equation Model for Structural Phase Transitions in Shape Memory Alloys, *M²AN*, **29**, No. 6, 1995, 629–655.
- [14] Huo, Y., Müller, I. and Seelcke, S., Quasiplasticity and Pseudoelasticity in Shape Memory Alloys, in *Phase Transition and Hysteresis*, ed. by Brakate et al, Springer-Verlag, 1994, 87–146.
- [15] Klein, O., Stability and Uniqueness Results for a Numerical Approximation of the Thermomechanical Phase Transitions in SMA, *Advances in Mathematical Sciences and Applications (Tokyo)*, **5**, No. 1, 1995, 91–116.
- [16] Melnik, R.V.N., Steklov's Operator Technique in Coupled Dynamic Thermoelasticity, *Numerical Methods in Thermal Problems*, Vol. X, Eds. R.W. Lewis & J.T. Cross, 1997, 139–150.
- [17] Moyne, S., Boubakar, M.L. & C. Lexcellent, Extension of a Linear Behaviour Model of SMA for Finite Strain Studies, *J. Phys. IV France, Colloque C5*, **7**, 1997, 83–88.
- [18] Müller, I. & T. Ruggeri, *Extended Thermodynamics*, Springer-Verlag, 1993.
- [19] Niezgodka, M. & Sprekels, J., Existence of Solutions for a Mathematical Model of Structural Phase Transitions in SMA, *Math. Methods in the Applied Sciences*, **10**, 1988, 197–223.

- [20] Niezgodka, M. & Sprekels, J., Convergent Numerical Approximations of the Thermomechanical Phase Transitions in SMA, *Numer. Math.*, **58**, 1991, 759–778.
- [21] Nittono, O & Y. Koyama, Japan. J. Appl. Phys., **21**, 1982, 680.
- [22] Renardy, M., Hrusa, W. J. and Nohel, J. A., *Mathematical Problems in Viscoelasticity*, Longman Scientific & Technical, 1987.
- [23] Racke, R. and Zheng, S., Global Existence and Asymptotic Behaviour in Nonlinear Thermoviscoelasticity, *Journal of Differential Equations*, **134**, 1997, 46–67.
- [24] Roberts, A.J., The invariant manifold of beam deformations. Part 1: the simple circular rod, *J. Elas.*, **30**, 1993, 1–54.
- [25] Roberts, A.J., Boundary Conditions for Approximate Differential Equations, *J. Austral. Math. Soc. Ser. B.*, **34**, 1992, 54–80.
- [26] Sijbrand, J., Properties of center manifolds, *Trans. Amer. Math. Soc.*, **289**, 1985, 431–469.
- [27] Sprekels, J., Global Existence for Thermomechanical Processes with Nonconvex Free Energies of Ginzburg-Landau Form, *J. of Math. Analysis and Appl.*, **141**, 1989, 333–348.
- [28] Sprekels, J., Shape Memory Alloys: Mathematical Models for a Class of First Order Solid-Solid Phase Transitions in Metals, *Control and Cybernetics*, **19**, No. 3–4, 1990, 287–308.

USQ



TOOWOOMBA

**MODELLING NONLOCAL PROCESSES
IN SEMICONDUCTOR DEVICES WITH
EXPONENTIAL DIFFERENCE
SCHEMES**

R V N Mehlak

Department of Mathematics & Computing, USQ

Hao He

Department of Theoretical Physics,

THE UNIVERSITY OF
S O U T H E R N
Q U E E N S L A N D

Working Papers

FACULTY
OF
SCIENCES

**MODELLING NONLOCAL PROCESSES
IN SEMICONDUCTOR DEVICES WITH
EXPONENTIAL DIFFERENCE
SCHEMES**

R V N Melnik

Department of Mathematics & Computing, USQ

Hao He

Department of Theoretical Physics,

School of Physics,

University of Sydney NSW

Faculty of Sciences Working Paper Series

SC-MC-9822

27 August 1998

MODELLING NONLOCAL PROCESSES IN SEMICONDUCTOR DEVICES WITH EXPONENTIAL DIFFERENCE SCHEMES

Part 1: Relaxation Time Approximations

R. V. N. Melnik *

Department of Mathematics and Computing,
University of Southern Queensland, QLD 4350, Australia

Hao He

Department of Theoretical Physics,
School of Physics, University of Sydney, NSW 2006

Abstract

In this paper we deal with nonlocal quasi-hydrodynamic mathematical models describing non-equilibrium physical processes in semiconductor devices. These processes cannot be adequately described with conventional drift-diffusion models. The primary numerical difficulty arises in the energy balance equation. Details of the discretisation for the continuity equations will be described along with a transformation of the energy balance equations to give computationally convenient forms. In a companion paper [20] we construct effective exponential difference schemes and apply them to modelling transport phenomena in semiconductors. Stability conditions, computational convergence and implementation of the proposed schemes are discussed with numerical examples.

Key words: time relaxation, quasi-hydrodynamic models, exponential difference schemes.

*Corresponding author, E-mail: melnik@usq.edu.au

1 Introduction

During recent years microelectronics has provided a wide range of challenging mathematical problems. Amongst them are problems in describing the electron-hole plasma in semiconductor devices, plasmo-chemical etching, ion lithography, fluid and gas epitaxy processes and crystal growth. From the mathematical physics viewpoint, a number of problems in computational microelectronics can be reduced to mathematical models involving stiff systems of ordinary differential equations and non-linear partial differential equations including systems of the Navier-Stokes type and the kinetic Boltzmann equations with its variants [28].

Technological advances in the field of microelectronics foster interdisciplinary research between mathematicians, physicists and engineers. The application of many classical algorithms to problems of computational electronics encounters serious mathematical difficulties and technological trends require continuous development of new and efficient numerical techniques. The problems of computational microelectronics become a challenge for applied mathematicians, and as a consequence, a great impetus to the further development of effective numerical methods.

The degree of integration in microelectronics and high configuration density with increasing power density of scattering lead to a situation where the problem of accounting for thermal regimes is critical in the design of microelectronic devices. This includes

- the analysis of thermoelectrical conditions of a device and the definition of functional characteristics accounting for local thermal regimes of each device on a substrate [18];
- accounting for the possibility of “self-heating” of devices.

Our main focus in this paper is the latter problem. The use of Extended Drift-Diffusion Models (EDDM) in the solution of this problem does not account for thermoflux of charge carriers. Typically, such models are obtained under the assumption of thermal equilibrium of charge carriers with the lattice. As a result, EDDM, similar to the classical drift-diffusion model, cannot describe today's semiconductor devices with sufficient accuracy.

In this work we consider and analyse non-local mathematical models which allow us to account for non-equilibrium effects and nonlocal processes in the electron-hole semiconductor plasma. However, an interplay between the oscillatory and diffusive character of transport processes causes major mathematical difficulties in studying transport phenomena which generally includes both parabolic and hyperbolic modes of dynamics. This requires nonlocal models that can describe a combined effect of long and short range forces.

We organise this paper as follows.

- In Section 2 we analyse mathematical models for the description of carrier transport in semiconductors as a hierarchy of models constructed on the basis of the relaxation-time concept.
 - In Section 3 we consider the quasi-hydrodynamic model and compare this model with the conventional drift-diffusion and kinetic models.
 - In Section 4 we focus on the normalisation procedure for the quasi-hydrodynamic model and give a review of some existing computational procedures.
 - Section 5 is devoted to problems of flux approximations for the continuity equation and Section 6 deals with extensions of such approximations to the energy balance equation.
 - Conclusions and future directions are discussed in Section 7.

2 Mathematical Models for Electron-Hole Plasma

2.1 Modelling transport phenomena in semiconductors

In the most general setting, mathematical modelling of transport phenomena, including transport phenomena in semiconductors, originated from the Liouville equation for the evolution of the position-velocity probability density $f(\mathbf{x}, \mathbf{v}, t)$:

$$\partial_t f + \mathbf{v} \cdot \nabla_{\mathbf{x}} f + \frac{1}{m^*} \mathcal{F} \cdot \nabla_{\mathbf{v}} f = 0, \quad t > 0, \quad \mathbf{x} \in \mathbb{R}_{\mathbf{x}}^{3M}, \quad \mathbf{v} \in \mathbb{R}_{\mathbf{v}}^{3M}, \quad (2.1)$$

where the position, $\mathbf{x} \in \mathbb{R}_{\mathbf{x}}^3$, and velocity vectors, $\mathbf{v} \in \mathbb{R}_{\mathbf{v}}^3$, of a charge carrier (say, the electron) are functions of time t , m^* is the effective carrier mass, M is the number of carriers in the system and \mathcal{F} is the driving force. Model (2.1) has to be supplemented by the conditions for the probability density

$$f(\mathbf{x}, \mathbf{v}, t) |_{t=0} \geq 0, \quad \int \int_{\mu} f(\mathbf{x}, \mathbf{v}, t) |_{t=0} d\mathbf{x} d\mathbf{v} = 1, \quad (2.2)$$

with the integration in (2.2) over the whole $6M$ dimensional (\mathbf{x}, \mathbf{v}) space, denoted by μ . The definition of the driving force in the context of semiconductor device theory typically has the following form [14]

$$\mathcal{F} = -q\mathbf{E} \text{ or } \mathcal{F} = -q(\mathbf{E} + \mathbf{v} \times \mathbf{B}), \quad (2.3)$$

where q is the elementary *positive* charge, \mathbf{E} is the electric field vector and \mathbf{B} is the magnetic induction vector.

In its essence, the model (2.1)–(2.3) is a reformulation of Newton's second law in terms of probability density. Using the Hamiltonian-canonical-system arguments this model can also be formulated in the (\mathbf{x}, \mathbf{p}) -space with $\mathbf{p} \in \mathbb{R}_{\mathbf{p}}^{3M}$ as the momentum vector. We also note that using the concept of primitive cells of the reciprocal lattice, known as Brillouin zones, we can incorporate quantum effects into the model. Such effects become important if the period of the crystal lattice is of the order of 10^{-8} cm. Then the ions in the crystal lattice induce a lattice potential which may significantly influence the motion of charged particles. Formally, the classical model (2.1)–(2.3) follows from quantum models in the limit of the reduced Planck constant $\hbar \rightarrow 0$. We assume that in model (2.1)–(2.3) quantum effects are taken indirectly into account by the effective mass of electrons.

If we exclude certain strong singularities of \mathcal{F} at finite \mathbf{x} , \mathbf{v} and t ("collisions of ensembles"), then the mathematical analysis of model (2.1)–(2.3) in reflexive Banach spaces (such as L^2) is known in the literature (see references in [14]). However in the general case, all Liouville-type models for semiconductor device modelling require the resolution of the following difficulties:

- $6M$ -dimensional μ -space is unrealistic for modelling many of today's devices;
- adequate models for the driving force as a combination of short-range and long-range interactions are not readily available [14].

In order to overcome these difficulties it is a common practice to use the Bogolubov-Born-Green-Kirkwood-Yvon (BBGKY) hierarchy for the description of transport phenomena.

First, in the BBGKY hierarchy we consider a limiting case of collisionless systems that mathematically corresponds to the situation when $M \rightarrow \infty$. In other words, those points that represent physical particles are smeared out, forming a “continuous” phase fluid in the space μ . Systems of this type are thought of as a large ensemble of weakly interacting particles and only long-range forces (like Coulomb forces) are considered. Mathematical model for such situations can be thought of as a single-particle Liouville equation supplemented by an effective field equation that represents the averaged effect of many-body physics. A typical model of this type is the Vlasov equation that can be written in terms of the probability of existence ($F(\mathbf{x}, \mathbf{v}, t)$) of a particle at the state (\mathbf{x}, \mathbf{v}) at time t as follows

$$\partial_t F + \mathbf{v} \cdot \nabla_{\mathbf{x}} F + \frac{1}{m^*} \mathcal{F}_{\text{eff}} \cdot \nabla_{\mathbf{v}} F = 0, \quad t > 0, \quad \mathbf{x} \in \mathbb{R}_{\mathbf{x}}^3, \quad \mathbf{v} \in \mathbb{R}_{\mathbf{v}}^3, \quad (2.4)$$

where \mathcal{F}_{eff} is defined analogously to (2.3). For example,

$$\mathcal{F}_{\text{eff}} = -q \mathbf{E}_{\text{eff}} \quad (2.5)$$

with the effective electric strain

$$\mathbf{E}_{\text{eff}}(\mathbf{x}, t) = \mathbf{E}_{\text{ext}}(\mathbf{x}, t) + \int_{\mathbb{R}_{\mathbf{x}}^3} n(\mathbf{x}, t) \mathbf{E}_{\text{int}}(\mathbf{x}, \bar{\mathbf{x}}) d\bar{\mathbf{x}}, \quad (2.6)$$

where $n(\mathbf{x}, t)$ is the number of charged particles per unit volume in an infinitesimal neighborhood of \mathbf{x} at time t , and \mathbf{E}_{int} and \mathbf{E}_{ext} are the internal and external electric strains respectively.

In the general case equations (2.4)–(2.6) are supplemented by the Maxwell system. If we consider only the Coulomb force, then

$$\mathbf{E}_{\text{int}} = -\frac{q}{4\pi\epsilon} \frac{\mathbf{x} - \mathbf{y}}{\|\mathbf{x} - \mathbf{y}\|^3}, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}_{\mathbf{x}}^3, \quad \mathbf{x} \neq \mathbf{y}. \quad (2.7)$$

Under the assumption that $\mathbf{E}_{\text{eff}} = -\nabla_{\mathbf{x}}\varphi$ (φ is the electric field potential), the effective field equation for an electron system is reducible to the Poisson equation

$$-\epsilon\epsilon_0\Delta\varphi = q(N - n), \quad (2.8)$$

where $N(\mathbf{x}, t)$ is the overall concentration of doping (accounting for background ions), and ϵ and ϵ_0 are the dielectric permittivities of the considered material and vacuum respectively.

We note that in the semiconductor context, function $F(\mathbf{x}, \mathbf{v}, t)$ from (2.4) may be interpreted as the number of charged particles per unit volume in an infinitesimal neighborhood of (\mathbf{x}, \mathbf{v}) at time t . Equation (2.4) is the nonlinear equation with nonlinearity defined by the effective field equation such as (2.8). The problem of dimensionality is overcome for model (2.4)–(2.8), but the difficulty with this model lies in the definition of the driving force. Indeed, the integration of equation (2.4) leads to an idealised macroscopic conservation law

$$q\partial_t n - \text{div}\mathbf{J} = 0, \quad (2.9)$$

where $\mathbf{J} = -q \int_{\mathbb{R}_{\mathbf{v}}^3} \mathbf{v} F d\mathbf{v}$ is the current density. Equation (2.9) can be satisfied only in a collisionless environment. In the general case, model (2.4)–(2.8) is not appropriate in the large-time scale modelling.

More realistic in the semiconductor device context is the consideration of systems with collisions. Indeed, for sufficiently large time scales, the motion of carriers decisively depends on scattering (i.e. on the short-range forces, which in classical situation leads to particle collisions). Scattering effects can only be included *statistically* in the collision operator $Q(F)$ of the Boltzmann equation written with respect to the distribution function (the number density) for which we use the same notation $F(\mathbf{x}, \mathbf{v}, t)$:

$$\partial_t F + \mathbf{v} \cdot \nabla_{\mathbf{x}} F + \frac{1}{m^*} \mathcal{F}_{\text{eff}} \cdot \nabla_{\mathbf{v}} F = Q(F). \quad (2.10)$$

Nonlinearity in equation (2.10) is defined by the effective field model and by the model for collisions. Originally introduced for dilute gases, equation (2.10) is based on the observation that the rate of change of F , caused by the effective force \mathcal{F}_{eff} , vanishes along the characteristics defined by Newton equations of motion

$$\frac{d\mathbf{x}}{dt} = \mathbf{v}, \quad \frac{d(m^*\mathbf{v})}{dt} = \mathcal{F}_{\text{eff}}. \quad (2.11)$$

As a result, we think about the collision effect as the *instantaneous* scattering from one state to another through a very fast change $\|\Delta\mathbf{v}\|$ in the velocity (momentum) vector and a very slow change $\|\Delta\mathbf{x}\|$ in the position vector. Formally, the situation when

$$\|\Delta\mathbf{v}\| \rightarrow \infty, \quad \text{and} \quad \|\Delta\mathbf{x}\| \rightarrow 0 \quad (2.12)$$

is not excluded. Strictly speaking, we have to show that the Pauli principle, asserting that *two electrons cannot occupy the same state (\mathbf{x}, \mathbf{v}) at the same time*, is not violated. In the general case it might be a difficult task which is intrinsically connected with the stability issues of the mathematical model. For example, if the considered system is a Coulomb system, then one has to show that there are at most two electrons per volume $(2\pi)^3$ in the phase space.

Fortunately, in applications the situation (2.12) is prohibited by allowing for *non-zero relaxation time*. Then the Pauli principle can be satisfied if the collision operator is chosen in the following form [14]:

$$Q(F) = \int_B \{ s(\mathbf{x}, \mathbf{v}', \mathbf{v}) F(\mathbf{x}, \mathbf{v}', t) [1 - F(\mathbf{x}, \mathbf{v}, t)] - s(\mathbf{x}, \mathbf{v}, \mathbf{v}') F(\mathbf{x}, \mathbf{v}, t) [1 - F(\mathbf{x}, \mathbf{v}', t)] \} d\mathbf{v}', \quad (2.13)$$

where s is the scattering (or transition) rate for a particle in the position \mathbf{x} from velocity \mathbf{v}' to velocity \mathbf{v} at time t . The velocity of a particle in (2.13) is a function of the wave vector $\mathbf{k} = \mathbf{p}/\hbar$ (\mathbf{p} is the momentum vector) which corresponds to a specific energy band of the Brillouin zone B (see [35]), and hence the rate of transition can be considered as a function of \mathbf{x} and \mathbf{k} . Then the relaxation time (average time between consecutive collisions at (\mathbf{x}, \mathbf{k})) can be introduced as follows

$$\tau(\mathbf{x}, \mathbf{k}) = 1/\lambda(\mathbf{x}, \mathbf{k}), \quad \mathbf{k} \in B \quad \text{where} \quad \lambda = \int_B s(\mathbf{x}, \mathbf{k}, \mathbf{k}') d\mathbf{k}' \quad (2.14)$$

where integral (2.14) is taken over the Brillouin zone of the lattice.

The well-posedness of model (2.10) with the effective field equation defined by (2.8), is typically discussed in the literature under quite excessive smoothness requirements on function s . However, in reality transition rates are highly non-regular functions. The type of non-linearity of the model, as well as the regularity of the transition rates, depend on the mechanism of scattering that can be formalised mathematically only using physical parametrisation. Today's technology develops devices with active regions of characteristic dimensions below $1 \mu\text{m}$ that operate in electric fields $\sim 10^6 \text{ V/cm}$, leading to serious mathematical modelling difficulties. Since quantum effects may not be negligible in these situations, in principle we have to describe arbitrary mixed quantum states that may not be represented by a single wave function. In addition, the energy-wave vector function may have several minima (energy-valleys) and we usually have to approximate the energy-band structure. As a result, common assumptions on the parabolic band structure (formally obtainable by scaling procedures) and on the equality of effective masses in the different directions cannot be rigorously justified in the semiconductor context.

Due to the sensitivity of model (2.10) to the definition of $Q(F)$, which is the subject of approximation, the success in modelling semiconductor devices essentially depends on the consistency of function $\tau(\mathbf{x}, \mathbf{k})$ to a specific practical situation. The applicability range of mathematical models in semiconductor device theory is eventually determined by certain functional relationships between function $\tau(\mathbf{x}, \mathbf{k})$ and other characteristics of semiconductor plasma.

2.2 Classification of models on the basis of relaxation-time concepts

Physical properties of semiconductor plasma are characterised by a number of fundamental lengths, such as

- De-Broglie wave length, $\lambda = h/(m^* \tilde{v})$, where \tilde{v} is the characteristic velocity of charge carrier motion and $h = 2\pi\hbar$;
- the length of momentum (impulse) relaxation, i.e. the length of the free mean path with respect to the momentum, $\lambda_p = \tilde{v}\tau_p$, where τ_p is the momentum relaxation time (the time that describes the exchange of (quasi-)momentum between carriers and the crystal lattice);
- the length of energy relaxation or the length of “cooling”, $\lambda_\omega = \tilde{v}\sqrt{\tau_p\tau_\omega}$, where τ_ω is the energy relaxation time (the time that describes the exchange of energy between carriers and the crystal lattice).

We consider devices with characteristic dimension l for which at least one of the following inequalities holds

$$l \gg \lambda, \quad l \gg \lambda_p, \quad l \gg \lambda_\omega. \quad (2.15)$$

Strictly speaking, if any of inequalities (2.15) is violated and l is commensurate with the fundamental lengths defined above, quantum effects may essentially influence the electric characteristics and parameters of devices such as hetero-structures with selective doping, devices with quantum holes and heterojunctions, and thin-layer MOS devices. In such cases,

model (2.10) with an effective-field equation (for example, (2.8)) has to be supplemented by the Schrödinger equation [31].

As follows from the definitions of $\lambda, \lambda_p, \lambda_\omega$, in a specific practical situation the choice of model strongly depends on values of \bar{v} , i.e. on the mechanism of scattering. Within a large range of temperatures in many applications $\lambda_p \ll \lambda_\omega$, for example, under scattering on acoustic phonons, we expect $\tau_p \ll \tau_\omega$. Surprisingly, the range of applicability of kinetic models may lie outside this inequality. Therefore, modelling semiconductor devices with kinetic models (the process that typically require the application of costly computational procedures) may not always be justified. In addition, the solution of kinetic models often contains a great deal of redundant information. Computation with the complete kinetic model is relatively efficient only when pair collisions of charge carriers weakly influence the charge transfer. However, if the frequency of pair collisions is fairly high (that is the case for large concentrations, $n \geq 10^{14} \text{ cm}^{-3}$ and higher), then modelling of devices using kinetic models involves considerable difficulties.

The relaxation of mathematical models i.e. semiconductor device theory may be provided by comparing the role of collisions with other scattering mechanisms. In this case we have to define the range of model applicability with respect to the mean time between the collision introduced by (2.14) that characterises the momentum-and-energy exchange speed. Initially, this consideration leads to two limiting cases that are discussed below.

- Kinetic models (KM) may be efficient in the case when

$$\tau_p \leq \tau_\omega \ll \tau. \quad (2.16)$$

In this case scattering of carriers on each other is not essential. Charge carriers cannot be considered as an independent thermodynamical system, because the scattering of carriers on imperfections of the lattice plays the dominant role. This may include momentum scattering on charged impurity ions, as well as on acoustic/piezoelectric and optical phonons.

- Hydrodynamic models (HDM) are confined to the case

$$\tau \ll \tau_p \ll \tau_\omega, \quad (2.17)$$

when carriers have enough time to exchange by energy and by momentum before the scattering on phonons (and other lattice impurities) becomes essential. In this case the electron-hole plasma (EHP) can be considered as an almost independent thermodynamical system that only weakly interacts with the lattice. We do not require that the temperature of lattice, T_l , should be equal to the carrier temperature (electron temperature, T_n , or hole temperature, T_p), but we think of the motion of the carrier system as a whole with respect to the lattice. Using analogy with fluid dynamics the models based on this reasoning are referred to as hydrodynamic. In these macroscopic models physical quantities are averaged over the whole carrier (electron/hole) population, and the sought-for information is substantially reduced compared to kinetic models. However, it is well known that locally such models may not correctly describe many important physical process such as impact ionisation caused by the influence of hot carrier subpopulations [30].

2.3 Stability issues for mathematical models with collision operators

As follows from (2.16) and (2.17), kinetic and hydrodynamic models belong to two distinct and generally non-overlapping classes of mathematical models for semiconductor device modelling. Strictly speaking, neither perturbation techniques nor the method of moments can lead to the rigorous derivation of hydrodynamic models from the Boltzmann kinetic equation unless simplified physical assumptions are made. Nevertheless, a reduction of kinetic equations to a low-dimensional system is computationally desirable.

The perturbation technique, used for such a reduction, is usually based on the expansion of the solution with respect to a dimensionless parameter, typically the scaled mean free path defined as $\alpha = \lambda_p/l$. This technique, known as the Hilbert expansion, works well for small electric fields. More precisely, the Boltzmann equation after rescaling has the following form

$$\alpha(\partial_t F + \mathbf{v} \cdot \nabla_{\mathbf{x}} F) - \mathbf{E}_{\text{eff}} \cdot \nabla_{\mathbf{v}} F = Q(F). \quad (2.18)$$

The chosen time scaling for the Boltzmann equation defines a relationship between the collision operator $Q(F)$ and the driving force \mathcal{F} [14]. In this case, F_0 , the leading term in the expansion $F = \sum_{i=0}^{\infty} \alpha^i F_i$, is the Fermi-Dirac distribution and has to satisfy the following equation

$$-\mathbf{E}_{\text{eff}} \cdot \nabla_{\mathbf{v}} F_0 = Q(F_0). \quad (2.19)$$

Unfortunately, this relationship may lead to the runaway phenomenon, occurrence of which depends on the collision frequency, i.e. on the physical mechanisms of scattering. This is a clear indication of the need to correct the leading term F_0 in the Hilbert expansion. However, equations (2.18) and (2.19) (or a modification of the latter) are not independent of each other and form a system of coupled equations. This leads to considerable difficulties in finding a correction to F_0 . Resolving these difficulties in a semiconductor device context ultimately leads to a hyperbolic-type equation for concentrations, the solution of which may have discontinuities. These discontinuities may be formally eliminated via “viscosity” arguments similar to the derivation of the Navier-Stokes system from the kinetic equations. Although in many applications such a parabolic smoothing is often acceptable from the engineering point of view [15], its rigorous justification in the semiconductor context requires a certain connection between diffusion coefficients, D_n, D_p , and the drift mobility of carriers μ_n, μ_p . For example, if thermal equilibrium is assumed with the absolute carrier temperature T , then such a connection is often described by the classical Einstein relation

$$D_n/\mu_n = D_p/\mu_p = \varphi_T, \quad (2.20)$$

where $\varphi_T = k_b T / q$ is referred to as the thermal voltage (or thermal potential) and k_b is the Boltzmann constant [20]. Using appropriate assumptions, different modifications and generalisations of (2.20) have been proposed in the literature [16]. However, for sufficiently high electron fields the dependency (2.20) as well as its modifications can be violated. Hence, it is quite natural to require that for the description of non-equilibrium and non-local processes in semiconductor plasma, high field phenomena have to be modelled in a way compatible with experiment [7].

During recent years attempts have been made to improve hydrodynamic models using the method of moments and taking into account moments of higher orders [39]. This allows us to take into account the energy flow. Attempts have also been made to obtain new improved expressions for the current density [7]. In the final analysis, in order to rigorously derive the hydrodynamic model or its modifications from kinetic equations one has to know some *a priori* information on the solution of the Boltzmann equation and to use mathematical assumptions compatible with the physical situation under investigation. Let us consider, for example, the electro-hydrodynamic model for an electron system (see [3, 26, 6] and references therein):

$$\frac{\partial \mathbf{z}}{\partial t} = \boldsymbol{\zeta} + \left(\frac{\partial \mathbf{z}}{\partial t} \right)_{\text{col}}, \quad (2.21)$$

where

$$\mathbf{z} = (n, \mathbf{v}, W)^T, \quad \boldsymbol{\zeta} = (\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3)^T, \quad \mathcal{F}_1 = -\nabla \cdot (n\mathbf{v}), \quad (2.22)$$

$$\mathcal{F}_2 = -\mathbf{v} \cdot \nabla \mathbf{v} - q\mathbf{E}_{\text{eff}}/m_n^* - \nabla(nT_n)/(m_n^* n), \quad (2.23)$$

$$\mathcal{F}_3 = -\nabla \cdot (\mathbf{v}W) - qn\mathbf{v} \cdot \mathbf{E}_{\text{eff}} - \nabla \cdot (\mathbf{v}nT_n) - \nabla \cdot \mathbf{q}, \quad (2.24)$$

T_n is the electron temperature given in energetic units, n is the electron concentration, \mathbf{v} is their averaged velocity, W is the energy density (typically modelled by $W = 3nT_n/2 + mn\|\mathbf{v}\|^2$), \mathbf{q} is the heat flow (typically modelled by the Fourier law $\mathbf{q} = -k\nabla T_n$) and m_n^* is the effective electron mass.

The system (2.21) couples electrical, mechanical and thermal fields. The first term in the RHS of (2.21) approximates the thermo-electromagnetic field effect, whereas the second term is of a “mechanical” nature and approximates collisions caused by lattice vibrations, impurities, crystal imperfections etc. Equations (2.21) have to be supplemented by the field equation, for example the Poisson equation (2.8). Although equations (2.21) are similar to the Euler equations, in contrast to the latter they include source terms modelling relaxation processes and electric field effects. In the general case, the type of the differential equations (2.21) changes with respect to the functional dependency between “collision terms”, $(\partial_t n)_{\text{col}}$, $(\partial_t \mathbf{v})_{\text{col}}$ and $(\partial_t T_n)_{\text{col}}$. These terms cannot be evaluated explicitly. In applications, the collision terms are typically approximated by relaxation-time approximations. In such cases hierarchy of the mathematical models with respect to the dependencies between τ , τ_ω and τ_p (see Section 2.2) is the most natural.

The roots of the main difficulties arising in semiconductor device modelling lie in the adequate definition of the approximate relationship between the collision operator and the driving force. In its essence, such an *approximate* relationship determines the range of the mathematical model applicability in the solution of practical problems. For example, the reduction of kinetic equations to hydrodynamic-type models such as (2.21) may be reasonably justified only for systems with strong collision. Assuming that this is the case, we can model semiconductor plasma using the analogy with an ensemble of interacting particles. A similar situation takes place in fluid dynamics when we model non-viscous non-compressible fluid (vortices). In the reduction procedure we ignore part of the information about the system. This results in the difficulties well known in Multiparticle Quantum Theory (MQT) [9].

When studying a large dynamic system it is important to have information on its ground state energy, i.e. information on the eigenfunction that corresponds to the lowest eigenvalue. In the reduction to hydrodynamic models we modify the original problem considering “particle cloud” motion instead of the interaction between particles. This may lead to a significant deviation between the results obtained with hydrodynamic models and experiments [30]. If we consider a large Coulomb system, the shape of the cloud is typically obtained assuming that the system of particles in its ground state behaves like a Thomas-Fermi gas, i.e. like a classical gas supplemented by the Pauli principle. However, the Thomas-Fermi theory gives only the leading asymptotic terms of the ground state system energy, which represents the quasi-classical energy, whereas the second term represents the quantum spectrum of Coulomb singularities. In this case, the well-posedness of the mathematical formulation of the problem can be established if one can show the positiveness of the system Hamiltonian. In turn, this can be established *only* under certain relationships between the fine structure constant, $\alpha_f = q^2/\hbar c$, or the scaled mean free path between two scattering events ([14], p.86), α , the number of nuclei, K , their charges, Z , and the number of electrons N in the system. From the practical point of view, if the perturbation technique is used, one has to obtain a lower bound for the “perturbation parameter” α , which is a function of Z , N and K . In the general case, without such a bound one cannot guarantee that the system is stable. Since in practice the value of N and K are typically finite, it is the connection between α (α_f) and Z that holds the key to the stability problem solution in applications (see, for example, [11, 12]).

A useful simplification of model (2.21), well investigated mathematically, provides the drift-diffusion model (DDM). However, the derivation of the DDM is usually based on a version of the Hilbert expansion and can be rigorously justified only for low carrier densities and small electric fields. The name of the model came from the type of dependence of the current densities on carrier densities and electric field. For the DDM the current densities are the sums of drift terms (with the mobilities μ_n and μ_p) and diffusion terms (with the diffusion coefficient D_n , D_p) which are connected by the Einstein-type relation (see (2.20)). As is the case for hydrodynamic-type models, mobility coefficients in the DDM cannot be evaluated explicitly. In addition, the scaling of the problem (and as a result the definition of a perturbation parameter α in the Hilbert expansion) decisively depends on the choice of the reference time and the reference field strength (i.e. on the ratio between the thermal voltage, φ_T , and the reference length, l_0). In modern applications of semiconductors the reference field strength may be very large. In such cases the analysis of the problem based on the Hilbert expansion cannot adequately account for high field effects. Strictly speaking, the DDM can only be applied when $\alpha \rightarrow 0$. However, through the technological advances connected with the miniaturization and the use of materials other than silicon the mean path becomes larger compared to the device size. As a result, the development of the next generation of mathematical models and numerical methods for their solutions has become an important challenging problem in applied mathematics.

3 Quasi-Hydrodynamic Models

3.1 Assumptions and the range of applicability

A certain compromise between the model types described in the previous section gives quasi-hydrodynamic models. In fact, there is a number of reasons in favor of the development of models other than hydrodynamic, kinetic and drift-diffusion types. Firstly, conditions for application of hydrodynamic models are quite restrictive from the physical point of view. Such conditions can be justified under conditions of strong injection or in application to low-bandgap materials when the intrinsic concentration of charge carriers is very large. Secondly, the application of kinetic models is connected with essential computational difficulties. Thirdly, although mathematical investigation of drift-diffusion type models has achieved some maturity, these models cannot predict important physical phenomena such as carrier heating or velocity overshoot. By now it is clear that drift-diffusion type models are not compatible with technological advances. Moving to the next generation of mathematical models in this field means accounting for *non-equilibrium* and *non-local* behaviour of semiconductor plasma.

A wide area of applications is confined to the situation, which may not overlap with (2.16) or (2.17), when

$$\tau_p \leq \tau \ll \tau_\omega. \quad (3.1)$$

From the physical point of view this means that charge carriers have enough time to repeatedly exchange by energy (but not by momentum!) before the scattering on phonons becomes essential. Plasma of charge carriers achieves its equilibrium after time τ , i.e. long before the time when the exchange between carriers and the lattice becomes noticeable. Hence, in this case our perception on carrier temperature is quite definite. Indeed, with respect to the energy, plasma of charge carriers can be considered as an *almost independent* thermodynamical system. Of course, it is not true any more with respect to the momentum because the scattering takes place mainly on impurities of the lattice. This approximation leads to mathematical models of quasi-hydrodynamic type. The area of applicability of such models is wider than that of hydrodynamic models, although physical simplifications connected with the application of quasi-hydrodynamic type models are similar to those for hydrodynamic models. Typically, we assume

- “small anisotropy”, i.e. we require

$$m_n^* \|v\|^2 / 2 \ll 3T_n / 2, \quad (3.2)$$

- quadratic law of dispersion and the parabolicity of the energy bands, i.e. that the effective masses of carriers are scalar constants (m_n^* and m_p^* for the electron and the hole respectively).

Both of these assumptions may be questionable. However, the first assumption allows us to represent an approximation to the average energies in the form

$$\langle \mathcal{E}_n \rangle = 3T_n / 2, \quad \langle \mathcal{E}_p \rangle = 3T_p / 2, \quad (3.3)$$

and implies that all quantities that depend on $\langle \mathcal{E}_n \rangle$ (or $\langle \mathcal{E}_p \rangle$) become functions of temperature (for example, $\mu_n = \mu_n(T_n)$, $D_n = D_n(T_n)$ etc). The second assumption helps us

to deal with multi-extrema situations in energy-wave vector-functions. In the general case, much complexity in this area of mathematical modelling stems from the multi-temperature character of semiconductor plasma. It is a well-known fact that the energy spectrum of electrons in semiconductors (including unipolar field-effect transistors of submicron-size or Si-MOS structures) has a multi-valley character. This leads to the difference in dynamic properties of carriers that belong to different valleys. Such a difference may be considerable in high electro-magnetic fields. This phenomenon can be explained physically with respect to a specific material. For example, for GaAs semiconductors a difference in effective masses and locations of energy minima of the central Γ -valley and lateral L - and X -valleys may play a crucial role. In this case, assumptions on the equality of effective masses in different directions can be hardly justified ([14], p.69). Another difficulty lies in the differences of dynamic properties of carriers that belong to different valleys. They may be connected with differences in the orientation of these valleys with respect to the electro-magnetic field (as is the case for Si-based devices). Whenever it is the case, we have several options. For example, we may use the kinetic equation for the description of carriers in different valleys of the conductance band. Alternatively, we may use a multi-temperature quasi-hydrodynamic model. In the latter case we assume that carrier collisions guarantee "almost Maxwellisation" of carriers that belong to different valleys and that EHP is non-degenerate, which means that the energy density of carriers (for example, electrons) in the i th valley is defined by $(\bar{\mathcal{E}}_n)_i = 3n_i(T_n)_i/2$ (as before the temperature is given in energy units). Since in this case collisions lead to the relaxation to a inter-valley balance (which is defined by the common effective temperature of carriers of each sign) it is reasonable to consider one-component (for each type of carriers) models. This agreement will be adopted for our further consideration.

3.2 Limiting cases and the comparison with other models

It is common practice to use an analogy with conservation laws for the physical interpretation of hydrodynamic equations such as (2.21). Mathematically such laws may be satisfied only approximately and the first three moments of the Boltzmann equation may not be sufficient for the adequate description of semiconductor physics. Different attempts to extend the hydrodynamic model by including energy flux conservation have been recently reported in the literature [39].

In this paper we follow a different direction. It is clear that for modelling of a number of non-equilibrium processes we have to use the energy balance law which can be derived as a third moment of the kinetic equation. It is reasonable to supplement the fundamental semiconductor system by this equation without the momentum conservation equation. Of course, in this case in the continuity equation we have to introduce thermal dependencies of coefficients on the energy variable (3.3). The first model of this type was introduced by Stratton [34] for taking into account "hot carrier" effects and for more accurate estimation of their contribution to the current. This model, often referred to as the Energy Balance Model (EBM) or the Hot-Carrier Transport Equations (HCTE), has been developed by many authors. Attempts have also been made to use different types of Simplified Hydrodynamic Models (SHDM) (see [38, 1, 13] and references therein). If relaxation times are obtained by fitting specified velocity-field and temperature-field characteristics, the distinction between EBM and SHDM loses its significance [1]. Both types of models, EBM and SHDM, may be referred as quasi-hydrodynamic models (QHDM). For models of *quasi-hydrodynamic* type

mobility becomes a function of carrier energy rather than local electric field strength as in the DDM. Mathematically this leads to a strong coupling between continuity and balance energy equations. As a result of this coupling serious mathematical and computational difficulties arise in the analysis of such models. These difficulties have their origin in physical models which require the description of essentially non-equilibrium behaviour of electron-hole plasma with parameters (such as concentrations, velocity, energy) that are non-locally connected with the electric field strength.

In practice we often observe a good agreement between the results obtained from HDM and QHD models [33]. This can be explained by the fact that in most practical situations the condition $\tau_p \ll \tau_\omega$ will be satisfied and this condition is the same for both models. Hence, although internal physical processes predicted by the classical drift-diffusion models and EBMs, HDMs, SHDMs are different in principle, such a difference may not manifest itself on the output characteristics of devices. As a result, there are many artificial approaches for the modification of DDM. In the majority of cases assumptions made under such approaches can be hardly justified (for example, the assumption on the equality between carrier temperature and lattice temperature or neglecting thermoflux of carriers). The area of application of DDM is restricted at least by elements with dimensions exceeding lengths of relaxation for momentum and energy. This restriction does not allow us to account non-equilibrium and non-locality of EHP. If effective temperatures of semiconductor structure are considered as local functions of electro-magnetic field (it may often be acceptable approximation if (2.15) holds), quasi-hydrodynamic model turns into DDM for which carrier velocity is a local function of electric field. For example, for the electron system we may have

$$\mathbf{v}_n = \mu_n(\mathbf{E}_{\text{eff}})\mathbf{E}_{\text{eff}}, \quad (3.4)$$

where $\mu_n = q\tau_p/m_n^*$ is the low-field mobility and τ_p , as above, the moment relaxation time. We recall that the mobility coefficient μ_n is determined by scattering mechanisms (i.e. carrier collisions with phonons and lattice vibrations, impurities and crystal imperfections), approximation of which is often tedious ([31], p.75 and Appendix 23). The linear dependence of the velocity on the electric field, expressed by (3.4), is a consequence of Newton's second law of motion (2.11). The RHS of (2.11), \mathcal{F}_{eff} , has to approximate a combined effect of thermo-mechanical and electromagnetic forces. For example, in the simplest case of a one-electron system evolving in a low electric field, we may use the following approximation

$$\mathcal{F}_{\text{eff}} \approx \mathcal{F}_0 = -q\mathbf{E}_{\text{eff}} - \frac{m_n^*\mathbf{v}}{\tau_p}. \quad (3.5)$$

Since this approximation is not appropriate in high electric fields, attempts have been made to improve it. One such attempt is expressed by the second equation of system (2.21), where two extra-terms were introduced to the model (3.5)

$$\mathcal{F}_2 = \mathcal{F}_0 - \mathbf{v} \cdot \nabla \mathbf{v} - \frac{1}{nm_n^*} \nabla(nT_n). \quad (3.6)$$

However, it is well known that in describing the dynamics of one-electron systems in high electric fields, the concept of energy-relaxation time, τ_ω , becomes inevitable (see [31], p.73). From the mathematical point of view we arrive at the evolutionary equation for the energy

density

$$\frac{\partial W}{\partial t} = \mathcal{F}_3 + \left(\frac{\partial W}{\partial t} \right)_{\text{col}}, \quad (3.7)$$

where the simplest approximations of the RHS of this equation can be written in the form

$$\mathcal{F}_3 \approx \mathcal{F}_3^0 = -q\mathbf{nv} \cdot \mathbf{E}_{\text{eff}}, \quad \left(\frac{\partial W}{\partial t} \right)_{\text{col}} \approx -\frac{W - 3nT_n/2}{\tau_\omega}. \quad (3.8)$$

An improved approximation of \mathcal{F}_3 is provided by the third equation in (2.21) (see [31, 3])

$$\mathcal{F}_3 \approx \mathcal{F}_3^0 - \nabla \cdot (\mathbf{v}W) - \nabla \cdot (\mathbf{v}nT_n) - \nabla \cdot \mathbf{q}. \quad (3.9)$$

The idea of replacing the equation for momentum by the energy balance equation is central to the development of quasi-hydrodynamic-type models. For the construction of such models we use the equations of conservation of mass and energy, but not momentum.

Quasi-hydrodynamic models have at least two limiting cases (for two-types carrier models):

- effective masses of different type carriers are of the same order (then it is reasonable to talk about common temperature of charge carriers which may be different from the lattice temperature);
- effective masses of different type carriers are sharply different, but times between collision for the same type of carriers are fairly small (then it is reasonable to talk about temperatures for carriers of different type which may not coincide between themselves as well as with the lattice temperature).

As we mention above, the similarity between hydrodynamic and quasi-hydrodynamic models manifests itself in the fact that the condition $\tau_p \ll \tau_\omega$ is the same for both models. In other words both models can be applied in the case of weakly *non-elastic* scattering. This happens if momentum scattering takes place on impurity atoms (under arbitrary energy scattering mechanism) or if energy scattering as well as (quasi-)momentum scattering is induced by interaction between charge carriers and acoustic/piezoelectric oscillation of the lattice. The scattering on optical phonons will be weakly non-elastic if the average energy of the charge carrier essentially exceeds the energy of optical phonons. In non-degenerate EHP the relaxation time is inversely proportional to the carrier concentration (up to slowly changing factor). Therefore, both hydrodynamic and quasi-hydrodynamic models may be applied only under high enough concentrations which exceed certain critical values, n_ω and n_p . Above these values scattering of energy and (quasi-)momentum (respectively) at the cost of between-electron (or between-hole) collisions becomes dominant. In order to evaluate these values in practice we scale the ratios τ_ω/τ and τ_p/τ to 1, leading to the approximate relation $n_\omega/n_p \approx \tau_p/\tau_\omega$.

Quasi-hydrodynamic models may be effectively applied even if not applicable in the rigorous sense of the words, i.e. when concentration of charge is not large enough compared to n_ω . This approach provides a reasonable approximation if we are interested only in physical quantities obtainable by averaging relatively smooth energy functions (the error of approximation is connected only with identification of the average energy with (3.3)). In

this case we chose the inertial system in such a way that the system of charge carriers is in equilibrium (see (3.3)). In any other inertial system the average energy would contain a term connected with the kinetic energy of motion of the charge carrier system as a whole (drift energy). In this inertial system the RHS of the energy balance equation takes a form similar to that from heat transfer theory, but with a relaxation factor. For example, for the electron system we have $\langle \partial_t \mathcal{E}_n \rangle = (T_n - T_l)/\tau_\omega$, where T_l denotes the temperature of the lattice. EHP and the crystal lattice play the role of subsystems that exchange energy. Whenever we expect $T_n > T_l$ the “hot carriers” terminology acquires the intuitive sense.

3.3 Mathematical model and physical parametrisation

Following [5, 2, 13, 19], in the space-time region $\bar{G}^R = \{(x, t) : 0 \leq x \leq L, 0 \leq t \leq T\}$ we consider the following quasi-hydrodynamic model for semiconductor device modelling

$$\left\{ \begin{array}{l} \partial_{xx}\varphi = q(n - p - N)/\epsilon\epsilon_0 \\ \partial_t n - \partial_x J_n/q = F, \\ \partial_t p + \partial_x J_p/q = F, \\ \partial_t \bar{\mathcal{E}}_n + \partial_x Q_n = -J_n \partial_x \varphi + P_n, \\ \partial_t \bar{\mathcal{E}}_p + \partial_x Q_p = -J_p \partial_x \varphi + P_p, \end{array} \right. \quad (3.10)$$

where expressions for densities of carrier currents, J_n , J_p and flux energies, Q_n and Q_p have the following form

$$J_n = -qn\mu_n\partial_x\varphi + q\partial_x(D_n n), \quad J_p = -qp\mu_p\partial_x\varphi - \partial_x(D_p p), \quad (3.11)$$

$$Q_n = \beta_n T_n n \mu_n \partial_x \varphi - \beta_n \partial_x [T_n D_n n]/q, \quad Q_p = -\beta_p T_p p \mu_p \partial_x \varphi - \beta_p \partial_x [T_p D_p p]/q, \quad (3.12)$$

$\bar{\mathcal{E}}_n = 3nT_n/2$, $\bar{\mathcal{E}}_p = 3pT_p/2$ are the average densities of the electron and hole systems respectively, and F is an approximation to the contribution of the generation-recombination (and, possibly, ionisation) processes.

The Peltier coefficients, β_n and β_p in (3.12) for typical Si and GaAs semiconductors take values between 2 and 3 and can be well approximated by the following formulae

$$\beta_n = 2.5 + \xi_n, \quad \beta_p = 2.5 + \xi_p, \quad (3.13)$$

where $\xi_n = d \ln \mu_n(T_n) / d \ln T_n$, $\xi_p = d \ln \mu_p(T_p) / d \ln T_p$ [2]. The first terms in the RHS of the energy balance equations represent the velocity of Joule heating/cooling. The second terms represent the velocity of energy losses induced by scattering on the lattice and are modelled by the formulae

$$P_n = n(T_l - T_n)/\tau_\omega^n, \quad P_p = p(T_l - T_p)/\tau_\omega^p, \quad (3.14)$$

where temperature is considered in energy units, and τ_ω^n , τ_ω^p are the average energy relaxation times for electrons and holes respectively. We can easily include the velocity of energy exchange between electrons and holes (and vice versa) as well as the energy of electron and

hole subsystems due to non-elastic collisions (recombination and ionization). Without loss of mathematical generality, we do not include these processes in our numerical procedures. As physical support for this simplification, we note that in many semiconductor structures like high-speed diodes and transistors, transition periods are small compared to characteristic times of energy exchange between carriers and the time of recombination/generation of carriers.

We also assume that carriers in different valleys have the same effective temperature. Then if n_i is the concentration of electrons in the i th valley we have $n = \sum_i n_i$. In this case properties of the carrier systems are characterized by the fact that average mobilities ($\mu_n = \mu_n(T_n)$, $\mu_p = \mu_p(T_p)$), diffusion coefficients ($D_n = D_n(T_n)$, $D_p = D_p(T_p)$), and times of energy relaxation ($\tau_\omega^n = \tau_\omega^n(T_n)$, $\tau_\omega^p = \tau_\omega^p(T_p)$) are dependent on carrier temperatures (see [2]). These dependencies can be approximated using a number of models known in the literature [6, 2, 22]. Typically, it is assumed that impulse scattering takes place mainly on acoustic phonons, so that the energy relaxation time can be approximated as a sum of two terms. The first one takes into account deformational acoustic phonons, and the second one is due to between-valley acoustic phonons. For example, for the electron system we have [2]

$$\frac{1}{\tau_\omega^n(T_n)} = \frac{1}{\tau_a(T_n/T_l)^{-1/2}} + \frac{\exp(-\hbar\omega_0/T_n)}{\tau_o(T_n/T_l)^{1/2}}, \quad (3.15)$$

where in the second term we use a "one"-phonon approximation ($\hbar\omega_0$ is the mean energy of an optical phonon), and τ_a , τ_o are temperature-dependent time-constants that characterise deformational and between-valley acoustic phonons. When the lattice temperature is close to 300°K, the contribution of the first term for Si devices becomes smaller. Another approximation often used in the literature (see [6] and references therein) has the following form

$$\tau_\omega^n = \frac{m_n \mu_n^0 T_0}{2q} \frac{T_0}{T_n} + \frac{3\mu_n^0}{2q(v_s^n)^2} \frac{T_n T_0}{T_n + T_0}, \quad (3.16)$$

where the velocity saturation, v_s^n , depends on the lattice temperature [26, 31] (typically it is of the order $10^6 - 10^7 \text{ cm} \times \text{s}^{-1}$). In order to avoid unnecessary technicalities we follow [22] by setting

$$\mu_n = \mu_n^0 (T_n/T_l)^q, \quad \mu_p = \mu_p^0 (T_p/T_l)^q, \quad (3.17)$$

$$\tau_\omega^n = \tau_{\omega,0}^n (T_l/T_n)^s, \quad \tau_\omega^p = \tau_{\omega,0}^p (T_l/T_p)^s, \quad (3.18)$$

where q and s are determined by the dominant relaxation mechanisms of the momentum and energy. Computational results, reported in the second part of this paper, were obtained for $q = s = 0$ with the low-field mobilities taken as $\mu_n^0 = 1300 \text{ cm}^2/\text{V}\cdot\text{s}$, $\mu_p^0 = 400 \text{ cm}^2/\text{V}\cdot\text{s}$ and the energy relaxation times set as $\tau_\omega^n = \tau_\omega^p = 0.4 \times 10^{-12} \text{ s}$ [5].

As for the dependencies of the diffusion coefficients on carrier temperatures, we admit that the numerical procedures developed in [20] can be easily generalized to the general type of dependence

$$D_n(T_n)/\mu_n(T_n) = \tilde{f}_1(T_n), \quad D_p(T_p)/\mu_p(T_p) = \tilde{f}_2(T_p). \quad (3.19)$$

To be specific, we developed the computational procedures under the assumption of the Einstein-type relationship, that is

$$D_n(T_n) \sim T_n \mu_n(T_n) \quad D_p(T_p) \sim T_p \mu_p(T_p) \quad (3.20)$$

with the constant of proportionality equal to $k_b/q = 8.61738 \times 10^{-5} \text{ eV/K}$.

Initial conditions for the model are

$$n(x, 0) = n_0(x), \quad p(x, 0) = p_0(x), \quad T_n(x, 0) = T_p(x, 0) = T_l, \quad 0 \leq x \leq L. \quad (3.21)$$

We assume that the functions $n_0(x)$ and $p_0(x)$ in the initial conditions are defined as equilibrium values of densities for electrons and holes, that is

$$p_0(x)n_0(x) = n_{ie}^2, \quad n_0(x) - p_0(x) - N = 0, \quad (3.22)$$

where n_{ie} is the effective intrinsic concentration of carriers.

In the general case, boundary conditions depend on the type of modelling structure. In this paper we require

- equality of carrier temperature and lattice temperature

$$T_n(0, t) = T_p(L, t) = T_l; \quad (3.23)$$

- conditions of quasi-neutrality and infinite velocity of recombination (thermodynamic equilibrium):

$$p - n + N = 0, \quad pn = n_{ie}^2, \quad x \in \partial G^R = \{0, L\}, \quad (3.24)$$

from where it is easy to get

$$n = \frac{N}{2} + \sqrt{\left(\frac{N}{2}\right)^2 + n_{ie}^2}, \quad p = -\frac{N}{2} + \sqrt{\left(\frac{N}{2}\right)^2 + n_{ie}^2}, \quad x \in \partial G^R = \{0, L\}. \quad (3.25)$$

For the potential, boundary conditions are standard [14]

$$\varphi(0, t) = 0, \quad \varphi(L, t) = U + \varphi_{cont} \quad (3.26)$$

where U is the applied voltage and φ_{cont} is the contact potential difference determined by the formula $\varphi_{cont} = \varphi_T \ln(n(t, L)/n_{ie})$ (obtained as a consequence of $n = n_{ie} \exp((\varphi - \varphi_n)/\varphi_T)$ by setting $\varphi_n = U$). In other words, we assume that the bias is applied at the right contact, while the left contact is grounded. In this case we require the conjugating conditions

$$\varphi(0, 0) = 0, \quad \varphi(L, 0) = U + \varphi_{cont} \quad (3.27)$$

to be satisfied. We note that if $\varphi_{cont} > 0$ then the case $U < 0$ corresponds to forward bias, and the case $U > 0$ corresponds to reverse bias.

For the effective intrinsic concentration, n_{ie} , in (3.22), (3.24), (3.25) and (3.26) one has to use an empirical model, for example (see [1, 31, 35] and references therein):

$$n_{ie} = n_i(T) \exp(q\Delta E_g/(2T)), \quad (3.28)$$

where T is the absolute temperature taken in energy units (multiply by the factor k_b/q), and ΔE_g is an *experimentally* measured parameter known as the effective bandgap narrowing. The energy gap itself, $E_g = E_c - E_v$, defined as the difference between the bottom of the conduction band, E_c , and the ceiling of the valence band, E_v , may change with different doping profiles and temperature. It is known, for example, that for highly doped material and high temperatures the bandgaps become smaller. Formula (3.28) is meant to take into account such changes. The intrinsic concentration, n_i , in formula (3.28) depends on the effective number of states in the conduction and valent zones (N_c and N_v respectively). It is common practice to use the following formula for its approximation

$$n_i = \sqrt{N_c N_v} \exp(-E_g/2k_b T), \quad (3.29)$$

where

$$N_c = 2 \left(\frac{2\pi m_{dn}^* k_b T}{h^2} \right)^{1.5} M_c, \quad N_v = 2 \left(\frac{2\pi m_{dp}^* k_b T}{h^2} \right)^{1.5}, \quad (3.30)$$

M_c is the number of equivalent minima in the conduction zone, and m_{dn}^* , m_{dp}^* are the density-of-state of effective masses of electrons and holes respectively (see [35], p.17). It is easy to see that when $\Delta E_g \rightarrow 0$, the effective intrinsic concentration can be well approximated by the intrinsic concentration. This assumption is used in our code where we set $n_{ie} \approx n_i = 1.45 \times 10^{10} \text{ cm}^{-3}$ (see [35], p.850; [8]).

The recombination model was chosen to account for recombination on defects induced by dopants (the Shockley-Read-Hall recombination) and between-zone Auger recombination:

$$F(n, p) = \frac{pn - n_{ie}^2}{\tau_n(p + n_{ie}) + \tau_p(n + n_{ie})} + (pn - n_{ie}^2)(c_n n + c_p p), \quad (3.31)$$

where the carrier life times and coefficients of Auger recombination are set as follows

$$\tau_n = 1.7 \times 10^{-5} \text{ s}, \quad \tau_p = 3.95 \times 10^{-4} \text{ s}, \quad c_n = 2.9 \times 10^{-31} \text{ cm}^6/\text{s}, \quad c_p = 1.2 \times 10^{-31} \text{ cm}^6/\text{s}.$$

For the problem where impact ionisation plays a significant role we have to add the velocity of ionization term

$$G_p - G_n = \alpha_p J_p - \alpha_n J_n, \quad (3.32)$$

where

$$J_n = -qnv_n, \quad J_p = qp v_p \quad (3.33)$$

and α_n , α_p are field-dependent carrier ionisation rates defined as the number of electron-hole pairs generated by an electron/hole per unit distance travelled [35]. Although in our numerical examples, presented in [20], only the Shockley-Read-Hall recombination was considered, our code is easily adaptable to account for other processes such as ionisation [8].

4 Normalization Procedure and Challenges in the Computational Treatment of Nonlocal Models

The magnitudes of dependent variables in the quasi-hydrodynamic model critically vary amongst each other, leading to a substantial computational cost of associated numerical procedures. In order to reduce the cost, effective normalisation procedures have to be implemented for the quasi-hydrodynamic model [1, 14].

Let us introduce the following dimensionless variables

$$\begin{aligned} x' &= x/L, \quad t' = t/t_*, \quad \varphi' = \varphi/\varphi_*, \quad n' = n/n_*, \quad p' = p/n_*, \quad T'_c = T_c/T_*, \\ N' &= N/n_*, \quad J'_c = J_c/J_*, \quad F' = F/F_*, \quad \mu'_n = \mu_n/\mu_*, \quad Q'_c = Q_c/Q_*, \\ \tau'_c &= \tau_c/t_*, \quad \alpha'_c = \alpha/\alpha_*, \quad c'_c = c_c/c_*, \end{aligned} \quad (4.1)$$

where quantities with the subindex “*” are critical values to be defined below, and variables with the subindex “c” (carriers) are equally applied to electrons or holes. We perform non-dimensionalisation of problem (3.10)–(3.12) in four steps.

- Step 1. From the Poisson equation written in the dimensionless variables

$$\frac{\epsilon\epsilon_0}{L^2}\varphi_* \frac{\partial^2 \varphi'}{\partial x'^2} = qn_*(n' - p' - N'), \quad (4.2)$$

we obtain

$$\frac{\epsilon\epsilon_0}{L^2}\varphi_* = qn_*. \quad (4.3)$$

- Step 2. The continuity equations (as an example we consider the continuity equation for electrons)

$$\frac{n_*}{t_*} \frac{\partial n'}{\partial t'} - \frac{1}{q} \frac{J_*}{L} \frac{\partial J'_n}{\partial x'} = F_* F' \quad (4.4)$$

leads to the relationships for two other normalised factors

$$\frac{n_* q L}{t_* J_*} = 1, \quad \frac{F_* q L}{J_*} = 1. \quad (4.5)$$

The formula for the current density

$$J'_n = -\frac{q n_* \mu_* \varphi_*}{L} n' \mu'_m \frac{\partial \varphi'}{\partial x'} + \frac{T_* \mu_* n_*}{L} \frac{\partial}{\partial x'} (T'_n \mu'_n n') \quad (4.6)$$

and relationships (4.3), (4.5) allow us to obtain

$$n_* = \frac{\epsilon\epsilon_0 T_*}{L^2 q^2}, \quad J_* = \frac{\epsilon\epsilon_0 T_*}{L q t_*}, \quad t_* = \frac{q L^2}{\mu_* T_*}, \quad \varphi_* = T_*/q. \quad (4.7)$$

- Step 3. Then, from the energy balance equation we get

$$\frac{3}{2} \frac{n_* T_* n' T'}{t_*} \frac{\partial t'}{\partial t'} + \frac{Q_*}{L} \frac{\partial Q'_n}{\partial x'} = -\frac{J_* \varphi_*}{L} J'_n \frac{\partial \varphi'}{\partial x'} + \frac{n_* T_* n'}{t_*} \frac{1 - T'_n}{\tau'_w} \quad (4.8)$$

that leads to only one relationship

$$Q_* = (n_* T_* L) / t_*. \quad (4.9)$$

(the other is satisfied automatically due to (4.7)).

The current density formula

$$Q_* Q'_n = \frac{T_* n_* \mu_* \varphi_*}{L} \beta_n T'_n n' \mu'_n \frac{\partial \varphi'}{\partial x'} - \frac{\beta_n}{q} \frac{T_* D_* n_*}{L} \frac{\partial}{\partial x'} [T'_n D'_n n'] \quad (4.10)$$

requires $D_* = \mu_* T_*$. The critical values of μ and T_c have been set to 1 (in mobility coefficient units) and T_l (in energy units) respectively.

• **Step 4.** Finally, we consider the recombination-generation-ionisation term in dimensionless variables

$$\begin{aligned} F_* F' = n_*^2 [p' n' - (n'_{ie})^2] & \left\{ \frac{1}{t_* n_* [\tau'_n (p' + n'_{ie}) + \tau'_p (n' + n'_{ie})]} + \right. \\ & \left. c_* n_* (c'_n n' + c_p p') \right\} + \frac{1}{q} J_* \alpha_* [\alpha'_p J'_p - \alpha'_n J'_n]. \end{aligned} \quad (4.11)$$

This leads to

$$\frac{n_* q L}{t_* J_*} = 1, \quad \frac{c_* n_*^3 q L}{J_*} = 1, \quad \frac{J_*}{q} \alpha_* \frac{q L}{J_*} = 1 \quad (4.12)$$

from which we easily obtain normalised factors for the Auger recombination coefficients and for the ionisation rates

$$c_* = 1/(n_*^2 t_*), \quad \alpha_* = 1/L. \quad (4.13)$$

The above deliberation leads to the following normalised factors that we use in this paper

$$\begin{aligned} T_* &= T_l k_b / q, \quad n_* = (\epsilon \epsilon_0 T_*) / (L^2 q^2), \quad \varphi_* = q n_* L^2 / (\epsilon \epsilon_0) = T_* / q, \quad D_* = \mu_* T_*, \quad t_* = q L^2 / (\mu_* T_*), \\ c_* &= 1/(n_*^2 t_*), \quad \alpha_* = 1/L, \quad J_* = \epsilon \epsilon_0 T_*/(L q t_*), \quad Q_* = n_* T_* L / t_*, \quad \mu_* = 1. \end{aligned} \quad (4.14)$$

The constants and normalised factors used in our computational experiments are summarised in [20] (see Appendix B).

Using (4.14), after simple transformations we get the following normalised system

$$\left\{ \begin{array}{l} \partial_{xx} \varphi = n - p - N, \\ \partial_t n - \partial_x J_n = F, \\ 3/2 \partial_t (n T_n) + \partial_x Q_n = -J_n \partial_x \varphi + P_n, \\ \partial_t p + \partial_x J_p = F, \\ 3/2 \partial_t (p T_p) + \partial_x Q_p = -J_p \partial_x \varphi + P_p, \end{array} \right. \quad (4.15)$$

where

$$J_n = -n \mu_n \partial_x \varphi + \partial_x (T_n \mu_n n), \quad J_p = -p \mu_p \partial_x \varphi - \partial_x (T_p \mu_p p), \quad (4.16)$$

$$Q_n = \beta_n T_n n \mu_n \partial_x \varphi - \beta_n \partial_x [T_n D_n n], \quad Q_p = -\beta_p T_p p \mu_p \partial_x \varphi - \beta_p \partial_x [T_p D_p p]. \quad (4.17)$$

All scaled quantities in (4.15)–(4.17) are denoted by the same symbols as their unscaled counterparts. The system (4.15)–(4.17) is considered in the normalised space-time region $\bar{G} = \{(x, t) : 0 \leq x \leq 1, 0 \leq t \leq T/t_*\}$ and is supplemented by the normalised initial and boundary conditions:

$$n(x, 0) = n_0(x)/n_*, \quad p(x, 0) = p_0(x)/n_*, \quad T_n(x, 0) = T_p(x, 0) = 1, \quad (4.18)$$

$$p - n + N = 0, \quad pn = n_{ie}^2, \quad T_n = T_p = 1, \quad x \in \partial G = \{0, 1\}, \quad (4.19)$$

$$\varphi(0, t) = 0, \quad \varphi(1, t) = (U + \varphi_{cont})/\varphi_*. \quad (4.20)$$

It is also assumed that the condition $J_n(x, 0) = J_p(x, 0) = 0$ and the normalised conjugating conditions $\varphi(0, 0) = 0, \varphi(1, 0) = (U + \varphi_{cont})/\varphi_*$ are satisfied.

In contrast to DDM, for which numerical methods have undergone extensive development, starting with Gummel's work (see, for example, references in [32, 14] and others), efficient numerical methods constructed for nonlocal-type models are at the beginning of their development. Modelling with nonlocal models incurs considerable mathematical difficulties, the overcoming of which is a challenging problem in applied mathematics [28]. For example, considering models of quasi-hydrodynamic type, we are dealing with fairly complex, strongly nonlinear problems of coupled field theory. Therefore, the development of effective numerical algorithms is required for the investigation of physical processes within the framework of such models. A large amount of publications is devoted to results of computations for specific devices [5, 6, 16, 17, 26, 30, 32, 33, 34, 37, 39]. However, the analysis of cost-effective algorithms is still lacking in the literature.

Mathematical modelling of non-local phenomena such as ballistic transfer and the velocity overshoot in semiconductor plasma has progressed since the early 1980's (although physical effects were described long before that time). The common feature of models for such phenomena is the accounting for macroscopic parameters for which balance laws are written with respect to the average energy. As a result, in contrast to DDM (where average energy is a local function of the field) such models are classified as nonlocal models.

A typical example of non-local mathematical models in semiconductor device theory is provided by the quasi-hydrodynamic model. A straightforward mathematical procedure for the solution of (4.15)–(4.20) is the Newton-Raphson method, applied to the discretised system of nonlinear equations [5, 37, 6]. This procedure is quite costly when applied to realistic semiconductor devices. Usually we have to apply special techniques in order to obtain convergence for the whole discretized system. For example, we may apply a sequentially-simultaneous algorithm which increases the convergence rate by conducting internal (adiabatic) iterations under fixed carrier temperature for 3 equations of DDM (prior to solving the coupled system of all 5 equations).

Alternative approaches to the solution of problem (4.15)–(4.20) are often based on different types of splitting algorithms [17]. The application of such approaches to strongly coupled problems encountered considerable difficulties in the context of semiconductor devices, especially for large electric fields. Another group of approaches uses different versions of the macro-particle method, where carrier collisions are modelled by Monte-Carlo-type

procedures [21]. The methods in this group are known to be typically costly and “noisy” in the computational sense. The principal problem with the macro-particle approach lies in the adequate modelling of pair collisions, a problem remains open to a large extent [15]. References to other recently developed computational procedures for semiconductor device models can be found in [10, 25].

In the next sections, using the quasi-hydrodynamic model as a typical example of non-local models, we demonstrate the main ideas of the construction of effective numerical schemes which can also be applied to hydrodynamic and classical drift-diffusion models.

5 Flux Approximations for Nonlocal Models of Quasi-hydrodynamic Type

One of the most important properties required by difference schemes in semiconductor device theory is monotonicity. Indeed, we have to guarantee that the solutions of the continuity and energy balance equations are nonnegative ($n, p, T_n, T_p \geq 0$) for any function of the potential φ . Let us consider these issues in some details.

First, we introduce a non-uniform grid in \bar{G}

$$\hat{\omega}_{h\tau} = \hat{\omega}_h \times \hat{\omega}_\tau, \quad (5.1)$$

where

$$\begin{aligned} \hat{\omega}_h &= \{x_{i+1} = x_i + h_i, \quad i = 0, \dots, N, x_0 = 0, x_{N+1} = 1, \sum_{i=0}^N h_i = 1\}, \\ \hat{\omega}_\tau &= \{t_j = t_{j-1} + \tau_j, \quad j = 1, \dots, K-1, \quad t_0 = 0, \quad t_K = T_f, \sum_{j=1}^{K-1} \tau_j = T_f\}. \end{aligned}$$

We will compute the values of φ , n , p , T_n , and T_p in the “whole” nodes (i.e. x_i , $i = 0, 1, \dots, N+1$), whereas the values of J_n , J_p , Q_n , Q_p , and $E = -\nabla\varphi$ will be computed in the data-driven (flux) nodes (i.e. $x_{i+1/2}$, $i=0, \dots, N$).

The approximation of fluxes is a long-standing problem in many applied problems for which solutions have steep gradients. In the context of semiconductor device modelling, we recall that even in the DDM, for which current density is defined as $J_n = -\mu_n n \nabla \varphi + \mu_n \nabla n$, the application of the approximation

$$J_{n,i+1/2} = -\mu_{i+1/2} n_{i+1/2} \frac{\varphi_{i+1} - \varphi_i}{h_{i+1}} + \mu_{i+1/2} \frac{n_{i+1} - n_i}{h_{i+1}} \quad (5.2)$$

is impeded because of the very restrictive condition on the space step discretisation which follows from the *maximum principle*. In order to obtain this condition we use the theorem on monotonicity of three-point difference operators (the Karetkina lemma, see [13, 19] and references therein), namely

Theorem 5.1 *If for the three-point-stencil difference operator*

$$(\Lambda^* n)_i \equiv A_i n_{i-1} - C_i n_i + B_i n_{i+1}, \quad 0 < i < N+1$$

the following conditions

$$(\Lambda^* n)_i \leq 0, \quad 0 < i < N + 1; \quad A_i > 0, \quad 1 \leq i \leq N + 1; \quad B_i > 0, \quad 0 \leq i \leq N;$$

$$C_i = A_{i+1} + B_{i-1} + G_i, \quad \text{where } G_i \geq 0, \quad 0 < i < N + 1; \quad n_0 \geq 0, n_{N+1} \geq 0.$$

are satisfied, then the monotonicity condition

$$n_i \geq 0, \quad \forall i = 0, \dots, N + 1$$

is guaranteed.

If we use the approximation for the current density in formula (5.2), the conditions of this theorem will be satisfied for

$$h < 2/E^*, \quad \text{where } E^* = \max_{i=1, \dots, N} |E_{i+1/2}|. \quad (5.3)$$

In the case of the QHDM (4.15)–(4.20), the monotonicity condition for the current-density approximation analogous to (5.2), that is

$$J_{n,i+1/2} = \frac{D(T_{i+1})n_{i+1} - D(T_i)n_i}{h_{i+1}} - \frac{n_{i+1}\mu(T_{i+1}) + n_i\mu(T_i)}{2} \frac{\varphi_{i+1} - \varphi_i}{h_{i+1}}, \quad (5.4)$$

coincides with (5.3).

The first monotone difference scheme in a semiconductor-device-modelling context was first reported by D. L. Scharfetter and H.K. Gummel (see references, for example, in [32, 24, 13] and others). These types of schemes, known to the mathematical community as *exponential*, constitute an important tool in the integration of stiff ordinary differential equations [4, 23]. For ODEs they are typically unconditionally stable and, what is very important, positivity of the solution is guaranteed if the solution of the differential problem is expected to be positive. They can also be constructed without major difficulties in the case of partial differential equations when the spatial differential operator can be reduced to the self-conjugate form [27]. For the continuity equations of the classical DDM, the idea of such a reduction has been intensively investigated. If the Boltzmann statistics is assumed, then the exponential change of variables

$$n = n_{ie} \exp(\varphi) \Phi_n, \quad p = n_{ie} \exp(-\varphi) \Phi_p, \quad (5.5)$$

leads to an essential simplification of the current densities which become linearly dependent on quasi-potentials Φ_n , Φ_p . From the mathematical point of view, this is a very attractive feature of the model that, in turn, leads to a number of effective algorithms [32]. Such algorithms were also constructed in the case of Fermi statistics [24]. However, it should be noted that the practical value of all such schemes is essentially dependent on the quality of approximation of the strongly nonlinear RHS of the continuity equation.

In the 70's and early 80's, works in the application of non-local models to semiconductor device simulation were conducted predominantly with Monte-Carlo type procedures, and in those papers where difference methods were used, questions of scheme quality have not been adequately explored. The turning point in the development of difference methods for the QHDM was the work of Tang [36], where the Scharfetter-Gummel approximation

was generalized to the case of a particular type of non-local model. The current density approximation was considered in the following form (we omit the indexes n and p for the simplicity):

$$J_{i+1/2} = \frac{1}{h_{i+1}} \left[D(T_{i+1}) n_{i+1} f_1 \left(\frac{\varphi_{i+1} - \varphi_i}{T_{i+1/2}} \right) - D(T_i) n_i f \left(\frac{\varphi_{i+1} - \varphi_i}{T_{i+1/2}} \right) \right], \quad (5.6)$$

where

$$f(x) = x \exp(x) / (\exp(x) - 1), \quad f_1(x) = x / (\exp(x) - 1) \quad (5.7)$$

are Bernoulli functions (bearing a computer code in mind, we recall that $f_1(x) = f(-x)$), and the quantities $T_{i+1/2}$ may be approximated by any value of temperature on the integration interval $[x_i, x_{i+1}]$, for example, T_i , $(T_i + T_{i+1})/2$, T_{i+1} . As we expect, such approximation turns into the Scharfetter-Gummel approximation when $T_{i+1} \rightarrow T_i$.

Unfortunately in the general case, even the Scharfetter-Gummel-type approximation cannot guarantee absolute stability neither for the DDM nor for nonlocal models. We can only claim the conservation of solution positiveness for certain “model” problems for the continuity equation, such as

$$(\mu_n \left(\frac{\partial n}{\partial x} - n \frac{\partial \varphi}{\partial x} \right))_x = 0. \quad (5.8)$$

We can also claim the conservation of positiveness property for a specific computational experiment. However, if we consider the problem with even the “simplest” recombination model (say, the Shockley-Read-Hall recombination), then the stability of the method depends on the method of linearization of the recombination term. If, for example, the recombination term is taken from the previous iteration, then the first condition of Theorem 5.1 ($(\Lambda^* n)_i \leq 0$) does not hold any longer. Indeed, for structures with a strong recombination and for small time lives of carriers, numerical experiments show the possibility of negative concentrations. Of course, there exist linearisation procedures for the recombination term (such as the Seidman-Choo procedure) that satisfy all conditions of the monotonicity theorem. However, if other processes such as ionization are dominant and therefore have to be taken into account, then the RHS linearization is typically a heuristic procedure, aimed at the achievement of numerical stability and algorithm convergence.

Challenging mathematical and computational problems also arise in the approximation of energy balance equations for nonlocal models. Since the approximation for the energy flux, analogous to the Scharfetter-Gummel approximation of the current density, is known [36, 38], the main challenge is to transform the energy balance equations to forms that are most suitable for an efficient computational implementation.

6 Transformation of the Energy Balance Equations to Forms Amenable to Computational Efficiency

In contrast to the continuity equation, the energy balance equation *cannot be readily reduced* to a “divergent” or “conservation” form [2, 27, 29]. In the semiconductor-device modelling context, the main problem with the energy balance equation lies with the presence

of the product between the current density and the electric field strength ($J_n \times E$ or $J_p \times E$), that has a “non-divergent” structure. Hence, one cannot immediately apply the general theory developed for the construction of monotone difference schemes [27, 29, 23]. However, since the product between the current density and the electric field strength provides the key to the nonlocal coupling between the effective carrier temperature and the electric field, the problem of its efficient approximation has to be dealt with. In this section we show how the energy balance equations can be reduced to an “almost-divergent” form which is used for their effective discretisation [20].

Following [36, 2, 13], let us transform the energy fluxes to forms where all derivatives of \mathcal{E}_n and \mathcal{E}_p are “covered” by the symbol of divergence. We start from the energy balance equation for the system of electrons

$$\partial_t \tilde{\mathcal{E}}_n = -\partial_x Q_n - J_n \partial_x \varphi + P_n. \quad (6.1)$$

The right hand side of (6.1) is transformed as follows

$$\begin{aligned} -\frac{\partial Q_n}{\partial x} - J_n \frac{\partial \varphi}{\partial x} + P_n &= \frac{\partial}{\partial x} \left[\beta_n \frac{\partial(T_n D_n n)}{\partial x} - \beta_n n T_n \mu_n \frac{\partial \varphi}{\partial x} \right] - \\ \frac{\partial}{\partial x} [\mu_n n T_n] \frac{\partial \varphi}{\partial x} + n \mu_n \left(\frac{\partial \varphi}{\partial x} \right)^2 + P_n &= \beta_n \frac{\partial}{\partial x} \left[\frac{\partial(T_n D_n n)}{\partial x} \right] - \\ \frac{\partial}{\partial x} [\mu_n n T_n] \frac{\partial \varphi}{\partial x} - \mu_n n T_n \frac{\partial^2 \varphi}{\partial x^2} + n T_n \mu_n \frac{\partial^2 \varphi}{\partial x^2} - \\ \beta_n \frac{\partial}{\partial x} [\mu_n n T_n] \frac{\partial \varphi}{\partial x} - \beta_n \mu_n n T_n \frac{\partial^2 \varphi}{\partial x^2} + n \mu_n \left(\frac{\partial \varphi}{\partial x} \right)^2 + P_n &= \\ \frac{\partial}{\partial x} \left[\beta_n \frac{\partial(D_n n T_n)}{\partial x} - (1 + \beta_n) \mu_n n T_n \frac{\partial \varphi}{\partial x} \right] + n T_n \left[\mu_n \frac{\partial^2 \varphi}{\partial x^2} + \right. \\ \left. \frac{\mu_n}{T_n} \left(\frac{\partial \varphi}{\partial x} \right)^2 + \frac{1 - T_n}{\tau_{\omega^n} T_n} \right] &= \frac{\partial Q_n^*}{\partial x} + S_n(T_n, \varphi) \mathcal{E}_n, \end{aligned} \quad (6.2)$$

where

$$\mathcal{E}_n = n T_n, \quad S_n = \mu_n(T_n) \partial_{xx} \varphi + \mu_n(T_n) (\partial_x \varphi)^2 + (1 - T_n) / (\tau_{\omega^n}^n(T_n) T_n), \quad (6.3)$$

$$Q_n^* = \beta_n \partial_x (D_n(T_n) \mathcal{E}_n) - (1 + \beta_n) \mu_n(T_n) \mathcal{E}_n \partial_x \varphi. \quad (6.4)$$

Therefore, equation (6.1) can be written in the form

$$3 \partial_t \mathcal{E}_n / 2 = \partial_x Q_n^* + S_n(T_n, \varphi) \mathcal{E}_n. \quad (6.5)$$

In a similar manner we transform the balance energy equation for the system of holes

$$\partial_t \tilde{\mathcal{E}}_p = -\partial_x Q_p - J_p \partial_x \varphi + P_p. \quad (6.6)$$

Since

$$\begin{aligned}
 -\frac{\partial}{\partial x} \left[-\beta_p T_p p \mu_p \frac{\partial \varphi}{\partial x} - \beta_p \frac{\partial}{\partial x} [T_p D_p p] \right] - \left[-p \mu_p \frac{\partial \varphi}{\partial x} - \frac{\partial}{\partial x} (D_p p) \right] \frac{\partial \varphi}{\partial x} + P_p = \\
 \beta_p \frac{\partial}{\partial x} \left[\frac{\partial (T_p D_p p)}{\partial x} \right] + \frac{\partial}{\partial x} (\mu_p T_p p) \frac{\partial \varphi}{\partial x} - \mu_p p T_p \frac{\partial^2 \varphi}{\partial x^2} + p T_p \mu_p \frac{\partial^2 \varphi}{\partial x^2} + \\
 \beta_p \frac{\partial}{\partial x} [\mu_p T_p p] \frac{\partial \varphi}{\partial x} + \beta_p \mu_p p T_p \frac{\partial^2 \varphi}{\partial x^2} + p \mu_p \left(\frac{\partial \varphi}{\partial x} \right)^2 + P_p = \\
 \frac{\partial}{\partial x} \left\{ \beta_p \frac{\partial (D_p T_p p)}{\partial x} + \mu_p p T_p \frac{\partial \varphi}{\partial x} + \beta_p T_p p \mu_p \frac{\partial \varphi}{\partial x} \right\} + \mu_p p T_p \frac{\partial^2 \varphi}{\partial x^2} - \\
 p \mu_p \left(\frac{\partial \varphi}{\partial x} \right)^2 + P_p = \frac{\partial}{\partial x} \left\{ \beta_p \frac{\partial (D_p T_p p)}{\partial x} + (1 + \beta_p) \mu_p p T_p \frac{\partial \varphi}{\partial x} \right\} - \\
 \mu_p p T_p \frac{\partial^2 \varphi}{\partial x^2} + p \mu_p \left(\frac{\partial \varphi}{\partial x} \right)^2 + P_p = \frac{\partial Q_p^*}{\partial x} + S_p(T_p, \varphi) \mathcal{E}_p,
 \end{aligned} \tag{6.7}$$

we rewrite equation (6.6) in the following form

$$3 \partial_t \mathcal{E}_p / 2 = \partial_x Q_p^* + S_p(T_p, \varphi) \mathcal{E}_n, \tag{6.8}$$

where

$$\mathcal{E}_p = p T_p, \quad S_p = -\mu_p(T_p) \partial_{xx} \varphi + \mu_p(T_p) (\partial_x \varphi)^2 + (1 - T_p) / (\tau_\omega^p(T_p) T_p), \tag{6.9}$$

$$Q_p^* = \beta_p \partial_x (D_p(T_p) \mathcal{E}_p) + (1 + \beta_p) \mu_p(T_p) \mathcal{E}_p \partial_x \varphi. \tag{6.10}$$

The representations (6.5) and (6.8) allow us to construct *monotone exponential* difference schemes [27, 29] for nonlocal models applied to semiconductor device simulation.

7 Conclusions and Future Directions

In this paper we considered a hierarchy of semiconductor device models using the relaxation time concept. In order to describe nonlocal, non-equilibrium processes in electron-hole semiconductor plasma, we focused on the quasi-hydrodynamic model which provides a reasonable compromise between kinetic, hydrodynamic and drift-diffusion models. The normalisation procedure, and issues of the approximation of fluxes for this model were discussed in detail. The energy balance equations for the electron and hole systems were reduced to convenient forms for computational implementation.

As a development of the work presented here, in [20] we construct effective exponential difference schemes based on the representations obtained in this paper. The problems of computational stability of the algorithmic realisations of the proposed schemes as well as their application to the modelling of transport phenomena in semiconductor devices are also discussed in [20].

Acknowledgements.

Authors were supported by grant USQ-PTRP 17989 and by Australian Research Council Small Grant 17906. We thank Dr David Smith for his assistance at the final stage of preparation of this paper.

References

- [1] Apanovich, Y., Lyumkis, E., Polksy, B.S. et al, Steady-State and Transient Analysis of Submicron Devices Using Energy Balance and Simplified Hydrodynamic Models, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 13, No. 6, 702–711, 1994.
- [2] Birukova, L.J. et al, Simulation algorithms for computing processes in electron plasma of submicron semiconductor devices, *Math. Modelling*, Vol. 1, No. 5, 11–22, 1989.
- [3] Blotekjaer, K., Transport equations for electrons in two-valley semiconductors, *IEEE Transactions on Electronic Devices*, Vol. ED-17, No. 1, 38–47, 1970.
- [4] Bui, T.D., A.K. Oppenheim, and D.T. Pratt, Recent advances in methods for numerical solution of ODE initial value problems, *J. Comp. Math.*, Vol. 11, 283–296, 1984.
- [5] Cook, R.K., Numerical Simulation of Hot-Carrier Transport in Silicon Bipolar Transistors, *IEEE Trans. Electron. Devices*, Vol. ED-30, No. 9, 1103–1110, 1983.
- [6] C.L. Gardner, J.W. Jerome and D.J. Rose, Numerical Methods for the Hydrodynamic Device Model: Subsonic Flow, *IEEE Transactions on Computer-Aided Design*, Vol. 8, No. 5, 501–507, 1989.
- [7] Hansch, W. and Miura-Mateush, A new current relation for hot electron transport, in *NEMACODE IV: Proceedings of the Fourth International Conference on the Numerical Analysis of Semiconductor Devices and Integrated Circuits: Trinity College, Dublin, Ireland*, Edited by J.J. H. Miller, Boole Press, 1985, 311–314.
- [8] He, H., Melnik, R.V.N., Numerical simulation of semiconductor devices with the quasi-hydrodynamic model: program manual, *Technical Report SC-MC-98, University of Southern Queensland*, 1998, 1–42.
- [9] Ivrii, V. Ja. and Sigal, I.M., Asymptotics of the ground state energies of large Coulomb systems, *Annals of Mathematics*, , Vol. 138, No. 2, 243–335, 1993.
- [10] Jerome, J.W., Algorithmic aspects of the hydrodynamic and drift-diffusion models, *Mathematical Modelling and Simulation of Electrical Circuits and Semiconductor Devices*, Eds. R.E. Bank, R. Burlirsch, K. Merten, Int. Series of Numerical Mathematics, Vol. 93, Birkhauser-Verlag, 217–236, 1990.
- [11] Lieb, E.H. and Yau, H.-T., The stability and instability of relativistic matter, *Commun. Math. Phys.*, Vol. 118, 177–213, 1988.
- [12] Lieb, E.H., Siedentop, H. and Solovej, J.P., Stability and instability of relativistic electrons in classical electromagnetic fields, *Journal of Statistical Physics*, Vol. 89, No. 1/2, 37–57, 1997.

- [13] Lyumkis, E.D. et al Transient Semiconductor Device Simulation including energy balance equation, *COMPEL - The International Journal for Computation and Mathematics in Electrical and Electronic Engineering*, Vol. 11, No. 2, 311–325, 1992.
- [14] Markowich, P.A., Ringhofer, C.A. and Schmeiser, C., *Semiconductor equations*, Springer-Verlag, 1991.
- [15] Markowich, P.A., Diffusion Approximation of Nonlinear Electron Phonon Collision Mechanisms, *Mathematical Modelling and Numerical Analysis (M²AN)*, Vol. 29, No. 7, 857–869, 1995.
- [16] Marshak, A.H. and van Vliet K.M., Electrical current in solids with position-dependent band structure, *Solid-State Electronics*, Vol. 21, No. 2, 417–427, 1978.
- [17] Mayorov, S.A., Melnikov, A.M. and Rudenko, A.A., Modelling semiconductor microstructures in strong electric fields taking into account collision ionisation, *Math. Modelling*, Vol. 1, No. 5, 23–32, 1989.
- [18] Melnik, R.V.N., Correction for nonstationarity and internal nonlinearity in the analysis of integrated-circuits thermal parameters, *Radioelectronics & Communications Systems*, Vol. 34, No. 6, 84–86, 1991.
- [19] Melnik, R.V.N., Semi-Implicit Finite-Difference Schemes with Flow Correction for Quasi-Hydrodynamic Models of Semiconductor Devices, *Engineering Simulation*, Gordon and Breach, Vol. 12, 856–865, 1995.
- [20] Melnik, R.V.N., He, H., Modelling nonlocal processes in semiconductor devices with exponential difference schemes. Part 2: Numerical Methods and Computational Experiments, *International Journal for Numerical Methods in Engineering*, submitted.
- [21] Moglestue, C., *Monte Carlo simulation of semiconductor devices*, Chapman & Hall, 1993.
- [22] Nikolaeva, V. A., Ryzhii, V. I. and B.N. Cheverushkin, A numerical method for the simulation of two-dimensional semiconductor structures in the quasi-hydrodynamic approximation, *Sov. Phys. Dokl*, Vol. 33(2), 110–112, 1988.
- [23] Oran, E.S. and J.P. Boris, *Numerical Simulation of Reactive Flow*, Elsevier, 1987.
- [24] Polksy, B.S. and Rimshans, J.S., Two-dimensional numerical simulation of bipolar semiconductor devices taking into account heavy doping effects and Fermi statistics, *Solid-State Electronics*, Vol. 26, No. 4, 275–279, 1983.
- [25] Ringhofer, C., Computational methods for semiclassical and quantum transport in semiconductor devices, *Acta Numerica*, Cambridge University Press, 1997, 485–521.
- [26] Rudan, M., Odeh, F. and J. White, Numerical solution of the hydrodynamic model for a one-dimensional semiconductor device, *COMPEL - The International Journal for Computation and Mathematics in Electrical and Electronic Engineering*, Vol. 6, No. 3, 151–170, 1987.
- [27] A.A. Samarskii, *Theorie der Differenzenverfahren*, (Series: Mathematik und ihre Anwendungen in Physik und Technik), Aufl. Leipzig: Academische Verlagesellschaft Geest & Portig, 1984.

- [28] Samarskii, A. A. and Chetverushkin, B.N., Microelectronics as a New Object of Investigation in Applied Mathematics, *Computational Mathematics and Cybernetics: Vestnik of the Moscow University*, No. 3, 9–20, 1986.
- [29] A.A. Samarskii and E.S. Nikolaev, *Numerical Methods for Grid Equations*, Basel, Boston: Birkhauser Verlag, 1989.
- [30] Scrobohaci, P. G. and T.-W. Tang, Modeling of the Hot Electron Subpopulation and its Application to Impact Ionization in Submicron Silicon Devices, *IEEE Transactions on Electron Devices*, Vol. 41, No. 7, 1197–1212, 1994.
- [31] Shur, M., *Physics of Semiconductor Devices*, Prentice Hall, 1990.
- [32] Slotboom, J.W. Computer aided two-dimensional analysis of bipolar transistor, *IEEE Trans. Electron. Devices*, Vol. ED-20, No. 8, 669–679, 1973.
- [33] Snowden, C.M. and D. Loret, Two-dimensional hot-electron models for short-gate-length GaAs MESFET's, *IEEE Trans. Electron. Devices*, Vol. ED-34, No. 2, 212–223, 1987.
- [34] Stratton, R., Diffusion of hot and cold electrons in semiconductor barriers, *Phys. Rev. B*, Vol. 126, No. 6, 2002–2014, 1962.
- [35] Sze, S. M., *Physics of Semiconductor Devices*, John Wiley & Sons, 1981.
- [36] Tang, T.-W., Extension of the Scharfetter-Gummel algorithm to the energy balance equation, *IEEE Transactions on Electronic Devices*, Vol. ED-31, No. 12, 1912–1914, 1984.
- [37] Tang, T.-W., Ou, X.X., Navon, D.X., Prediction of the velocity overshoot by a nonlocal hot-carrier transport model, in *NEMACODE IV: Proceedings of the Fourth International Conference on the Numerical Analysis of Semiconductor Devices and Integrated Circuits: Trinity College, Dublin, Ireland*, Edited by J.J. H. Miller, Boole Press, 1985, 519–524.
- [38] Tang, T.-W. and Leong, M.-K., Discretization of Flux Densities in Device Simulations Using Optimum Artificial Diffusivity, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 14, No. 11, 1309–1315, 1995.
- [39] Zhang, Y., M.El. Nokali, A Hydrodynamic transport model and its applications in semiconductor device simulation, *Solid-State Electronics*, Vol. 36, No. 12, 1689–1696, 1993.

USQ



TOOWOOMBA

**OPTIMAL-BY-ACCURACY AND
OPTIMAL-BY-ORDER CUBATURE
FORMULAE IN CLASS $C_{1,L,N}^2$.**

K N Mehlak

Department of Computer Science
Flinders University, Adelaide, SA

R V N Melnik

Department of Mathematics & Computing, USQ

THE UNIVERSITY OF
S O U T H E R N
Q U E E N S L A N D

Working Papers

FACULTY
OF
SCIENCES

**OPTIMAL-BY-ACCURACY AND
OPTIMAL-BY-ORDER CUBATURE
FORMULAE IN CLASS $C_{1,L,N}^2$.**

K N Melnik

Department of Computer Science
Flinders University, Adelaide, SA

R V N Melnik

Department of Mathematics & Computing, USQ
Faculty of Sciences Working Paper Series
SC-MC-9821
20 August 1998

OPTIMAL-BY-ACCURACY AND OPTIMAL-BY-ORDER CUBATURE FORMULAE IN CLASS $C_{1,L,N}^2$.

K. N. Melnik *

Department of Computer Science,
Flinders University, Adelaide, SA 5001, Australia

R. V. N. Melnik †

Department of Mathematics and Computing,
University of Southern Queensland, QLD 4350, Australia

Abstract

In this article we use the method of limit functions for the construction of optimal-by-accuracy and optimal-by-order (with constant not exceeding two) cubature formulae for the integration of fast oscillatory functions given by their values at a finite number of fixed nodes. Main results of the paper are obtained for the interpolational class $C_{1,L,N}^2$. The close connection between numerical integration procedures and the optimal recovery of functions from interpolational classes provides a rigorous mathematical basis for the constructive solution of the considered problems.

Key words: fast oscillatory functions, numerical integration, optimal recovery, interpolational classes.

1 Introduction.

In this paper we consider the problem of construction of optimal-by-accuracy and optimal-by-order cubature formulae for computing integrals

$$I^2(f) = \int_0^1 \int_0^1 f(x_1, x_2) \varphi_1(x_1) \varphi_2(x_2) dx_1 dx_2, \quad (1.1)$$

where $\varphi_1(x_1), \varphi_2(x_2)$ are known integrable functions, and $f(x_1, x_2)$ belongs to a certain given class F_N .

*Currently with the Electronic Data Systems, 60 Waymouth Street, Adelaide, 5000, Australia

†Corresponding author, E-mail: melnik@usq.edu.au

Integrals of the form (1.1) often arise in the context of the Fourier or Fourier-Bessel integral transforms [27] and occur in many applications including signal and image processing, radioastronomy, crystallography and modelling automatic regulation systems. In these applications, functions $\varphi_1(x_1)$ and/or $\varphi_2(x_2)$ in (1.1) often exhibit highly oscillatory behaviour. This leads to the essential mathematical difficulties in computing integrals (1.1) (see [1, 5, 8, 4, 10, 11, 12, 14] and references therein). Such difficulties are encountered even in a relatively simple one-dimensional case where we have to integrate the product $f(x)\exp(-i\omega x)$ on an interval (a, b) , where $\omega(b - a) \gg 1$ [7, 16]. On the interval (a, b) the functions $\text{Re}(f(x)\exp(-i\omega x))$ and $\text{Im}(f(x)\exp(-i\omega x))$ have approximately $\omega(b - a)/\pi$ zeros. Therefore, even if $f(x)$ is a smooth function, in order to achieve an adequate level of approximation we have to choose a polynomial of degree $n \gg \omega(b - a)/\pi$. The use of such a high degree polynomial may eventually lead to instability of computations [11]. The situation becomes more complicated in the two-dimensional case [12, 19, 6, 17]. Moreover, in the majority of practical situations only approximate information about the integrand is given as a result of measurements or physical experiments. Therefore, interpolational classes become the appropriate functional classes for the study of problems in numerical integration of fast oscillatory functions.

In this paper we construct optimal-by-accuracy and optimal-by-order cubature formulae for computing integrals (1.1) in interpolational class $C_{1,L,N}^2$. The class $C_{1,L,N}^2$ is the class of functions that are defined in the domain π_2 , $\pi_2 = \{X = (x_1, x_2) : 0 \leq x_i \leq 1, i = 1, 2\}$, satisfy the Lipschitz condition with constant L ,

$$|f(X) - f(Y)| \leq L\|X - Y\| = L \max_{i=1,2} |x_i - y_i| \quad (1.2)$$

and take in fixed nodes X_1, \dots, X_N of arbitrary grid corresponding fixed values $f(X_1) = f_1, \dots, f(X_N) = f_N$. It is assumed that this functional class is non-empty.

In order to obtain optimal-by-accuracy and optimal-by-order solutions of problem (1.1), we use the method of limit functions [24, 21, 15, 16]. Namely, we define the upper, $I^+(F_N)$, and the lower, $I^-(F_N)$, limits of the set of possible values of the integral (1.1) on functions from class F_N as

$$I^+(F_N) = \sup_{f \in F_N} I^2(f), \quad I^-(F_N) = \inf_{f \in F_N} I^2(f), \quad (1.3)$$

and then determine the value

$$I^*(F_N) = \frac{I^+(F_N) + I^-(F_N)}{2}, \quad (1.4)$$

which is taken as the optimal-by-accuracy value of the integral $I^2(f)$. In this case $I^*(F_N)$ is the Chebyshev center of uncertainty domain of values $I^2(f)$ on class F_N [16]. The Chebyshev radius coincides with $\delta(F_N)$ and is defined as follows

$$\delta(F_N) = \frac{1}{2} (I^+(F_N) - I^-(F_N)). \quad (1.5)$$

When $\varphi_1(x_1) = \varphi_2(x_2) = 1$, the problem becomes one of computing optimal-by-accuracy value $I_1^*(F_N)$ for integrals

$$I_1^2(f) = \int \int_{\pi_2} f(X) dX \quad (1.6)$$

with $f \in F_N$ and $X = (x_1, x_2)$. We notice that the problem of optimal-by-accuracy integration on class F_N is closely connected to the problem of optimal-by-accuracy recovery of $f(X) \in F_N$ at point $X = (x_1, x_2) \in \pi_2$ (see, for example, [3, 21] and references therein). In particular, an optimal-by-accuracy recovery $f^*(X)$ at point $X \in \pi_2$ for functions from class $C_{1,L,N}^2$ is of special interest. In order to explore the connection between the above two problems we recall the definition of majorant (minorants) of functional classes.

Definition 1.1 Let F_N be a class of functions defined in a domain D . Then a function $A_{F_N}^+(X)$ ($A_{F_N}^-(X)$) is called a majorant (minorant) of the class F_N , if the conditions

- (a) $A_{F_N}^+(X) \geq f(X)$ ($A_{F_N}^-(X) \leq f(X)$) for all $f \in F_N$, $X = (x_1, x_2) \in D$ and
- (b) $A_{F_N}^+(X) \in F_N$ ($A_{F_N}^- \in F_N$)

are satisfied.

The value of

$$f^*(X) = \frac{1}{2} (A_{F_N}^+(X) + A_{F_N}^-(X)) \quad (1.7)$$

(with $A_{F_N}^+(X)$, $A_{F_N}^-(X)$ majorant and minorant of class F_N respectively) is taken as the optimal-by-accuracy recovery of $f(X)$ at $X \in \pi_2$. Further in this paper we assume that $F_N = C_{2,L_1,L_2,N}^2$ or $F_N = C_{2,L,L,N}^2$. The error $\bar{\delta}(F_N, X)$ of the recovery of function $f(X) \in F_N$ at point X has the form

$$\bar{\delta}(F_N, X) = \frac{A_{F_N}^+(X) - A_{F_N}^-(X)}{2}. \quad (1.8)$$

Then, the optimal-by-accuracy cubature formulae for computing (1.6) is [24, 21, 17]

$$I_1^*(F_N) = \int \int_{\pi_2} f^*(X) dX \quad (1.9)$$

with the Chebyshev radius, $\bar{\delta}(F_N)$, of the domain of undefinability of integral (1.6) in the form

$$\bar{\delta}(F_N) = \int \int_{\pi_2} \bar{\delta}(F_N, X) dX. \quad (1.10)$$

For a constructive solution of problems (1.7), (1.8) and (1.9), (1.10), as well as for the construction of efficient cubature formulae for computing integral (1.1) we have to consider the properties of majorants and minorants of functional classes that are investigated.

We organise this paper as follows.

- In Sections 2 and 4 we consider the problem of optimal-by-accuracy recovery of a function.
- In Section 3 we deal with problems connected with the choice of the grid in the interpolational class $C_{1,L,N}^2$.
- In Section 5 we construct optimal-by-accuracy cubature formula for computing integral $I_1^2(f)$.
- Finally, in Section 6 we construct optimal-by-order cubature formulae.

2 On Majorant and Minorant of the Functional Class $C_{1,L,N}^2$

The important special cases of integral $I^2(f)$ are integrals of the form

$$I_2^2(f) = \int_0^1 \int_0^1 f(x_1, x_2) \sin \omega_1 x_1 \sin \omega_2 x_2 dx_1 dx_2, \quad (2.1)$$

$$I_3^2(f) = \int_0^1 \int_0^1 f(x_1, x_2) \cos \omega_1 x_1 \cos \omega_2 x_2 dx_1 dx_2, \quad (2.2)$$

where $\omega_i, i = 1, 2$ are certain real numbers, $|\omega_i| \geq 2\pi, i = 1, 2$. We aim to construct optimally-order cubature formulae for computing integrals (2.1), (2.2) and obtain error estimates for these formulae.

For the solution of the above problems we need to know the form and certain properties of the majorant $A_{C_{1,L,N}^2}^+(X)$ and the minorant $A_{C_{1,L,N}^2}^-(X)$ of the class that is being investigated.

It can be shown [13, 2] that

$$A_{C_{1,L,N}^2}^+(X) = \sup_{f \in C_{1,L,N}^2} f(X) = \min_{\mu=1,\dots,N} (f_\mu + L\|X - X_\mu\|_1), \quad (2.3)$$

$$A_{C_{1,L,N}^2}^-(X) = \inf_{f \in C_{1,L,N}^2} f(X) = \max_{\mu=1,\dots,N} (f_\mu - L\|X - X_\mu\|_1), \quad (2.4)$$

where $X = (x_1, x_2)$, $D = \pi_2$ (see Definition 1.1), $\|X\|_1 = \max_{i=1,2}(|x_1|)$. Let us make more precise the notion of the collection of nodes X_1, \dots, X_N . For class $C_{1,L,N}^2$ we denote it by $\Delta = \{X_\mu\}_{\mu=1,\dots,N}$, $N = m^2$. For the definition of Δ we consider an auxiliary grid

$$\begin{cases} \Delta' = \{X'_v\}_{v=1,\dots,N'}, N' = (m+1)^2, X'_v = (x'_{1,i}; x'_{2,j}), \\ v = (i-1)(m+1) + j, x_{1,i} = (i-1)\frac{1}{m}, x_{2,j} = (j-1)\frac{1}{m}, \end{cases} \quad (2.5)$$

with $i = 1, \dots, m+1, j = 1, \dots, m+1$. The grid Δ' splits the region π_2 into m^2 equal squares K_μ , $\mu = 1, \dots, m^2$ with sides $h = 1/m$.

Let us consider the grid $\Delta = \{X\}_{\mu=1,\dots,N}$, nodes of which are the centers of squares K_μ , $\mu = 1, \dots, m^2$. Closest to the sides of K_μ we have rows of nodes of the uniform grid located at distance $1/(2m)$ (in metrics $\|\cdot\|_1$). We note that for any other grid $\tilde{\Delta} = \{X_\mu\}_{\mu=1,\dots,N}$, $N = m^2$ we cannot claim that for any $X \in \pi_2$,

$$\min_{\mu=1,\dots,N} \|X - X_\mu\|_1 \leq \frac{1}{2m}. \quad (2.6)$$

Then we can claim the grid Δ provides an optimal cover of π_2 (see [21] and references therein). Our proofs below will be conducted only for function $A_{C_{1,L,N}^2}^+(X)$. For function $A_{C_{1,L,N}^2}^-(X)$ all proofs are analogous.

For the constructive representation of $A_{C_{1,L,N}^2}^+(X)$ in π_2 we first consider the case when the function is given at nodes $\bar{\Delta} = \Delta \cup \Delta'$, $\bar{\Delta} = \{X_s\}_{s=1,\dots,N}, \bar{N} = N + N'$.

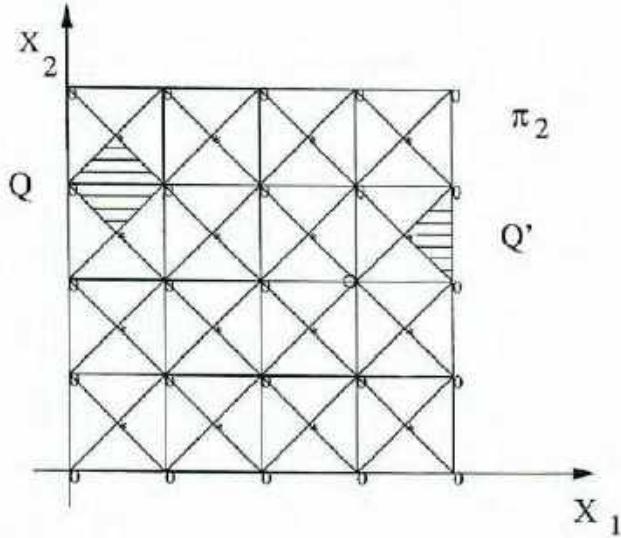


Figure 1: Splitting of the domain π_2 by the grid Δ .

The grid $\bar{\Delta}$ splits square π_2 into regions of the forms Q and Q' (see Fig. 1).

Nodes of the grid Δ' are denoted by circles (\circ) on Fig. 1 and nodes of the grid Δ are denoted by stars (*). It can be shown (see, [20, 21] and references therein), that values of the limit functions from class $C_{1,L,N}^2$ in regions Q and Q' are defined by the values of these functions at vertices of the regions Q and Q' . Let, for functions $f(X)$ from class $C_{1,L,N}^2$,

$$f(X_\mu) = f_\mu, \quad X_\mu \in \Delta \text{ and } f(X'_v) = f_v, \quad X'_v \in \Delta'. \quad (2.7)$$

Then the majorant $A_{F_N}^+(X)$ for $F_N = C_{1,L,N}^2$ has the form

$$A_{C_{1,L,N}^2}^+(X) = \min_{\mu=1,\dots,m; v=1,\dots,(m+1)^2} (f_\mu + L\|X - X_\mu\|_1, f_v + L\|X - X'_v\|_1). \quad (2.8)$$

Let us single out regions of linearity of $A_{C_{1,L,N}^2}^+(X)$. We split region Q into sub-regions $\Omega_l^*, l = 1, \dots, 4$ (Fig. 2) and place the origin at the vertex X_{μ_1} of the region Q . Here

$$\begin{cases} X_{\mu_1} = (0; 0), X'_{v_1} = (h/2; -h/2), X_{\mu_2} = (h; 0), X'_{v_2} = (h/2; h/2), \\ v_1 = (i-1)(m+1) + j, v_2 = (i-1)(m+1) + j + 1, \mu_1 = (i-2)m + j, \\ \mu_2 = (i-1)m + j, i = 1, \dots, m+1, j = 1, \dots, m+1. \end{cases} \quad (2.9)$$

For definiteness, let $f_{v_1} + f_{v_2} > f_{\mu_1} + f_{\mu_2}$ and $f_{\mu_2} > f_{\mu_1}$ (see Fig. 2). Then equations of the lines that split region Q into sub-regions $\Omega_l^*, l = 1, \dots, 4$ have the form

- for the line through O_1, O_2

$$x_1 = \frac{f_{\mu_2} - f_{\mu_1}}{2L} + \frac{h}{2}; \quad (2.10)$$

- for O_1, O'_1

$$x_2 = \frac{f_{v_2} - f_{\mu_1}}{L} + \frac{h}{2} - x_1; \quad (2.11)$$

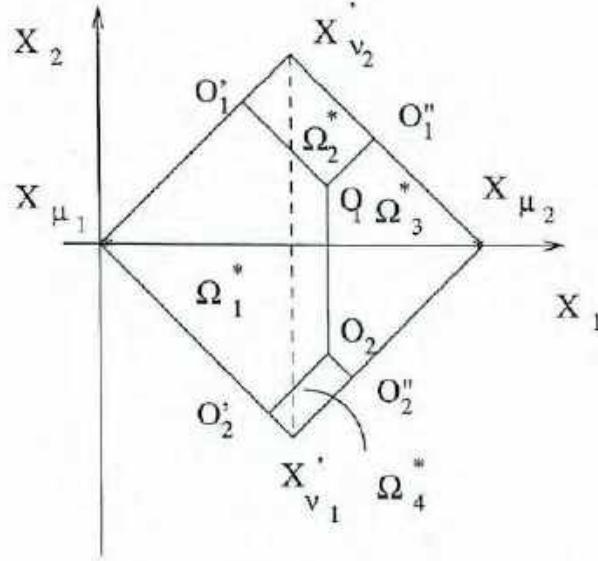


Figure 2: Regions of linearity of the domain Q .

- for O_1, O_1''

$$x_2 = \frac{f_{v_2} - f_{\mu_2}}{L} - \frac{h}{2} + x_1; \quad (2.12)$$

- for O_2, O_2''

$$x_2 = \frac{f_{\mu_1} - f_{v_1}}{L} + \frac{h}{2} - x_1; \quad (2.13)$$

- for O_2, O_2'

$$x_2 = \frac{f_{\mu_1} - f_{v_1}}{L} - \frac{h}{2} + x_1; \quad (2.14)$$

Lemma 2.1 Majorant of class $C_{1,L,N}^2$ for $X \in Q = \bigcup_{i=1}^4 Q_i^*$ has the form

$$A_{C_{1,L,N}^2}^+(X) = \begin{cases} f_{\mu_1} + L\|X - X_{\mu_1}\|_1, & X \in \Omega_1^*, \\ f_{v_2} + L\|X - X'_{v_2}\|_1, & X \in \Omega_2^*, \\ f_{\mu_2} + L\|X - X_{\mu_2}\|_1, & X \in \Omega_3^*, \\ f_{v_1} + L\|X - X'_{v_1}\|_1, & X \in \Omega_4^*. \end{cases} \quad (2.15)$$

Proof. Let

$$\tilde{g}_{v_1}(X) = f_{v_1} + L\|X - X'_{v_1}\|_1, \quad g_{\mu_1}(X) = f_{\mu_1} + L\|X - X_{\mu_1}\|_1. \quad (2.16)$$

Then it is easy to show that

$$g_{\mu_1}(X) \leq g_{\mu_2}(X), \quad g_{\mu_1}(X) \leq \tilde{g}_{v_1}(X), \quad l = 1, 2, X \in \Omega_1^*. \quad (2.17)$$

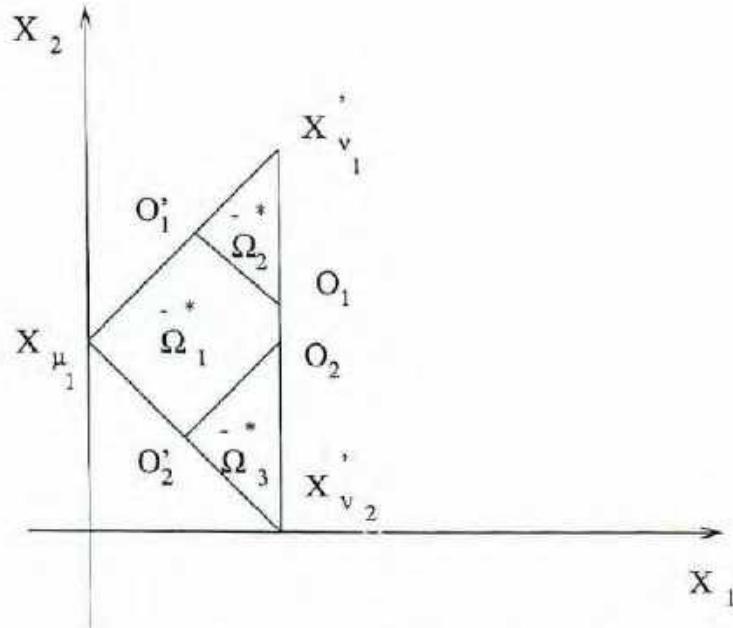


Figure 3: Splitting of the domain Q' when $f_{\mu_1} < \frac{1}{2}(f_{v_1} + f_{v_2})$.

Let us prove this inequality for $g_{\mu_2}(X)$, for example. We have

$$\begin{aligned} g_{\mu_1}(X) - g_{\mu_2}(X) &= f_{\mu_1} + L\|X - X_{\mu_1}\|_1 - f_{\mu_2} - L\|X - X_{\mu_2}\|_1 = \\ f_{\mu_1} - f_{\mu_2} + L|x_1 - x_{1,\mu_1}| - L|x_1 - x_{1,\mu_2}| &= f_{\mu_1} - f_{\mu_2} - Lh \leq 0. \end{aligned} \quad (2.18)$$

We note that our proof is valid for $f_{v_1} + f_{v_2} > f_{\mu_1} + f_{\mu_2}$. In the case $f_{v_1} + f_{v_2} \leq f_{\mu_1} + f_{\mu_2}$ the proof is analogous. ■

Now we consider the region Q' . We split it into three sub-regions as shown in Fig. 3 when $f_{\mu_1} < \frac{1}{2}(f_{v_1} + f_{v_2})$, or as shown in Fig. 4 when $f_{\mu_1} > \frac{1}{2}(f_{v_1} + f_{v_2})$. In the case $f_{\mu_1} = f_{v_1} = f_{v_2}$, the interval O_1O_2 contracts to a point. We have to note that intervals $O_1O'_1$, $O_2O'_2$ can contract to a point under certain relationships between $f_{v_1}, f_{v_2}, f_{\mu_1}$. This remark is also relevant to Fig. 2.

It is easy to see that the equations of the lines that split the region Q' into $\bar{\Omega}_l^*$, $l = 1, \dots, 3$ in the case $f_{\mu_1} < \frac{1}{2}(f_{v_1} + f_{v_2})$ (see Fig. 3) have the form

$$x_2 = \frac{f_{v_2} + f_{\mu_1}}{L} + \frac{h}{2} - x_1 \quad (2.19)$$

for the line through points O_1, O'_1 ;

$$x_2 = \frac{f_{\mu_1} + f_{v_1}}{L} + \frac{h}{2} + x_1 \quad (2.20)$$

O_2, O'_2 ; and in the case $f_{\mu_1} > \frac{1}{2}(f_{v_1} + f_{v_2})$ (see Fig. 4) - the form

$$x_2 = (f_{v_2} - f_{v_1})/(2L) \quad (2.21)$$

for the line through points O_1, O_2 ; and the form of equations of the lines that pass points O_1, O'_1 and O_1, O''_1 as defined by formulae (2.19), (2.20) respectively.

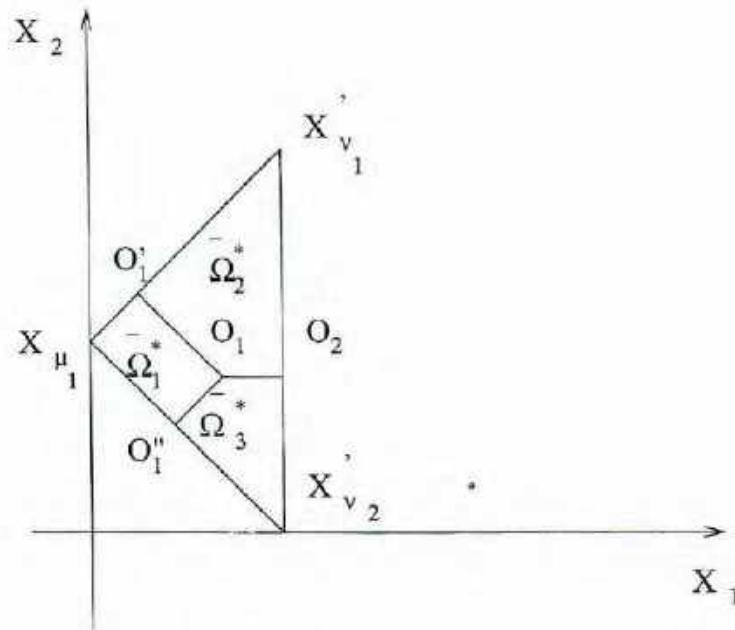


Figure 4: Splitting of the domain Q' when $f_{\mu_1} > \frac{1}{2}(f_{v_1} + f_{v_2})$.

Lemma 2.2 Majorant $A_{C_{1,L,N}^2}^+(X)$ of class $C_{1,L,N}^2$ for $X \in Q' = \bigcup_{l=1}^3 \bar{Q}_l^*$ has the form

$$A_{C_{1,L,N}^2}^+(X) = \begin{cases} f_{\mu_1} + L\|X - X_{\mu_1}\|_1, & X \in \bar{\Omega}_1^*, \\ f_{v_2} + L\|X - X'_{v_2}\|_1, & X \in \bar{\Omega}_2^*, \\ f_{v_1} + L\|X - X'_{v_1}\|_1, & X \in \bar{\Omega}_3^*. \end{cases} \quad (2.22)$$

Proof is analogous to the proof of Lemma 2.1.

Therefore, Lemmas 2.1 and 2.2 allow us to present majorant $A_{C_{1,L,N}^2}^+(X)$ in π_2 in a sufficiently simple form. Indeed, (2.15) and (2.22) can be written as follows

$$A_{C_{1,L,N}^2}^+(X) = \begin{cases} f_{\mu_1} + L|x_1 - x_{1,\mu_1}|, & X \in \Omega_1^*, \\ f_{v_2} + L|x_2 - x'_{2,v_2}|, & X \in \Omega_2^*, \\ f_{\mu_2} + L|x_1 - x_{1,\mu_2}|, & X \in \Omega_3^*, \\ f_{v_1} + L|x_2 - x'_{2,v_1}|, & X \in \Omega_4^*, \end{cases} \quad (2.23)$$

$$A_{C_{1,L,N}^2}^+(X) = \begin{cases} f_{\mu_1} + L|x_1 - x_{1,\mu_1}|, & X \in \bar{\Omega}_1^*, \\ f_{v_2} + L|x_2 - x'_{2,v_2}|, & X \in \bar{\Omega}_2^*, \\ f_{v_1} + L|x_2 - x'_{2,v_1}|, & X \in \bar{\Omega}_3^* \end{cases} \quad (2.24)$$

with $Q = \bigcup_{l=1}^4 Q_l^*$ and $Q' = \bigcup_{l=1}^3 \bar{Q}_l^*$.

An analogous representation can be obtained for the minorant of this class.

Therefore the proposed approach allows us to single out regions of linearity of $A_{C_{1,L,N}}^+(X)$ and $A_{C_{1,L,N}}^-(X)$ and constructively solve the problem of optimal-by-accuracy recovery $f^*(X)$ at point $X \in \pi_2$ of the functions from $C_{1,L,N}^2$. The choice of the grid is implied by the form of regions of linearity of majorant and minorant of class $C_{1,L,N}^2$.

3 The Choice of the Grid in Class $C_{1,L,N}^2$.

Let us now consider the case when function values are only given at the nodes of the grid Δ . In this case for the constructive representation of majorant $A_{C_{1,L,N}}^+(X)$ of class $C_{1,L,N}^2$ we have to perform some additional computations. Let the functions from the class $C_{1,L,N}^2$ take fixed values at the nodes of the grid $\tilde{\Delta} = \Delta \cup \Delta'$ and let the following relationship hold

$$C_{1,L,N}^2 = \{f(X) : |f(X_1) - f(X_2)| \leq L\|X_1 - X_2\|_1, \quad (3.1)$$

$$f(X_\mu) = f(X_\mu), X_\mu \in \Delta, f(X'_v) = A_{C_{1,L,N}}^+(X), X'_v \in \Delta'. \quad (3.2)$$

Then it is easy to see that

$$A_{C_{1,L,N}}^+(X) = A_{C_{1,L,N}}^+(X), X \in \pi_2 \quad (3.3)$$

and

$$A_{C_{1,L,N}}^-(X) \neq A_{C_{1,L,N}}^-(X). \quad (3.4)$$

Therefore, regions of linearity of majorant of the class $C_{1,L,N}^2$ can be singled out analogously to class $C_{1,L,N}^2$. Let us consider the question of computation of values $A_{C_{1,L,N}}^+(X)$ at points $X'_v \in \Delta'$ in detail.

Let $\tilde{\Delta} \subset \Delta'$ be the set of the grid Δ' which lies on the sides of π_2 , and $\Delta^* \subset \tilde{\Delta}$ be the set of vertices of π_2 . Let us choose in the region π_2 elementary regions $\tilde{K}_p, p = 1, \dots, (m_1)^2$, i.e. such squares whose vertices are nodes of the grid Δ . Points $X'_v \in \{\Delta' \setminus \tilde{\Delta}\}$ are centers of these squares (see Fig. 5). In Fig. 5 vertices of regions $\tilde{K}_p, p = 1, \dots, (m-1)^2$ are denoted by (x), and points $X'_v \in \{\Delta' \setminus \tilde{\Delta}\}$ by (o).

Let $\sigma(\tilde{K}_p) = \{s_l\}_{l=1,\dots,4}$, where $s_1 = (i-1)m + j, s_2 = (i-1)m + j + 1, s_3 = im + j + 1, s_4 = im + j$ be the numbers of nodes $X_{s_l}, l = 1, \dots, 4$ that correspond to the vertices of elementary square $\tilde{K}_p, p = (i-1)m + j, i = 1, \dots, m, j = 1, \dots, m$.

Theorem 3.1 *Let $X'_v \in \{\Delta' \setminus \tilde{\Delta}\}$. Then if $X'_v \in \tilde{K}_p$, we have that*

$$A_{C_{1,L,N}}^+(X'_v) = \min_{s \in \sigma(\tilde{K}_p)} f_s + \frac{Lh}{2}, p = 1, \dots, (m-1)^2. \quad (3.5)$$

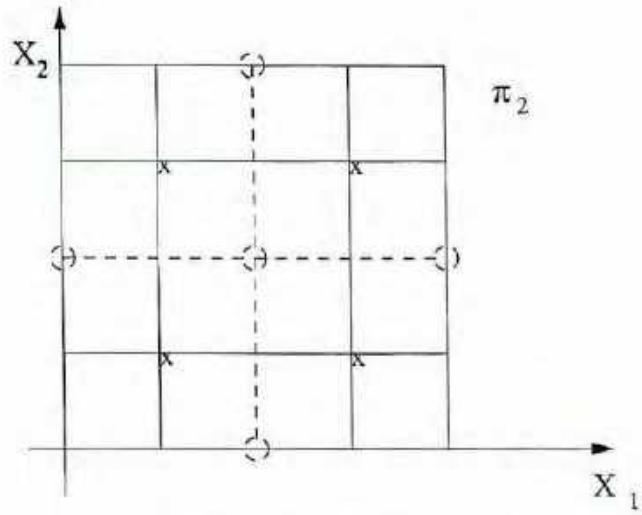


Figure 5: Elementary subregions in π_2 .

Proof. Consider the elementary square \tilde{K}_p . Its vertices are the following nodes of the grid Δ :

$$\begin{cases} X_{s_1} = ((\bar{i} - \frac{1}{2})h; (\bar{j} - \frac{1}{2})h), & X_{s_2} = ((\bar{i} - \frac{1}{2})h; (\bar{j} + \frac{1}{2})h), \\ X_{s_3} = ((\bar{i} + \frac{1}{2})h; (\bar{j} + \frac{1}{2})h), & X_{s_4} = ((\bar{i} + \frac{1}{2})h; m(\bar{j} - \frac{1}{2})h). \end{cases} \quad (3.6)$$

Then the point $X'_{\bar{v}} = (\bar{i}h, \bar{j}h)$ is the center of \tilde{K}_p , $p = 1, \dots, (m-1)^2$, $\bar{i} = 1, \dots, m$, $\bar{j} = 1, \dots, m$. We will show that

$$A_{C_{1,L,N}^2}^+(X'_{\bar{v}}) = \min_{s \in \sigma(\tilde{K}_p)} (f_s + L\|X'_{\bar{v}} - X_s\|_1). \quad (3.7)$$

First, we introduce the following function

$$g_\mu(X) = f_\mu + L\|X - X_\mu\|_1, \quad \mu \in \sigma(\Delta) \quad (3.8)$$

and show that

$$g_{\bar{\mu}}(X'_{\bar{v}}) \geq g_{s_l}(X'_{\bar{v}}), \quad \bar{\mu} \in \sigma(\Delta), \quad \sigma(\tilde{K}_p), l = 1, \dots, 4. \quad (3.9)$$

Let, for example, $X_{\bar{\mu}} = ((\bar{i} - \frac{1}{2} - k_1)h, (\bar{j} - \frac{1}{2} - k_2)h)$, $k_1 = 0, \dots, \bar{i} - 1$, $k_2 = 0, \dots, \bar{j} - 1$.

Then

$$\begin{aligned} g_{\bar{\mu}}(X'_{\bar{v}}) - g_{s_1}(X'_{\bar{v}}) &= f_{\bar{\mu}} + L\|X'_{\bar{v}} - X_{\bar{\mu}}\|_1 - f_{s_1} - L\|X'_{\bar{v}} - X_{s_1}\|_1 = \\ &= f_{\bar{\mu}} - f_{s_1} + L(\max(\bar{i}h - (\bar{i} - \frac{1}{2} - k_1)h, \bar{j}h - (\bar{j} - \frac{1}{2} - k_2)h) - \max(\bar{i}h - \\ &\quad (\bar{i} - \frac{1}{2})h, \bar{j}h - (\bar{j} - \frac{1}{2})h)) = f_{\bar{\mu}} - f_{s_1} + Lh \max(k_1, k_2) \geq \\ &\geq -Lh \max(k_1, k_2) + Lh \max(k_1, k_2) = 0. \end{aligned} \quad (3.10)$$

Analogously, it can be shown that $g_\mu(X'_v) - g_{s_1}(X'_v) \geq 0$ for $l = 2, 3, 4$. It is easy to see that this inequality is satisfied for all others $\mu \in \sigma(\Delta) \setminus \sigma(\tilde{K}_p)$. This means that for all functions $g_\mu(X)$, $\mu \in \sigma(\Delta) \setminus \sigma(\tilde{K}_p)$ the following inequality holds

$$g_\mu(X'_v) \geq \min_{s \in \sigma(\tilde{K}_p)} g_s(X'_v), \quad X'_v \in \tilde{K}_p. \quad (3.11)$$

From (3.11) the statement of the theorem for elementary square \tilde{K}_p , $p = 1, \dots, (m-1)^2$ follows.

Corollary 3.1 *Let $X'_v \in \tilde{\Delta}$, then the following relationship holds*

$$A_{C_{1,L,N}^2}^+(X'_v) = \min(f_{s_1}, f_{s_2}) + \frac{Lh}{2}, \quad (3.12)$$

where s_1, s_2 are numbers of nodes X_{s_1}, X_{s_2} of the grid Δ for which $\|X'_v - X_{s_1}\|_1 = \frac{h}{2}$, $l = 1, 2$.

Corollary 3.2 *Let $X'_v \in \Delta'$, then the following relationship holds*

$$A_{C_{1,L,N}^2}^-(X'_v) = f_{\bar{s}} + \frac{Lh}{2}, \quad (3.13)$$

where \bar{s} is the number of node $X_{\bar{s}} \in \Delta$ for which $\|X'_v - X_{\bar{s}}\|_1 = \frac{h}{2}$.

Proofs of Corollaries 3.1, 3.2 are analogous to the proof of Theorem 3.1.

As mentioned before, the choice of the grid $\tilde{\Delta}$ is dictated by the form of the regions of linearity of functions $A_{C_{1,L,N}^2}^+(X), A_{C_{1,L,N}^2}^-(X)$. The difficulty of the realisation of the approach (1.3)–(1.5) lies with the constructive computation of $I_i^+(C_{1,L,N}^2), I_i^-(C_{1,L,N}^2)$, $i = 1, 2, 3$. In computing $I_i^+(C_{1,L,N}^2)$ we extend $f(X) = A_{C_{1,L,N}^2}^+(X)$, $X \in \Delta'$ and for computing $I_i^-(C_{1,L,N}^2)$ we extend $f(X) = A_{C_{1,L,N}^2}^-(X)$, $X \in \Delta'$, $i = 1, 2, 3$.

Lemma 3.1 *For the majorant of class $C_{1,L,N}^2$ the following relationship holds*

$$A_{C_{1,L,N}^2}^+(X) = \min_{\mu=1,\dots,m; v=1,\dots,(m+1)^2} (f_\mu + L\|X - X_\mu\|_1, f_v + L\|X - X'_v\|_1), \quad (3.14)$$

where $f_\mu = f(X_\mu)$ for $X_\mu \in \Delta$ and $f_v = A_{C_{1,L,N}^2}^+(X'_v)$ for $X'_v \in \Delta'$.

The proof of Lemma 3.1 is conducted analogously to the proof of Lemma 4.1 from [17]. Analogously it can also be shown that

$$A_{C_{1,L,N}^2}^-(X) = \max_{\mu=1,\dots,m; v=1,\dots,(m+1)^2} (f_\mu - L\|X - X_\mu\|_1, f_v - L\|X - X'_v\|_1). \quad (3.15)$$

Therefore, Theorem 3.1, its Corollaries and Lemma 3.1 allow us to substantially reduce the amount of *a priori* information that is necessary for the application of the approach proposed in Section 2. Instead of $2m^2 + 2m + 1$ function values at the nodes of the grid $\tilde{\Delta}$ it is sufficient to provide function values only at m^2 nodes of the grid Δ which is the optimal covering of π_2 . Then the values of majorant and minorant of class $C_{1,L,N}^2$ at nodes of the grid Δ' are computed by the relationships (3.5), (3.12) and (3.13). This allows us to constructively solve the problem of optimal-by-accuracy recovery $f(X)$ at point $X \in \pi_2$ of functions from $C_{1,L,N}^2$. The results analogous to those developed in [17] also hold.

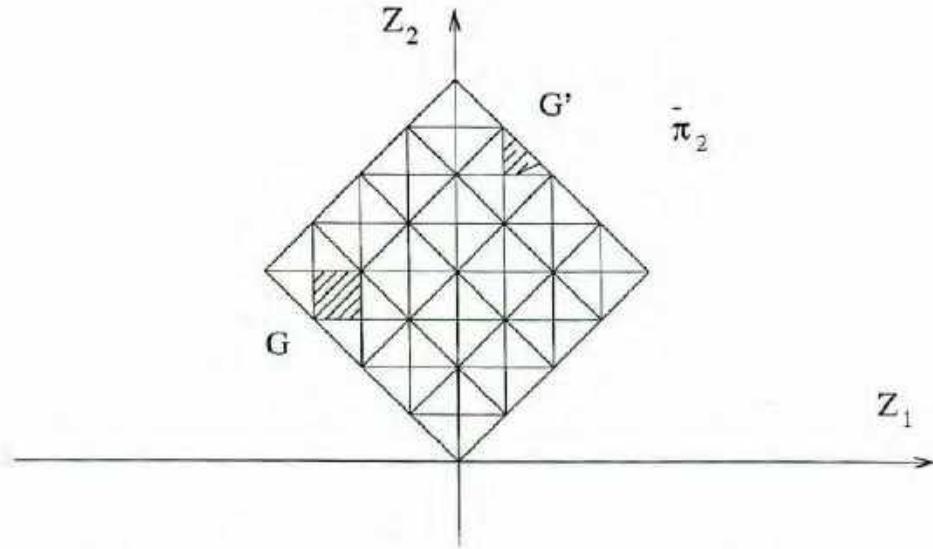


Figure 6: Domain π_2 in the new system of coordinates.

4 On Optimal Integration of Product of Functions in Class $C_{1,L,N}^2$.

Let us consider integrals of the form (1.1) where $f(x_1, x_2) \in X_{1,L,N}^2$ and $\varphi_1(x_1), \varphi_2(x_2)$ are known integrable functions.

In previous sections we studied properties of the functions $A_{C_{1,L,N}^2}^+(X), A_{C_{1,L,N}^2}^-(X)$. Each of the regions of the form Q and Q' were split into sub-regions, in which functions $A_{C_{1,L,N}^2}^+(X), A_{C_{1,L,N}^2}^-(X)$ were linear. We note that in the case when $\varphi_1(x_1), \varphi_2(x_2)$ have simple forms (for example, when $\varphi_1(x_1) = \varphi_2(x_2) = 1$ or $\varphi_1(x_1) = \sin \omega_1 x_1, \varphi_2(x_2) = \sin \omega_2 x_2$ or $\varphi_1(x_1) = \cos \omega_1 x_1, \varphi_2(x_2) = \cos \omega_2 x_2$), the computing $I^+(C_{1,L,N}^2)$ and $I^-(C_{1,L,N}^2)$ is a relatively easy procedure.

In order to simplify computations, we pass from the Cartesian system of coordinates (x_1, x_2) to a new system of coordinates by rotating coordinate axes about the angle $\alpha = -45^\circ$ (see Fig. 6):

$$Z_1 = x_1 \cos \alpha + x_2 \sin \alpha, \quad Z_2 = -x_1 \sin \alpha + x_2 \cos \alpha. \quad (4.1)$$

Therefore

$$\begin{cases} Z_1 = \frac{\sqrt{2}}{2}(x_1 - x_2), \\ Z_2 = \frac{\sqrt{2}}{2}(x_1 + x_2), \end{cases} \Rightarrow \begin{cases} x_1 = \frac{1}{\sqrt{2}}(Z_1 + Z_2), \\ x_2 = \frac{1}{\sqrt{2}}(Z_2 - Z_1). \end{cases} \quad (4.2)$$

In the transfer to the new coordinate system, the region π_2 (see Fig. 1) is transformed into the region $\bar{\pi}_2$ (see Fig. 6), and regions Q and Q' (Fig. 1) - into regions G and G' respectively (Fig. 6).

Let

- $\tilde{G}_p, p = 1, \dots, m(m-1)$ be regions of the form G that are located on the right of axis Oz_2 , and $\tilde{G}_p, p = 1, \dots, m(m-1)$ be regions of the form G that are located on the left of the axis Oz_2 ;
- $G'_l, l = 1, \dots, 2m$ be regions of the form G' that are located on the right of axis Oz_2 , and $\tilde{G}'_l, l = 1, \dots, 2m$ be regions of the form G' that are located on the left of the axis Oz_2 ;

Let also the points $Z_{p_k}, k = 1, \dots, 4$ be vertices of elementary region \tilde{G}_p , and $f_{p_k}, k = 1, \dots, 4$ be values of function $A_{C_{1,L,N}^2}^+(X)$ at these vertices. Then

$$Z_{p_1} = (z_{1,j_1}; z_{2,j_2}), Z_{p_2} = (z_{1,j_1}; z_{2,j_2+1}), Z_{p_3} = (z_{1,j_1+1}; z_{2,j_2+1}), Z_{p_4} = (z_{1,j_1+1}; z_{2,j_2}), \quad (4.3)$$

where

$$z_{1,j_1} = (j_1 - 1)h_1, z_{2,j_2} = (j_2 - 1)h_1, h_1 = h/\sqrt{2}, \quad (4.4)$$

and

$$f_{p_1} = f_{j_1,j_2}, f_{p_2} = f_{j_1,j_2+1}, f_{p_3} = f_{j_1+1,j_2+1}, f_{p_4} = f_{j_1+1,j_2}. \quad (4.5)$$

Here $j_1 = 1, \dots, j_2 - 1$ for $j_2 = 2, \dots, m$ and $j_1 = 1, \dots, 2m - j_2$ for $j_2 = m + 1, \dots, 2m - 1$, and

$$p = \begin{cases} (j_2 - 2)(j_2 - 1)/2 + j_1, & j_1 = 1, \dots, j_2 - 1, j_2 = 2, \dots, m; \\ (j_2 - 1)(4m - j_2)/2 - m^2 + j_1, & j_1 = 1, \dots, 2m - j_2, j_2 = m + 1, \dots, 2m - 1. \end{cases} \quad (4.6)$$

Setting $h_1 = h/\sqrt{2}$ we analogously determine vertices of the region \tilde{G}_p , where p is computed by (4.6) ($p = 1, \dots, m(m-1)$).

Let $Z'_{t_k}, k = 1, 2, 3$ be vertices of the elementary region G'_l , and $f_{t_k}, k = 1, 2, 3$ be values of the function $A_{C_{1,L,N}^2}^+(X)$ at these vertices. Then

$$\begin{cases} Z'_{t_1} = (z_{1,j_1}; z_{2,j_2}), Z'_{t_2} = (z_{1,j_1}; z_{2,j_2+1}), Z'_{t_3} = (z_{1,j_1+1}; z_{2,j_2+1}), \\ f_{t_1} = f_{j_1,j_2}, f_{t_2} = f_{j_1,j_2+1}, f_{t_3} = f_{j_1+1,j_2+1} \end{cases} \quad (4.7)$$

for $j_1 = j_2, j_2 = 1, \dots, m$ and

$$\begin{cases} Z'_{t_1} = (z_{1,j_1}; z_{2,j_2}), Z'_{t_2} = (z_{1,j_1}; z_{2,j_2+1}), Z'_{t_3} = (z_{1,j_1+1}; z_{2,j_2}) \\ f_{t_1} = f_{j_1,j_2}, f_{t_2} = f_{j_1,j_2+1}, f_{t_3} = f_{j_1+1,j_2} \end{cases} \quad (4.8)$$

for $j_2 = m + 1, \dots, 2m, j_1 = 2m - j_2 + 1$ and $l = j_2, j_2 = 1, \dots, 2m$.

Setting $h_1 = -h/\sqrt{2}$, in a similar way we determine the vertices of the region $\tilde{G}'_l, l = 1, \dots, 2m$.

We introduce the following notation

$$\begin{cases} \tilde{I}_p^* = \frac{1}{2} \int \int_{\tilde{G}_p} \left(\tilde{A}_{C_{1,L,N}^2}^+(Z) + \tilde{A}_{C_{1,L,N}^2}^-(Z) \right) \tilde{\varphi}_1(z_1) \tilde{\varphi}_2(z_2) dZ, \\ \tilde{I}_p^* = \frac{1}{2} \int \int_{\tilde{G}_p} \left(\tilde{A}_{C_{1,L,N}^2}^+(Z) + \tilde{A}_{C_{1,L,N}^2}^-(Z) \right) \bar{\varphi}_1(z_1) \bar{\varphi}_2(z_2) dZ, \quad p = 1, \dots, m(m-1); \\ \tilde{I}_l^* = \frac{1}{2} \int \int_{\tilde{G}'_l} \left(\tilde{A}_{C_{1,L,N}^2}^+(Z) + \tilde{A}_{C_{1,L,N}^2}^-(Z) \right) \tilde{\varphi}_1(z_1) \tilde{\varphi}_2(z_2) dZ, \\ \tilde{I}_l^* = \frac{1}{2} \int \int_{\tilde{G}'_l} \left(\tilde{A}_{C_{1,L,N}^2}^+(Z) + \tilde{A}_{C_{1,L,N}^2}^-(Z) \right) \bar{\varphi}_1(z_1) \bar{\varphi}_2(z_2) dZ, \quad l = 1, \dots, 2m \end{cases} \quad (4.9)$$

where $\tilde{A}_{C_{1,L,N}^2}^\pm(Z)$ are majorant and minorant of the functional class $C_{1,L,N}^2$ in region $\bar{\pi}_2$ and $\tilde{\varphi}_1(z_1)\tilde{\varphi}_2(z_2) = \varphi_1\left(\frac{1}{\sqrt{2}}(z_1 + z_2)\right)\varphi_2\left(\frac{1}{\sqrt{2}}(z_2 - z_1)\right)$, $(z_1, z_2) \in \bar{\pi}_2$. It is easy to see that

$$\tilde{A}_{C_{1,L,N}^2}^+(Z) = \begin{cases} f_{j_1,j_2} + \frac{L}{\sqrt{2}}(z_1 + z_2 - h(j_1 + j_2 - 2)), & Z \in \Gamma_1^*, \\ f_{j_1,j_2+1} + \frac{L}{\sqrt{2}}(z_1 - z_2 + h_1(j_2 + 1 - j_1)), & Z \in \Gamma_2^*, \\ f_{j_1+1,j_2+1} + \frac{L}{\sqrt{2}}(h_1(j_1 + j_2) - z_1 - z_2), & Z \in \Gamma_3^*, \\ f_{j_1+1,j_2} + \frac{L}{\sqrt{2}}(z_2 - z_1 + h_1(j_1 + 1 - j_2)), & Z \in \Gamma_4^*, \end{cases} \quad (4.10)$$

where $G = \bigcup_{k=1}^4 \Gamma_k^*$, and $\Gamma_k^*, k = 1, 2, 3, 4$ are regions of linearity of the function $\tilde{A}_{C_{1,L,N}^2}^*(Z)$ in the region $G \subset \bar{\pi}_2$ which are defined by transformations of the regions $\Omega_l^*, l = 1, 2, 3, 4$, $\bigcup_{l=1}^4 \Omega_l^* = Q$. Further we have

$$\tilde{A}_{C_{1,L,N}^2}^+(Z) = \begin{cases} f_{j_1,j_2} + \frac{L}{\sqrt{2}}(z_1 + z_2 - h_1(j_1 + j_2 - 2)), & Z \in \Gamma_1^*, \\ f_{j_1,j_2+1} + \frac{L}{\sqrt{2}}(z_1 - z_2 + h_1(j_2 + 1 - j_1)), & Z \in \tilde{\Gamma}_2^*, \\ f_{j_1+1,j_2+1} + \frac{L}{\sqrt{2}}(h_1(j_1 + j_2) - z_1 - z_2), & Z \in \tilde{\Gamma}_3^* \end{cases} \quad (4.11)$$

with $\bigcup_{k=1}^3 \tilde{\Gamma}_k^* = \tilde{G}'_l, l = 1, \dots, m$ or $\bigcup_{k=1}^3 \tilde{\Gamma}_k^* = \tilde{G}_l, l = 1, \dots, m$ and

$$\tilde{A}_{C_{1,L,N}^2}^+(Z) = \begin{cases} f_{j_1,j_2} + \frac{L}{\sqrt{2}}(z_1 + z_2 - h_1(j_1 + j_2 - 2)), & Z \in \tilde{\Gamma}_1^*, \\ f_{j_1,j_2+1} + \frac{L}{\sqrt{2}}(z_1 - z_2 + h_1(j_2 + 1 - j_1)), & Z \in \tilde{\Gamma}_2^*, \\ f_{j_1+1,j_2} + \frac{L}{\sqrt{2}}(z_2 - z_1 + h_1(j_1 + 1 - j_2)), & Z \in \tilde{\Gamma}_3^* \end{cases} \quad (4.12)$$

with $\bigcup_{k=1}^3 \tilde{\Gamma}_k^* = \tilde{G}'_l, l = m+1, \dots, 2m$ or $\bigcup_{k=1}^3 \tilde{\Gamma}_k^* = \tilde{G}_l, l = m+1, \dots, 2m$. Here $\tilde{\Gamma}_k^*, \tilde{\Gamma}_k, k = 1, 2, 3$ are regions of linearity of the functions $\tilde{A}_{C_{1,L,N}^2}^\pm(Z)$ in $\tilde{G}_l, \tilde{G}'_l, l = 1, \dots, 2m$ from $\bar{\pi}_2$, which are determined by transformations of the regions $\tilde{\Omega}_l^*, l = 1, 2, 3, \bigcup_{l=1}^3 \tilde{\Omega}_l^* = Q'$.

It is easy to see that in the new system of coordinates, the majorant of the class $C_{1,L,N}^2$ in the elementary regions $\tilde{G}_p, \tilde{G}'_p, p = 1, \dots, m(m-1)$ has the form analogous to the form of the function $A_{C_{2,L,N}^2}^+(X)$ in the elementary region $K_p, p = 1, \dots, m^2$. Furthermore, the equations of the lines that split the regions $\tilde{G}_p, \tilde{G}'_p, p = 1, \dots, m(m-1)$ into the regions $\Gamma_k^*, k = 1, 2, 3, 4$ are analogous to the equations of the lines that split $K_p, p = 1, \dots, m^2$ into the regions $\Omega_l^+, l = 1, 2, 3, 4$. Indeed, let us consider the elementary region $\tilde{G}_p, p = 1, \dots, m(m-1)$. We place the origin at the left lower vertex of this region. Then $Z_{p_1} = (0, 0), Z_{p_2} = (0, h_1), Z_{p_3} = (h_1, h_1), Z_{p_4} = (h_1, 0)$. Let for the sake of definiteness $f_{p_2} + f_{p_4} > f_{p_1} + f_{p_3}, f_{p_3} > f_{p_1}$. Then the

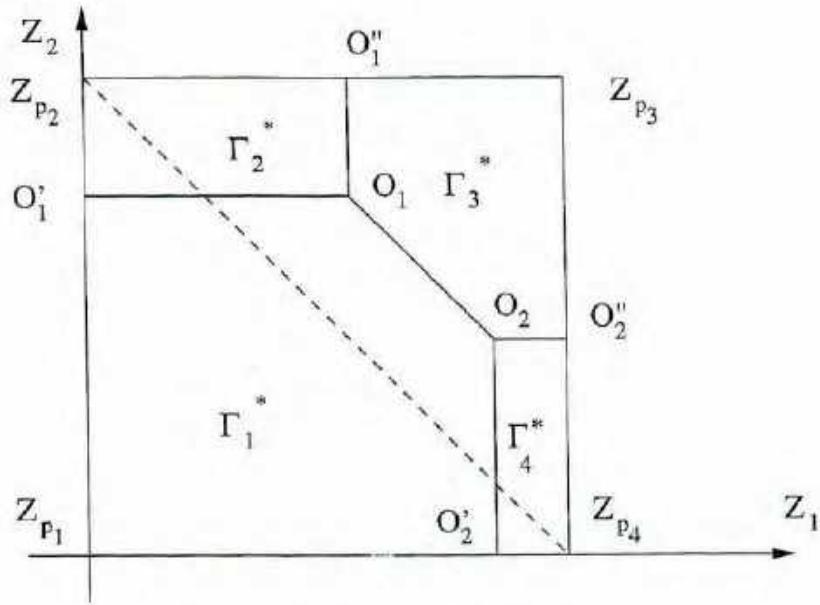


Figure 7: Splitting the domain \tilde{G}_p .

splitting of the region \tilde{G}_p into subregions Γ_k^* , $k = 1, 2, 3, 4$ is performed in the way shown in Fig. 7, and the equations of the lines that split \tilde{G}_p into these sub-regions have the following forms

$$z_2 = -z_1 + \frac{f_{p_3} - f_{p_1}}{\sqrt{2L}} + h_1 \quad (4.13)$$

for the line through points O_1, O_2 ;

$$z_2 = \frac{f_{p_2} - f_{p_1}}{\sqrt{2L}} + \frac{h_1}{2} \quad (4.14)$$

for O_1, O_1' ;

$$z_1 = \frac{f_{p_3} - f_{p_2}}{\sqrt{2L}} + \frac{h_1}{2} \quad (4.15)$$

for O_1, O_1'' ;

$$z_2 = \frac{f_{p_3} - f_{p_4}}{\sqrt{2L}} + \frac{h_1}{2} \quad (4.16)$$

for O_2, O_2'' ;

$$z_1 = \frac{f_{p_4} - f_{p_1}}{\sqrt{2L}} + \frac{h_1}{2} \quad (4.17)$$

for O_2, O_2' .

Therefore the problems set in Section 2 are reducible to the solution of problems in the functional class $C_{2,L,L,N}^2$ and all results obtained in [17] for $C_{2,L,L,N}^2$ hold for the functional class $C_{1,L,N}^2$.

Theorem 4.1 Let $f(X) \in C_{1,L,N}^2$. Optimal-by-accuracy cubature formula for computing integrals (1.1) in the case when functions $\tilde{\varphi}_1(z_1), \tilde{\varphi}_2(z_2)$ do not change their signs for $(z_1, z_2) \in \bar{\pi}_2$ has the form

$$I^*(C_{1,L,N}^2) = \sum_{p=1}^{m(m-1)} (\bar{I}_p^* + \bar{\bar{I}}_p^*) + \sum_{l=1}^{2m} (\bar{I}_l^* + \bar{\bar{I}}_l^*), \quad (4.18)$$

and

$$\delta(C_{1,L,N}^2) = \frac{1}{2} \int \int_{\bar{\pi}_2} \left(\tilde{A}_{C_{1,L,N}^2}^+(Z) - \tilde{A}_{C_{1,L,N}^2}^-(Z) \right) |\tilde{\varphi}_1(z_1)\tilde{\varphi}_2(z_2)| dZ. \quad (4.19)$$

Proof. It is clear that for $\tilde{\varphi}_1(z_1)\tilde{\varphi}_2(z_2) > 0$ for all $Z = (z_1, z_2) \in \bar{\pi}_2$ we have

$$I^\pm(C_{1,L,N}^2) = \int \int_{\bar{\pi}_2} \tilde{A}_{C_{1,L,N}^2}^\pm(Z) \tilde{\varphi}_1(z_1)\tilde{\varphi}_2(z_2) dZ, \quad (4.20)$$

and for $\tilde{\varphi}_1(z_1)\tilde{\varphi}_2(z_2) < 0$ we have

$$I^\pm(C_{1,L,N}^2) = \int \int_{\bar{\pi}_2} \tilde{A}_{C_{1,L,N}^2}^\pm(Z) \tilde{\varphi}_1(z_1)\tilde{\varphi}_2(z_2) dZ. \quad (4.21)$$

Taking into consideration (4.9), the statement of the theorem follows from (4.21) ■

Corollary 4.1 Let functions $\tilde{\varphi}_1(z_1), \tilde{\varphi}_2(z_2)$ change their signs for $(z_1, z_2) \in \bar{\pi}_2$. Then the error of the cubature formula (4.18) will not exceed more than 2 times the optimal.

The proof of this result is analogous to the proof of Corollary 3.1 from [17] (see also [22, 23]).

We note that for the construction of cubature formulae for computing integrals $I^2(f)$ in class $C_{1,L,N}^2$ (as well as for classes $C_{2,L_1,L_2,N}^2$ and $C_{2,L,L,N}^2$) we can use the method of approximation of integrand by a linear spline [9, 25, 26, 18]. However, if the zeros of the functions $\tilde{\varphi}_1(z_1), \tilde{\varphi}_2(z_2)$ are located in a relatively sparse manner with respect to the nodes of the grid, then the approach described in this paper has advantages. Indeed, in regions of constant sign of functions $\tilde{\varphi}_1(z_1), \tilde{\varphi}_2(z_2)$, formula (4.18) is optimal-by-accuracy. In addition the proposed approach allows us to construct error estimates for formula (4.18)

$$v(C_{1,L,N}^2, I^*, \bar{f}) \leq \frac{1}{2} \int \int_{\bar{\pi}_2} \left(\tilde{A}_{C_{1,L,N}^2}^+(Z) - \tilde{A}_{C_{1,L,N}^2}^-(Z) \right) |\tilde{\varphi}_1(z_1)\tilde{\varphi}_2(z_2)| dZ. \quad (4.22)$$

5 Optimal-By-Accuracy Cubature Formula for Functions from Class $C_{1,L,N}^2$.

Let us consider the problem of construction of optimal-by-accuracy cubature formula for computing integral (1.6) which is an important special case of the integral $I^2(f)$. As before we perform the rotation of the Cartesian system of coordinates about the angle $\alpha = -45^\circ$. Then

$$I_1^2(f) = \int \int_{\bar{\pi}_2} \bar{f}(Z) dZ, \quad (5.1)$$

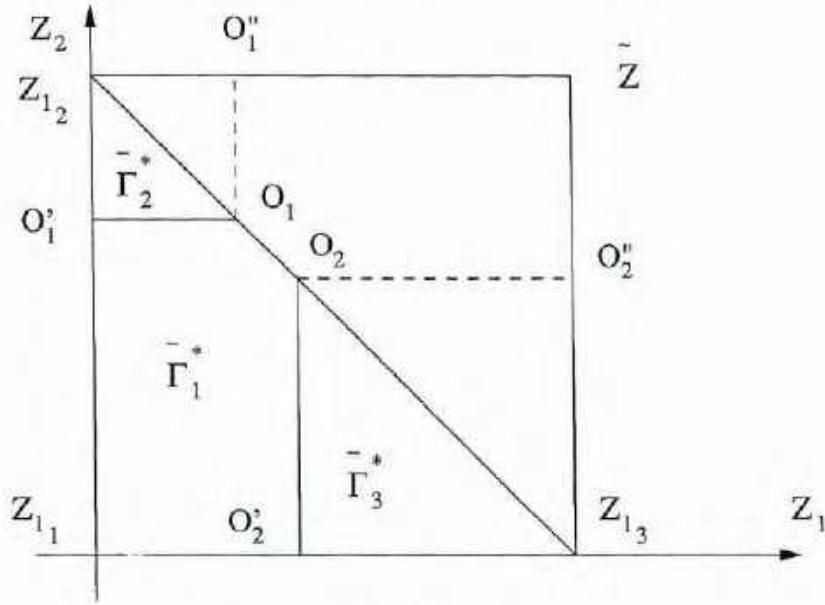


Figure 8: Expansion of \tilde{G}'_l when $f_{l_1} < \frac{1}{2}(f_{l_2} + f_{l_3})$.

where

$$\bar{f}(Z) = f\left(\frac{1}{\sqrt{2}}(z_2 + z_1), \frac{1}{\sqrt{2}}(z_2 - z_1)\right). \quad (5.2)$$

For convenience in computing integrals in regions of the form G' we expand all G' to regions of the form G , setting the value of the majorant in each additional vertex \tilde{Z} equal to the value at that vertex of G' which is symmetric to \tilde{Z} with respect to the center of the region G (see Fig. 8 and Fig. 9).

In Fig. 8 the expansion of elementary region $\tilde{G}'_l, l = m+1, \dots, 2m$ to the region of the form G is shown in the case when $f_{l_1} < \frac{1}{2}(f_{l_2} + f_{l_3})$. In Fig. 9 such an expansion is shown in the case when $f_{l_1} > \frac{1}{2}(f_{l_2} + f_{l_3}), f_{l_1} = f_{j_1, j_2}, f_{l_2} = f_{j_1, j_2+1}, f_{l_3} = f_{j_1+1, j_2}, j_1 = 2m - j_2 + 1, j_2 = m+1, \dots, 2m$. The expansion of the regions $\tilde{G}'_l, l = 1, \dots, m$ and $\tilde{G}'_l, l = 1, \dots, 2m$ is performed analogously. Corresponding representations can also be obtained for the minorant of this class. Let

$$\bar{R}_p^\pm = \int \int_{\tilde{G}_p} \tilde{A}_{C_{1,L,N}^2}^\pm(Z) dZ, \quad \bar{\bar{R}}_p^\pm = \int \int_{\bar{\tilde{G}}_p} \tilde{A}_{C_{1,L,N}^2}^\pm(Z) dZ \quad (5.3)$$

where $p = 1, \dots, m(m+1)$ and

$$\tilde{R}_l^\pm = \int \int_{\tilde{G}'_l} \tilde{A}_{C_{1,L,N}^2}^\pm(Z) dZ, \quad \bar{\bar{R}}_l^\pm = \int \int_{\bar{\tilde{G}}'_l} \tilde{A}_{C_{1,L,N}^2}^\pm(Z) dZ \quad (5.4)$$

where $l = 1, \dots, 2m$.

It is easy to see that

$$\bar{R}_l^\pm = \frac{1}{2} \int \int_{\tilde{G}'_l} \tilde{A}_{C_{1,L,N}^2}^\pm(Z) dZ, \quad \bar{\bar{R}}_l^\pm = \frac{1}{2} \int \int_{\bar{\tilde{G}}'_l} \tilde{A}_{C_{1,L,N}^2}^\pm(Z) dZ \quad (5.5)$$

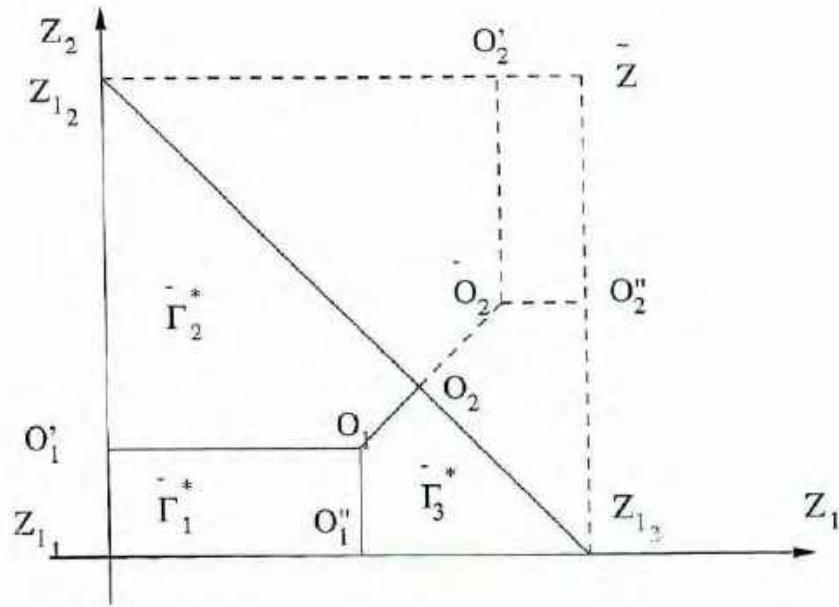


Figure 9: Expansion of \bar{G}'_l when $f_{l_1} > \frac{1}{2}(f_{l_2} + f_{l_3})$.

where $l = 1, \dots, 2m$ and $\bar{G}'_l, \bar{\bar{G}}'_l$ represent regions of the form $\bar{G}'_l, \bar{\bar{G}}'_l$ respectively expanded to the regions of the form G as explained above.

Let

$$\begin{cases} \bar{\xi}_{1,j_1} = z_{1,j_1} + \frac{h_1}{2} + \sigma \left(\frac{f_{j_1+1,j_2} - f_{j_1,j_2}}{\sqrt{2L}} \mu_1 + \frac{f_{j_1+1,j_2+2} - f_{j_1,j_2+1}}{\sqrt{2L}} \mu \right), \\ \bar{\bar{\xi}}_{1,j_1} = z_{1,j_1} + \frac{h_1}{2} + \sigma \left(\frac{f_{j_1+1,j_2+2} - f_{j_1,j_2+1}}{\sqrt{2L}} \mu_1 + \frac{f_{j_1+1,j_2} - f_{j_1,j_2}}{\sqrt{2L}} \mu \right), \\ \bar{\xi}_{2,j_2} = z_{2,j_2} + \frac{h_1}{2} + \sigma \frac{f_{j_1+1,j_2+1} - f_{j_1+1,j_2}}{\sqrt{2L}}, \quad \bar{\bar{\xi}}_{2,j_2} = z_{2,j_2} + \frac{h_1}{2} + \sigma \frac{f_{j_1,j_2+1} - f_{j_1,j_2}}{\sqrt{2L}}, \end{cases} \quad (5.6)$$

where

$$\mu_1 = \frac{1}{2}((1-\sigma)\gamma_2 + (1+\sigma)\gamma_1), \quad \mu_2 = \frac{1}{2}((1-\sigma)\gamma_1 + (1+\sigma)\gamma_2), \quad (5.7)$$

and

$$\begin{cases} \gamma_1 = \frac{1}{2}(1 - \text{sign}(f_{j_1,j_2} + f_{j_1+1,j_2+1} - f_{j_1,j_2+1} - f_{j_1+1,j_2})), \\ \gamma_2 = \frac{1}{2}(1 + \text{sign}(f_{j_1,j_2} + f_{j_1+1,j_2+1} - f_{j_1,j_2+1} - f_{j_1+1,j_2})), \end{cases} \quad (5.8)$$

$\sigma \in \{-1, 1\}$; $j_1 = 1, \dots, j_2 - 1$ for $j_2 = 2, \dots, m$ and $j_1 = 1, \dots, 2m - j_2$ for $j_2 = m + 1, \dots, 2m - 1$. Then the equation of the line that passes through points $(\bar{\xi}_{1,j_1}, \bar{\xi}_{2,j_2}), (\bar{\bar{\xi}}_{1,j_1}, \bar{\bar{\xi}}_{2,j_2})$ has the form

$$\frac{z_1 - \bar{\xi}_{1,j_1}}{\bar{\xi}_{1,j_1} - \bar{\bar{\xi}}_{1,j_1}} = \frac{z_2 - \bar{\xi}_{2,j_2}}{\bar{\xi}_{2,j_2} - \bar{\bar{\xi}}_{2,j_2}}, \quad (5.9)$$

or equivalently

$$(z_2 - \bar{\xi}_{2,j_2})(\bar{\xi}_{1,j_1} - \bar{\xi}_{1,j_1}) = (z_1 - \bar{\xi}_{1,j_1})(\bar{\xi}_{2,j_2} - \bar{\xi}_{2,j_2}). \quad (5.10)$$

Since

$$\bar{\xi}_{1,j_1} - \bar{\xi}_{1,j_1} = \sigma(\mu_1 - \mu_2) \frac{1}{\sqrt{2L}} (f_{j_1+1,j_2+1} - f_{j_1,j_2+1} - f_{j_1+1,j_2} - f_{j_1,j_2}) \quad (5.11)$$

and

$$\bar{\xi}_{2,j_2} - \bar{\xi}_{2,j_2} = \sigma(\mu_1 - \mu_2) \frac{1}{\sqrt{2L}} (f_{j_1,j_2+1} + f_{j_1+1,j_2} - f_{j_1+1,j_2+1} - f_{j_1,j_2}), \quad (5.12)$$

then

$$(\bar{\xi}_{2,j_2} - \bar{\xi}_{2,j_2}) / (\bar{\xi}_{1,j_1} - \bar{\xi}_{1,j_1}) = \mu_2 - \mu_1. \quad (5.13)$$

Finally, from (5.13) we get

$$z_2 = \bar{\xi}_{2,j_2} + (\mu_1 - \mu_2)(\bar{\xi}_{1,j_1} - z_1), \quad (5.14)$$

$j_1 = 1, \dots, j_2 - 1$ for $j_2 = 2, \dots, m$ and $j_1 = 1, \dots, 2m - j_2$ for $j_2 = m + 1, \dots, 2m - 1$.

It is obvious that equation (5.14) is the equation of the line O_1O_2 for $\sigma = 1$ (see Fig. 7). It is easy to show that for $\sigma = -1$ points $(\bar{\xi}_{1,j_1}, \bar{\xi}_{2,j_2})$ and $(\bar{\xi}_{1,j_1}, \bar{\xi}_{2,j_2})$ are the points that determine the splitting of the region of the form G' into regions of linearity of the function $\bar{A}_{C_{1,L,N}^2}(Z)$, $j_1 = 1, \dots, j_2 - 1$ for $j_2 = 2, \dots, m$ and $j_1 = 1, \dots, 2m - j_2$ for $j_2 = m + 1, \dots, 2m - 1$. From the relationships (4.22) and (4.13)–(4.17) it immediately follows

$$\begin{aligned} \bar{R}_p^\pm &= \int_{z_{1,j_1}}^{\bar{\xi}_{1,j_1}} \int_{z_{2,j_2}}^{\bar{\xi}_{2,j_2}} (f_{j_1,j_2} + \frac{\sigma L}{\sqrt{2}}(z_1 - z_{1,j_1} + z_2 - z_{2,j_2})) dz_2 dz_1 + \\ &\quad \int_{z_{1,j_1}}^{\bar{\xi}_{1,j_1}} \int_{\bar{\xi}_{2,j_2}}^{z_{2,j_2+1}} (f_{j_1,j_2+1} + \frac{\sigma L}{\sqrt{2}}(z_1 - z_{1,j_1} + z_{2,j_2+1} - z_2)) dz_2 dz_1 + \\ &\quad \int_{\bar{\xi}_{1,j_1}}^{z_{1,j_1+1}} \int_{\bar{\xi}_{2,j_2}}^{z_{2,j_2+1}} (f_{j_1+1,j_2+2} + \frac{\sigma L}{\sqrt{2}}(z_{1,j_1+1} + z_{2,j_2+1} - z_1 - z_2)) dz_2 dz_1 + \\ &\quad \int_{\bar{\xi}_{1,j_1}}^{z_{1,j_1+1}} \int_{z_{2,j_2}}^{\bar{\xi}_{2,j_2}} (f_{j_1+1,j_2} + \frac{\sigma L}{\sqrt{2}}(z_{1,j_1+1} - z_{2,j_2} - z_1 + z_2)) dz_2 dz_1 + \\ &\quad \int_{\bar{\xi}_{1,j_1}}^{\bar{\xi}_{1,j_1}} \int_{z_{2,j_2}}^{\bar{\xi}_{2,j_2} + (\mu_1 - \mu_2)(\bar{\xi}_{2,j_2} - z_1)} \left(\mu_1 f_{j_1,j_2} + \mu_2 f_{j_1+1,j_2} + \frac{\sigma L}{\sqrt{2}} ((\mu_1 - \mu_2) z_1 - \right. \\ &\quad \left. \mu_1 z_{1,j_1} + \mu_2 z_{1,j_1+1} + z_2 - z_{2,j_2})) dz_2 dz_1 + \int_{\bar{\xi}_{1,j_1}}^{\bar{\xi}_{1,j_1}} \int_{\bar{\xi}_{2,j_2} + (\mu_1 - \mu_2)(\bar{\xi}_{1,j_1} - z_2)}^{z_{2,j_2+1}} (\mu_1 f_{j_1+1,j_2+1} + \right. \\ &\quad \left. \mu_2 f_{j_1,j_2+1} + \frac{\sigma L}{\sqrt{2}} ((\mu_2 - \mu_1) z_1 - \mu_2 z_{1,j_1} - \mu_1 z_{1,j_1+1} + z_{2,j_2+1} - z_2)) \right) dz_2 dz_1 = \\ &\quad \sum_{k=1}^3 \bar{p}_k^\pm, \quad p = 1, \dots, m(m-1), \end{aligned} \quad (5.15)$$

where

$$\begin{aligned}\bar{p}_1^\pm &= (\bar{\xi}_{1,j_1} - z_{1,j_1})(z_{2,j_2+1}f_{j_1,j_2+1} - z_{2,j_2}f_{j_1,j_2} + \frac{\sigma L h_1}{\sqrt{2}}\bar{\xi}_{1,j_1}) + \\ &(z_{1,j_1+1} - \bar{\xi}_{1,j_1})(z_{2,j_2+1}f_{j_1+1,j_2+1} - z_{2,j_2}f_{j_1+1,j_2} + \frac{\sigma L h_1}{\sqrt{2}}\bar{\xi}_{1,j_1});\end{aligned}\quad (5.16)$$

$$\begin{aligned}\bar{p}_2^\pm &= (\bar{\xi}_{1,j_1} - \bar{\xi}_{1,j_1})((\bar{\xi}_{2,j_2} - z_{2,j_2})(\mu_1(f_{j_1,j_2} - \frac{\sigma L}{\sqrt{2}}z_{1,j_1}) + \mu_2(f_{j_1+1,j_2} + \frac{\sigma L}{\sqrt{2}}z_{1,j_1+1})) + \\ &(z_{2,j_2+1} - \bar{\xi}_{2,j_2})(\mu_1(f_{j_1+1,j_2+1} + \frac{\sigma L}{\sqrt{2}}z_{1,j_1+1}) + \mu_2(f_{j_1,j_2+1} - \frac{\sigma L}{\sqrt{2}}z_{1,j_1}));\end{aligned}\quad (5.17)$$

$$\bar{p}_3^\pm = \frac{\sigma L}{2\sqrt{2}}(4(\mu_1 - \mu_2)\bar{\xi}_{1,j_1}\bar{\xi}_{2,j_2} + z_{2,j_2+1}^2 + 2\bar{\xi}_{2,j_2}^2 - \frac{2}{3}(\bar{\xi}_{1,j_1}^2 + \bar{\xi}_{1,j_1}\bar{\xi}_{1,j_1} - 2\bar{\xi}_{1,j_1}^2)), \quad (5.18)$$

$j_1 = 1, \dots, j_2 - 1$ for $j_2 = 2, \dots, m$ and $j_1 = 1, \dots, 2m - j_2$ for $j_2 = m + 1, 2m - 1$. Setting $\sigma = 1$ and $h_1 = h/\sqrt{2}$ in (5.15)–(5.18), we obtain $\bar{R}_p^+, p = 1, \dots, m(m-1)$, and setting $\sigma = -1$ and $h_1 = h/\sqrt{2}$, we obtain $\bar{R}_p^-, p = 1, \dots, m(m-1)$. It is easy to see that values $\bar{R}_p^\pm, p = 1, \dots, m(m-1)$ can also be determined from formulae (5.15)–(5.18). Indeed, setting $\sigma = 1, h = -h_1/\sqrt{2}$ in (5.15)–(5.18), we obtain \bar{R}_p^+ , and setting $\sigma = -1, h = -h_1/\sqrt{2}$ we obtain \bar{R}_p^- , $p = 1, \dots, m(m-1)$.

Lemma 5.1 *Optimal-by-accuracy cubature formula for computing integral $I_1^2(f)$ in class $C_{1,L,N}^2$ has the form*

$$I_1^*(C_{1,L,N}^2) = \frac{1}{2} \left(\sum_{p=1}^{m(m-1)} (\bar{R}_p^* + \bar{R}_p^*) + \sum_{l=1}^{2m} (\bar{R}_l^* + \bar{R}_l^*) \right), \quad (5.19)$$

where

$$\begin{cases} \bar{R}_p^* = \frac{1}{2}(\bar{R}_p^+ + \bar{R}_p^-), \bar{R}_p^* = \frac{1}{2}(\bar{R}_p^+ + \bar{R}_p^-), & p = 1, \dots, m(m-1), \\ \bar{R}_l^* = \frac{1}{2}(\bar{R}_l^+ + \bar{R}_l^-), \bar{R}_l^* = \frac{1}{2}(\bar{R}_l^+ + \bar{R}_l^-), & l = 1, \dots, 2m, \end{cases} \quad (5.20)$$

and

$$\bar{\delta}(C_{1,L,N}^2) = \frac{1}{2} \left(\sum_{p=1}^{m(m-1)} (\bar{R}_p^+ + \bar{R}_p^+ - \bar{R}_p^- - \bar{R}_p^-) + \sum_{l=1}^{2m} (\bar{R}_l^+ - \bar{R}_l^+ - \bar{R}_l^- - \bar{R}_l^-) \right), \quad (5.21)$$

and the values $\bar{R}_p^\pm, \bar{R}_p^\pm, p = 1, \dots, m(m-1)$ and $\bar{R}_l^\pm, \bar{R}_l^\pm, l = 1, \dots, 2m$ are determined by the formulae (5.15)–(5.18).

Proof. Optimality-by-accuracy of the cubature formula (5.19) follows immediately from Theorem 4.1.

We also showed that \tilde{R}_p^\pm is determined by formulae (5.15)–(5.18). Taking into account (5.17) from (4.22), (4.12) it follows that computing $\tilde{R}_l^\pm, l = 1, \dots, 2m$ can also be performed using relationships (5.15)–(5.18) for $j_1 = j_2$ when $j_2 = 1, \dots, m$ and for $j_1 = 2m - j_2 + 1$ when $j_2 = m + 1, 2m$. Assuming that $h_1 = -h/\sqrt{2}$ from (5.15)–(5.18) we analogously obtain $\tilde{R}_p^\pm, p = 1, \dots, m(m-1)$ and $\tilde{R}_l^\pm, l = 1, \dots, 2m$.

6 Optimal-By-Order Cubature Formulae for Computing Integrals of Fast-Oscillatory Functions in Class $C_{1,L,N}^2$.

In this section the problem of computing integrals $I_2^2(f), I_3^2(f)$ is considered in the case when $\omega_1 = \omega_2 = \omega, |\omega| \geq 2\pi$. It is easy to see that in the transformation to the new system of coordinates (4.2) we have

$$I_2^2(f) = \int \int_{\pi_2} f(X) \sin \omega x_1 \sin \omega x_2 dX = \frac{1}{2} \left(\int \int_{\tilde{\pi}_2} \bar{f}(Z) \cos \tilde{\omega} z_1 dZ - \int \int_{\tilde{\pi}_2} \bar{f}(Z) \cos \tilde{\omega} z_2 dZ \right) \quad (6.1)$$

and

$$I_3^2(f) = \int \int_{\pi_2} f(X) \cos \omega x_1 \cos \omega x_2 dX = \frac{1}{2} \left(\int \int_{\tilde{\pi}_2} \bar{f}(Z) \cos \tilde{\omega} z_1 dZ + \int \int_{\tilde{\pi}_2} \bar{f}(Z) \cos \tilde{\omega} z_2 dZ \right), \quad (6.2)$$

where

$$\bar{f}(Z) = \left(\frac{1}{\sqrt{2}}(z_1 + z_2), \frac{1}{\sqrt{2}}(z_2 - z_1) \right), \quad \tilde{\omega} = \sqrt{2}\omega. \quad (6.3)$$

Let

$$S(\bar{f}) = \int \int_{\tilde{\pi}_2} \bar{f}(Z) \cos \tilde{\omega} z_1 dZ, \quad E(\bar{f}) = \int \int_{\tilde{\pi}_2} \bar{f}(Z) \cos \tilde{\omega} z_2 dZ. \quad (6.4)$$

We note that the problem of computing integrals (6.1), (6.2) reduces to the problem of computing integrals (6.3), (6.4).

For many classes of problems, for which we have to compute both integrals $I_2^2(f)$ and $I_3^2(f)$, this reduction allows us to substantially increase efficiency of computations. This advantage is especially important in those cases when computing $I_2^2(f), I_3^2$ have to be performed many times as it is the case in the solution of problems in recognition and classification of images.

Let us construct a cubature formula for computing the first integral in (6.3). Let us introduce the following notation:

$$\tilde{S}_p^\pm = \int \int_{\tilde{G}_p} \tilde{A}_{C_{1,L,N}^2}^\pm(Z) \cos \tilde{\omega} z_1 dZ, \quad \tilde{S}_p^\pm = \int \int_{\tilde{G}_p} \tilde{A}_{C_{1,L,N}^2}^\pm(Z) \cos \tilde{\omega} z_2 dZ, \quad (6.5)$$

with $p = 1, \dots, m(m-1)$ and

$$\tilde{S}_l^\pm = \frac{1}{2} \int \int_{\tilde{G}'_l} \tilde{A}_{C_{1,L,N}^2}^\pm(Z) \cos \tilde{\omega} z_1 dZ, \quad \tilde{S}_l^\pm = \frac{1}{2} \int \int_{\tilde{G}'_l} \tilde{A}_{C_{1,L,N}^2}^\pm(Z) \cos \tilde{\omega} z_2 dZ, \quad (6.6)$$

with $l = 1, \dots, 2m$.

Lemma 6.1 *The cubature formula*

$$\hat{S}(\bar{f}) = \sum_{p=1}^{m(m-1)} (\bar{S}'_p + \bar{\bar{S}}'_p) + \sum_{l=1}^{2m} (\bar{S}'_l + \bar{\bar{S}}'_l), \quad (6.7)$$

with

$$\begin{cases} \bar{S}'_p = \frac{1}{2}(\bar{S}_p^+ + \bar{S}_p^-), & \bar{\bar{S}}'_p = \frac{1}{2}(\bar{\bar{S}}_p^+ + \bar{\bar{S}}_p^-), \quad p = 1, \dots, m(m-1), \\ \bar{S}'_l = \frac{1}{2}(\bar{S}_l^+ + \bar{S}_l^-), & \bar{\bar{S}}'_l = \frac{1}{2}(\bar{\bar{S}}_l^+ + \bar{\bar{S}}_l^-), \quad p = 1, \dots, 2m \end{cases} \quad (6.8)$$

is optimal-by-order with a constant not exceeding 2. The error estimate of cubature formula (6.7) is determined from the following relationship

$$\begin{aligned} v(C_{1,L,N}^2, \hat{S}, \bar{f}) \leq \frac{1}{2} \left(\sum_{p=1}^{m(m-1)} \left(\max(\bar{S}_p^+, \bar{S}_p^-) + \max(\bar{\bar{S}}_p^+, \bar{\bar{S}}_p^-) - \right. \right. \\ \left. \min(\bar{S}_p^+, \bar{S}_p^-) - \min(\bar{\bar{S}}_p^+, \bar{\bar{S}}_p^-) \right) \sum_{l=1}^{2m} \left(\max(\bar{S}_l^+, \bar{S}_l^-) + \max(\bar{\bar{S}}_l^+, \bar{\bar{S}}_l^-) - \right. \\ \left. \left. \min(\bar{S}_l^+, \bar{S}_l^-) - \min(\bar{\bar{S}}_l^+, \bar{\bar{S}}_l^-) \right) \right). \end{aligned} \quad (6.9)$$

The statement of this lemma follows directly from Corollary 3.1. In analogy with relationships (5.15)–(5.18) it can be shown that

$$\bar{S}_p^\pm = \sum_{k=1}^3 P_k, \quad p = 1, \dots, m(m-1), \quad (6.10)$$

where

$$\begin{aligned} \bar{P}_1 = \frac{1}{\omega} \left((\sin \bar{\omega} \bar{\xi}_{1,j_1} - \sin \bar{\omega} z_{1,j_1}) \left(f_{j_1,j_2} (\bar{\bar{\xi}}_{2,j_2} - z_{2,j_2}) + f_{j_1,j_2+1} (z_{2,j_2+1} + \bar{\bar{\xi}}_{2,j_2}) + \frac{\sigma L}{\sqrt{2}} \times \right. \right. \\ \left. \left(\bar{\bar{\xi}}_{2,j_2}^2 - \bar{\bar{\xi}}_{2,j_2} (z_{2,j_2} + z_{2,j_2+1}) + \frac{1}{2}(z_{2,j_2}^2 + z_{2,j_2+1}^2) \right) \right) + (\sin \bar{\omega} z_{1,j_1+1} - \sin \bar{\omega} \bar{\xi}_{1,j_1}) \times \\ \left(f_{j_1+1,j_2+1} (z_{2,j_2} - \bar{\bar{\xi}}_{2,j_2}) + f_{j_1+1,j_2} (\bar{\bar{\xi}}_{2,j_2} - z_{2,j_2}) + \frac{\sigma L}{\sqrt{2}} (\bar{\bar{\xi}}_{2,j_2}^2 - \bar{\bar{\xi}}_{2,j_2} (z_{2,j_2} + z_{2,j_2+1}) + \right. \\ \left. \left. \frac{1}{2}(z_{2,j_2}^2 + z_{2,j_2+1}^2) \right) \right); \end{aligned} \quad (6.11)$$

$$\begin{aligned} \bar{P}_2 = \frac{1}{\omega} (\sin \bar{\omega} \bar{\xi}_{1,j_1} - \sin \omega \bar{\bar{\xi}}_{1,j_1}) \left(z_{2,j_2+1} \left(\mu_1 f_{j_1+1,j_2+1} + \mu_2 f_{j_1,j_2+1} + \frac{\sigma L}{\sqrt{2}} (\mu_1 z_{1,j_1+1} - \right. \right. \\ \left. \left. \mu_2 z_{1,j_1} - (\mu_1 - \mu_2) \bar{\bar{\xi}}_{1,j_1} - \bar{\bar{\xi}}_{2,j_2}) \right) - z_{2,j_2} \left(\mu_1 f_{j_1,j_2} + \mu_2 f_{j_1+1,j_2} + \frac{\sigma L}{\sqrt{2}} (\mu_1 z_{1,j_1+1} - \mu_2 z_{1,j_1} + \right. \\ \left. \left. (\mu_1 - \mu_2) \bar{\bar{\xi}}_{1,j_1} + \bar{\bar{\xi}}_{2,j_2}) \right) + \frac{\sigma L}{\sqrt{2}} \left(\frac{1}{2}(z_{2,j_2+1}^2 + z_{2,j_2}^2) + (\bar{\bar{\xi}}_{2,j_2} + (\mu_1 - \mu_2) \bar{\bar{\xi}}_{1,j_1})^2 \right) \right); \end{aligned} \quad (6.12)$$

$$\begin{aligned}
 \tilde{P}_3 = & \frac{1}{\omega} \left(\left(\mu_1(f_{j_1,j_2} - f_{j_1+1,j_2+1}) + \mu_2(f_{j_1+1,j_2} - f_{j_1,j_2+1}) + \frac{\sigma L}{\sqrt{2}}(\mu_2 - \mu_1)(z_{1,j_1+1} + z_{1,j_1}) \right) \right. \\
 & \times \left(\bar{\xi}_{2,j_2} \sin \bar{\omega} \bar{\xi}_{1,j_1} - \bar{\bar{\xi}}_{2,j_2} \sin \bar{\omega} \bar{\bar{\xi}}_{1,j_1} - \frac{1}{\omega}(\mu_1 - \mu_2)(\cos \bar{\omega} \bar{\xi}_{1,j_1} - \cos \bar{\omega} \bar{\bar{\xi}}_{1,j_1}) \right) + \frac{\sigma L h_1}{\sqrt{2}} \times \\
 & \left. \left(\sin \bar{\omega} \bar{\bar{\xi}}_{1,j_1} (\bar{\bar{\xi}}_{1,j_1} - z_{1,j_1}) + (\bar{\xi}_{1,j_1} - z_{1,j_1+1}) \sin \bar{\omega} \bar{\xi}_{1,j_1} + \frac{1}{\omega} (\cos \bar{\omega} \bar{\xi}_{1,j_1} - \cos \bar{\omega} z_{1,j_1} - \right. \right. \\
 & \left. \left. \cos \bar{\omega} z_{1,j_1+1} + \cos \bar{\omega} \bar{\xi}_{1,j_1}) \right) \right), \tag{6.13}
 \end{aligned}$$

and $h_1 = h/\sqrt{2}$, $\bar{\xi}_{1,j_1}, \bar{\bar{\xi}}_{1,j_1}, \bar{\xi}_{2,j_2}, \bar{\bar{\xi}}_{2,j_2}$, μ_1, μ_2 are determined from relationships (5.6)–(5.14) where $j_1 = 1, \dots, j_2 - 1$ for $j_2 = 2, \dots, m$ and $j_1 = 1, \dots, 2m - j_2$ for $j_2 = m + 1, \dots, 2m - 1$. If in (6.10)–(6.13) we set $\sigma = 1$ we obtain \tilde{S}_p^+ , and by setting $\sigma = -1$ we obtain $\tilde{S}_p^-, p = 1, \dots, m(m-1)$. Calculations of $\tilde{S}_l^\pm, l = 1, \dots, 2m$ are also performed with the help of relationships (6.10)–(6.13), for $j_1 = j_2$ when $j_2 = 1, \dots, m$ and for $j_1 = 2m - j_2 + 1$ when $j_2 = m + 1, \dots, 2m$. Setting $h_1 = -h/\sqrt{2}$ from (6.10)–(6.13) we obtain $\tilde{S}_p^\pm, p = 1, \dots, m(m-1)$ and $\tilde{S}_l^\pm, l = 1, \dots, 2m$.

Optimal-by-order cubature formulae for computing the second integral in (6.4) are constructed in an analogous way.

Let

$$\tilde{E}_p^\pm = \int \int_{\tilde{G}_p} \tilde{A}_{C_{1,L,N}^2}^\pm(Z) \cos \bar{\omega} z_2 dZ \quad \bar{\tilde{E}}_p^\pm = \int \int_{\tilde{G}_p} \tilde{A}_{C_{1,L,N}^2}^\pm(Z) \cos \bar{\omega} z_2 dZ, \tag{6.14}$$

for $p = 1, \dots, m(m-1)$ and

$$\tilde{E}_l^\pm = \frac{1}{2} \int \int_{\tilde{G}_l} \tilde{A}_{C_{1,L,N}^2}^\pm(Z) \cos \bar{\omega} z_2 dZ \quad \tilde{E}_l^\pm = \frac{1}{2} \int \int_{\bar{G}_l} \tilde{A}_{C_{1,L,N}^2}^\pm(Z) \cos \bar{\omega} z_2 dZ, \tag{6.15}$$

for $l = 1, \dots, 2m$.

Then the following result holds.

Lemma 6.2 *The cubature formula*

$$\hat{E}(f) = \sum_{p=1}^{m(m-1)} (\tilde{E}'_p + \bar{\tilde{E}}'_p) + \sum_{l=1}^{2m} (\tilde{E}'_l + \tilde{E}'_l), \tag{6.16}$$

with

$$\begin{cases} \tilde{E}'_p = \frac{1}{2}(\tilde{E}_p^+ + \tilde{E}_p^-), \quad \bar{\tilde{E}}'_p = \frac{1}{2}(\bar{\tilde{E}}_p^+ + \bar{\tilde{E}}_p^-), & p = 1, \dots, m(m-1), \\ \tilde{E}'_l = \frac{1}{2}(\tilde{E}_l^+ + \tilde{E}_l^-), \quad \bar{\tilde{E}}'_l = \frac{1}{2}(\bar{\tilde{E}}_l^+ + \bar{\tilde{E}}_l^-), & p = 1, \dots, 2m \end{cases} \tag{6.17}$$

is optimal-by-order with a constant not exceeding 2. The error estimate of cubature formula (6.16) is determined from the following relationship

$$\begin{aligned}
 v(C_{1,L,N}^2, \hat{E}, \tilde{f}) \leq & \frac{1}{2} \left(\sum_{p=1}^{m(m-1)} (|\tilde{E}_p^+ - \tilde{E}_p^-| + |\bar{\tilde{E}}_p^+ - \bar{\tilde{E}}_p^-|) + \sum_{l=1}^{2m} (|\tilde{E}_l^+ - \tilde{E}_l^-| + \right. \\
 & \left. |\bar{\tilde{E}}_l^+ - \bar{\tilde{E}}_l^-|) \right). \tag{6.18}
 \end{aligned}$$

The statement of the lemma follows directly from Corollary 4.1. In analogy with relationships (4.6)–(4.22) it can be shown that

$$\bar{E}_p^{\pm} = \sum_{k=1}^3 \bar{Q}_k, \quad p = 1, \dots, m(m-1), \quad (6.19)$$

where

$$\begin{aligned} \bar{Q}_1 = & \frac{1}{\bar{\omega}} \left((\bar{\xi}_{1,j_1} - z_{1,j_1}) (f_{j_1,j_2} (\sin \bar{\omega} \bar{\xi}_{2,j_2} - \sin \bar{\omega} z_{2,j_2}) + f_{j_1,j_2+1} (\sin \bar{\omega} z_{2,j_2+1} - \right. \\ & \left. \sin \bar{\omega} \bar{\xi}_{2,j_2})) + \frac{\sigma L}{\sqrt{2}} \left((2 \bar{\xi}_{2,j_2} - z_{2,j_2} - z_{2,j_2+1}) \sin \bar{\omega} \bar{\xi}_{2,j_2} + \frac{1}{\bar{\omega}} \left(2 \cos \bar{\omega} \bar{\xi}_{2,j_2} - \cos \bar{\omega} z_{2,j_2} - \right. \right. \\ & \left. \left. \cos \bar{\omega} z_{2,j_2+1}) \right) + (z_{1,j_1+1} - \bar{\xi}_{1,j_1}) (f_{j_1+1,j_2+1} (\sin \bar{\omega} z_{2,j_2+1} - \sin \bar{\omega} \bar{\xi}_{2,j_2}) + f_{j_1+1,j_2} \times \right. \\ & \left. (\sin \bar{\omega} \bar{\xi}_{2,j_2} - \sin \bar{\omega} z_{2,j_2}) + \frac{\sigma L}{\sqrt{2}} \left((2 \bar{\xi}_{2,j_2} - z_{2,j_2} - z_{2,j_2+1}) \sin \bar{\omega} \bar{\xi}_{2,j_2} + \frac{1}{\bar{\omega}} \left(2 \cos \bar{\omega} \bar{\xi}_{2,j_2} - \right. \right. \\ & \left. \left. \cos \bar{\omega} z_{2,j_2} - \cos \bar{\omega} z_{2,j_2+1}) \right) \right); \end{aligned} \quad (6.20)$$

$$\begin{aligned} \bar{Q}_2 = & \frac{1}{\bar{\omega}} (\bar{\xi}_{1,j_1} - \bar{\xi}_{1,j_1}) (\sin \bar{\omega} z_{2,j_2+1} (\mu_1 f_{j_1+1,j_2+1} + \mu_2 f_{j_1,j_2+1} + \frac{\sigma L}{\sqrt{2}} (\mu_1 (z_{1,j_1+1} - \right. \\ & \left. \frac{1}{2} (\bar{\xi}_{1,j_1} + \bar{\xi}_{1,j_1})) - \mu_2 (z_{1,j_1} - \frac{1}{2} (\bar{\xi}_{1,j_1} + \bar{\xi}_{1,j_1}))) \right) - \sin \bar{\omega} z_{2,j_2} (\mu_1 f_{j_1,j_2} + \mu_2 f_{j_1+1,j_2+1} - \\ & \frac{\sigma L}{\sqrt{2}} (\mu_1 (z_{1,j_1} - \frac{1}{2} (\bar{\xi}_{1,j_1} + \bar{\xi}_{1,j_1})) - \mu_2 (z_{1,j_1+1} - \frac{1}{2} (\bar{\xi}_{1,j_1} + \bar{\xi}_{1,j_1}))) \right) - \frac{\sigma L}{\sqrt{2} \bar{\omega}} \times \\ & (\cos \bar{\omega} z_{2,j_2+1} + \cos \bar{\omega} z_{2,j_2})); \end{aligned} \quad (6.21)$$

$$\begin{aligned} \bar{Q}_3 = & \frac{1}{\bar{\omega}} \left(\left(\mu_1 (f_{j_1,j_2} - f_{j_1+1,j_2+1}) + \mu_2 (f_{j_1+1,j_2} - f_{j_1,j_2+1}) + \frac{\sigma L}{\sqrt{2}} ((\mu_2 - \mu_1) \times \right. \right. \\ & (z_{1,j_1+1} + z_{1,j_1}) - z_{2,j_2+1} - z_{2,j_2} + 2(\bar{\xi}_{2,j_2} + (\mu_1 - \mu_2) \bar{\xi}_{1,j_1})) \right) ((\mu_1 - \mu_2) \times \\ & \cos \bar{\omega} \bar{\xi}_{2,j_2} \left(1 - \cos \bar{\omega} (\bar{\xi}_{1,j_1} - \bar{\xi}_{1,j_1}) \right) + \sin \bar{\omega} \bar{\xi}_{2,j_2} \sin \bar{\omega} (\bar{\xi}_{1,j_1} - \bar{\xi}_{1,j_1})) + \\ & \frac{\sigma L}{\sqrt{2}} \left((\sin \bar{\omega} z_{2,j_2+1} - \sin \bar{\omega} z_{2,j_2}) \left(\bar{\xi}_{1,j_1} (\bar{\xi}_{1,j_1} - z_{1,j_1}) + \bar{\xi}_{1,j_1} (\bar{\xi}_{1,j_1} - z_{1,j_1+1}) + \right. \right. \\ & \left. \frac{1}{2} (z_{1,j_1}^2 + z_{1,j_1+1}^2) \right) + \frac{2}{\bar{\omega}^2} \left(\cos \bar{\omega} \bar{\xi}_{2,j_2} \sin \bar{\omega} (\bar{\xi}_{1,j_1} - \bar{\xi}_{1,j_1}) - (\mu_1 - \mu_2) \sin \bar{\omega} \bar{\xi}_{2,j_2} \times \right. \\ & \left. (1 - \cos \bar{\omega} (\bar{\xi}_{1,j_1} - \bar{\xi}_{1,j_1})) \right) \right) \end{aligned} \quad (6.22)$$

and $h_1 = h/\sqrt{2}$ and $\bar{\xi}_{1,j_1}, \bar{\bar{\xi}}_{1,j_1}, \bar{\xi}_{2,j_2}, \bar{\bar{\xi}}_{2,j_2}, \mu_1, \mu_2$ are determined from relationships (5.6)–(5.14), $j_1 = 1, \dots, j_2 - 1$ for $j_2 = 2, \dots, m$ and $j_1 = 1, \dots, 2m - j_2$ for $j_2 = m + 1, \dots, 2m - 1$. If we set in (6.19)–(6.22) $\sigma = 1$ we obtain \bar{E}_p^+ , and by setting $\sigma = -1$ we obtain \bar{E}_p^- , $p = 1, \dots, m(m - 1)$. Computing \bar{E}_l^\pm , $l = 1, \dots, 2m$ is also performed with the help of relationships (6.19)–(6.22) for $j_1 = j_2$ when $j_2 = 1, \dots, m$ and for $j_1 = 2m - j_2 + 1$ when $j_2 = m + 1, \dots, 2m$. Setting $h_1 = -h/\sqrt{2}$, from (6.19)–(6.22) we obtain \bar{E}_p^\pm , $p = 1, \dots, m(m - 1)$ and \bar{E}_l^\pm , $l = 1, \dots, 2m$.

Therefore, in this paper we have constructively solved the problem of optimal-by-accuracy recovery of functions from the class $C_{1,L,N}^2$. We also constructed optimal-by-accuracy cubature formulae for these functions as well as optimal-by-order (with a constant not exceeding 2) cubature formulae for computing integrals from fast-oscillatory functions in the class $C_{1,L,N}^2$.

7 Acknowledgements

The authors are grateful to Prof. V. Zadiraka and Dr T. Sag, for fruitful discussions and the Australian Research Council for a partial support (Grant 179406). We also thank Michael Simpson for his helpful assistance at the final stage of preparation of this paper.

References

- [1] Alaylioglu, A., Numerical Evaluation of Finite Fourier Integrals, *J. Comput. Appl. Math.*, **9**, 1983, 305–313.
- [2] Berezovskii, A.I., Ivanov, V.V., On Optimal-By-Accuracy Uniform Spline Approximation, *S. Mathematics (Iz. VUZ)*, No. **10**, 1977, 14–24.
- [3] Berezovskii, A.I., Nechiporenko, N.E., Optimal Accuracy Approximation of Functions and Their Derivatives, *Journal of S. Mathematics*, **54**, 1991, 799–812.
- [4] Blahut, R.E., *Fast Algorithms for Digital Signal Processing*, Addison-Wesley, 1987.
- [5] Chan, S.C. and Ho, K.L., A New Two-Dimensional Fast Cosine Transform, *IEEE Trans. Signal Process.*, **39**, No. 2, 1991, 481–485.
- [6] Cools, R., Constructing Cubature Formulae: The Science Behind the Art, *Acta Numerica*, Cambridge University Press, 1997, 1–54.
- [7] Davis, P. and Rabinowitz, P., *Methods of Numerical Integration*, Academic Press, 1984.
- [8] Drachman, B. and Ross, J., Approximation of Certain Functions Given by Integrals with Highly Oscillatory Integrands, *IEEE Transactions on Antennas and Propagation*, **42**, No. 9, 1994, 1355–1356.
- [9] Einarson, B., Numerical Calculation of Fourier Integrals with Cubic Splines, *BIT*, **8**, No. 3, 1968, 279–286.
- [10] Ersoy, O.K., *Fourier-Related Transforms, Fast Algorithms, and Applications*, Prentice Hall, 1997.

- [11] Haider, Q. and Liu, L.C., Fourier and Bessel Transformations of Highly Oscillatory Functions, *J. Phys. A: Math. Gen.*, **25**, 1992, 6755–6760.
- [12] Hopkins, H.H., Numerical Evaluation of a Class of Double Integrals of Oscillatory Functions, *IMA J. of Numerical Analysis*, **9**, 1989, 61–80.
- [13] Ivanov, V.V., On Optimal Algorithms for Minimisation of Functions from Certain Classes, *Cybernetics*, No. 4, 1972, 81–94.
- [14] Levin, D., Fast Integration of Rapidly Oscillatory Functions, *J. Comput. Appl. Math.*, **67**, 1996, 95–101.
- [15] Melnik, K.N. and Melnik, R.V.N., On Computational Aspects of Certain Optimal Digital Signal Processing Algorithms, *Proc. of Computational Technique and Applications Conference: CTAC97*, Eds. J. Noye, M. Teubner and A. Gill, World Scientific, 1998, 433–440.
- [16] Melnik, K.N. and Melnik, R.V.N., Optimal-By-Order Quadrature Formulae for Fast Oscillatory Functions with Inaccurately Given A Priori Information, *Technical Report SC-MC-9810, Department of Mathematics and Computing, University of Southern Queensland, 1998*, submitted.
- [17] Melnik, K.N. and Melnik, R.V.N., Optimal-By-Accuracy and Optimal-By-Order Cubature Formulae in Interpolational Classes, *Technical Report SC-MC-9810, Department of Mathematics and Computing, University of Southern Queensland, 1998*, submitted.
- [18] Melnik, K.N. and Melnik, R.V.N., A Note on Optimal-By-Order Cubature Formulae for Fast Oscillatory Functions in Lipschitz Classes, *Technical Report SC-MC-9811, Department of Mathematics and Computing, University of Southern Queensland, 1998*, submitted.
- [19] Mysovskih, I.P., *Interpolatorische Kubaturformulen*, Institut für Geometrie und Praktische Matematik der RWTH Aachen, Bericht 74, 1992.
- [20] Ostapenko, O.S., On Optimal Algorithms for Minimisation of Functions in Classes $C_{1,L,N}^n$, $C_{1,L,N,e}^n$, $\tilde{C}_{1,L,N,\delta}^n$, *Cybernetics*, No. 5, 1983, 88–95.
- [21] Sucharev, A., *Minimax Models in the Theory of Numerical Methods*, Kluwer Academic Publishers, 1992.
- [22] Traub, J.F. and Wozniakowski, H., *A General Theory of Optimal Algorithms*, New York: Academic Press, 1980.
- [23] Zadiraka, V.K. and Kasenov, S. Z., Optimal-By-Accuracy Quadrature Formulae for Computing Fourier Transform of Finite Functions from $C_{L,N}$, *Ukr. Mathematical Journal*, **38**, No. 2, 1986, 233–237.
- [24] Zadiraka, V.K., Abatov, N.T., Optimally Exact Algorithms for Solutions of a Certain Numerical Integration Problem, *Ukr. Mathematical Journal*, **43**, 1991, 43–54.
- [25] Zheludev, V.A., Periodic Splines and the Fast Fourier Transform, *Computational Mathematics and Mathematical Physics*, **32**, No. 2, 1992, 149.
- [26] Zheludev, V.A., Processing of Periodic Signals Using Spline-Wavelets, *Radioelectronics and Communications Systems*, **38**, No. 3, 1995, 1.
- [27] Zhileikin, Ya.M. and Kukarkin, A.B., A Fast Fourier-Bessel Transformation Algorithm, *Computational Mathematics and Mathematical Physics*, **35**, No. 7, 1995, 901.

USQ



TOOWOOMBA

**OPTIMAL-BY-ACCURACY AND
OPTIMAL-BY-ORDER CUBATURE
FORMULAE IN INTERPOLATIONAL
CLASSES**

K.N.Melnik

Department of Computer Science
Flinders University, Adelaide

R.V.N Melnik

THE UNIVERSITY OF
S O U T H E R N
Q U E E N S L A N D

Working Papers

FACULTY
OF
SCIENCES

**OPTIMAL-BY-ACCURACY AND
OPTIMAL-BY-ORDER CUBATURE
FORMULAE IN INTERPOLATIONAL
CLASSES**

K.N.Melnik

Department of Computer Science
Flinders University, Adelaide

R.V.N Melnik

Department of Mathematics & Computing, USQ
Faculty of Sciences Working Paper Series
SC-MC-9817
11 June 1998

OPTIMAL-BY-ACCURACY AND OPTIMAL-BY-ORDER CUBATURE FORMULAE IN INTERPOLATIONAL CLASSES

K. N. Melnik *

Department of Computer Science,
Flinders University, Adelaide, SA 5001, Australia

R. V. N. Melnik †

Department of Mathematics and Computing,
University of Southern Queensland, QLD 4350, Australia

Abstract

In this paper we constructively solve the problem of optimal integration for fast oscillatory functions of two variables when *a priori* information is limited. We explore the connection of this problem with the problem of optimal recovery of a function from interpolational classes.

Key words: fast oscillatory functions, interpolational classes, Chebyshev centre, optimal-by-order cubature formulae.

1 Introduction.

In the solution of many classes of problems such as statistical processing of experimental data, boundary problems for PDEs, signal processing, modelling systems of automotive regulation and image recognition we often have to compute integrals of the form

$$I^2(f) = \int_0^1 \int_0^1 f(x_1, x_2) \varphi_1(x_1) \varphi_2(x_2) dx_1 dx_2, \quad (1.1)$$

where $\varphi_1(x_1)$, $\varphi_2(x_2)$ are known integrable functions and $f(x_1, x_2)$ belongs to a given functional class F_N . An important partial case of this problem is the computation of integrals

$$I_2^2(f) = \int_0^1 \int_0^1 f(x_1, x_2) \sin(\omega_1 x_1) \sin(\omega_2 x_2) dx_1 dx_2, \quad (1.2)$$

*Currently with the Electronic Data Systems, 60 Waymouth Street, Adelaide, 5000, Australia

†Corresponding author, E-mail: melnik@usq.edu.au

$$I_3^2(f) = \int_0^1 \int_0^1 f(x_1, x_2) \cos(\omega_1 x_1) \cos(\omega_2 x_2) dx_1 dx_2, \quad (1.3)$$

where ω_1, ω_2 are real numbers with $|\omega_i| \geq 2\pi$, $i = 1, 2$.

Integrands in (1.2), (1.3) are typical examples of rapidly oscillatory functions that occur in various applications, often in the context of the Fourier or Fourier-Bessel integral transforms. Integrating fast oscillatory functions is beset with difficulties even in the one-dimensional case (see, for example, [1, 7, 9, 11, 14, 15, 25]). Indeed, assume that we have to integrate the product $f(x) \exp(-i\omega x)$ on an interval (a, b) , where $\omega(b-a) \gg 1$. Since $\Re(f(x) \exp(-i\omega x))$ and $\Im(f(x) \exp(-i\omega x))$ have approximately $\omega(b-a)/\pi$ zeros on the interval (a, b) , even if $f(x)$ is a smooth function, we have to choose a polynomial of degree $n \gg \omega(b-a)/\pi$ in order to achieve an adequate level of approximation. It is well known that the use of such a high degree polynomial may lead to instability [12], a difficulty which is exacerbated in the two-dimensional case [13, 18, 6]. Moreover, since *a priori* information about the integrand is typically given inaccurately in the majority of practical problems, optimisation issues in numerical integration of fast oscillatory functions become of primary importance.

In this paper we construct optimal-by-accuracy and optimal-by-order cubature formulae for computing integrals (1.1)–(1.3) in interpolational classes $C_{2,L_1,L_2,N}$ and $C_{2,L,L,N}^2$. These classes are defined as follows

- $C_{2,L_1,L_2,N}^2$ is the class of functions defined in the domain π_2 , $\pi_2 = \{\mathbf{x} = (x_1, x_2) : 0 \leq x_i \leq 1, i = 1, 2\}$, satisfying the Lipschitz condition with constant L_1 and L_2 in each variable,

$$|f(\bar{x}_1, x_2) - f(\bar{\bar{x}}_1, x_2)| \leq L_1 |\bar{x}_1 - \bar{\bar{x}}_1|, \quad |f(x_1, \bar{x}_2) - f(x_1, \bar{\bar{x}}_2)| \leq L_2 |\bar{x}_2 - \bar{\bar{x}}_2|, \quad (1.4)$$

and taking fixed values $f(x_1) = f_1, \dots, f(x_N) = f_N$ at fixed nodes x_1, \dots, x_N respectively;

- $C_{2,L,L,N}^2$ is the class of functions defined in the domain π_2 , $\pi_2 = \{\mathbf{x} = (x_1, x_2) : 0 \leq x_i \leq 1, i = 1, 2\}$, satisfying the Lipschitz condition with constant L in both variables:

$$|f(\bar{x}_1, x_2) - f(\bar{\bar{x}}_1, x_2)| \leq L |\bar{x}_1 - \bar{\bar{x}}_1|, \quad |f(x_1, \bar{x}_2) - f(x_1, \bar{\bar{x}}_2)| \leq L |\bar{x}_2 - \bar{\bar{x}}_2|, \quad (1.5)$$

and taking fixed values $f(x_1) = f_1, \dots, f(x_N) = f_N$ at fixed nodes x_1, \dots, x_N respectively.

In what follows it is assumed that these functional classes are non-empty.

In order to obtain optimal-by-accuracy and optimal-by-order solutions of problems (1.1)–(1.3) we use the method of limit functions [22, 19, 15, 16]. The method consists of the definition of upper, $I^+(F_N)$, and lower, $I^-(F_N)$, limits of the set of possible values of the integral (1.1) (and, hence, (1.2), (1.3) as a special case) on functions from class F_N by the following formula

$$I^+(F_N) = \sup_{f \in F_N} I^2(f), \quad I^-(F_N) = \inf_{f \in F_N} I^2(f), \quad (1.6)$$

and the determination of the value

$$I^*(F_N) = \frac{I^+(F_N) + I^-(F_N)}{2}, \quad (1.7)$$

taken as the optimal-by-accuracy value of the integral $I^2(f)$. In this case $I^*(F_N)$ is the Chebyshev center of undefinability domain of values $I^2(f)$ on class F_N . The Chebyshev radius coincides with $\delta(F_N)$, defined as follows

$$\delta(F_N) = \frac{1}{2} (I^+(F_N) - I^-(F_N)). \quad (1.8)$$

In a special case, where $\varphi_1(x_1) = \varphi_2(x_2) = 1$, we come to the problem of computing the optimal-by-accuracy value $I_1^*(F_N)$ for integrals

$$I_1^2(f) = \int \int_{\pi_2} f(\mathbf{x}) d\mathbf{x} \quad (1.9)$$

with $f \in F_N$ and $X = (x_1, x_2)$.

It is known (see, for example, [3, 19] and references therein), that the problem of optimal-by-accuracy integration on class F_N is closely connected with the problem of optimal-by-accuracy recovery of $f(X) \in F_N$ at point $X = (x_1, x_2) \in \pi_2$.

Definition 1.1 Let F_N be a class of functions defined in a domain D . Then a function $A_{F_N}^+(X)$ ($A_{F_N}^-(X)$) is called a majorant (minorant) of the class F_N , if the conditions

- (a) $A_{F_N}^+(X) \geq f(X)$ ($A_{F_N}^-(X) \leq f(X)$) for all $f \in F_N$, $X = (x_1, x_2) \in D$ and
- (b) $A_{F_N}^+(X) \in F_N$ ($A_{F_N}^-(X) \in F_N$)

are satisfied.

The value of

$$f^*(X) = \frac{1}{2} (A_{F_N}^+(X) + A_{F_N}^-(X)) \quad (1.10)$$

(with $A_{F_N}^+(X)$, $A_{F_N}^-(X)$ majorant and minorant of class F_N respectively) is taken as the optimal-by-accuracy recovery of $f(X)$ at $X \in \pi_2$. Further in this paper we assume that $F_N = C_{2,L_1,L_2,N}^2$ or $F_N = C_{2,L,L,N}^2$. The error $\bar{\delta}(F_N, X)$ of the recovery of function $f(X) \in F_N$ at point X has the form

$$\bar{\delta}(F_N, X) = \frac{A_{F_N}^+(X) - A_{F_N}^-(X)}{2}. \quad (1.11)$$

Then, the optimal-by-accuracy cubature formulae for computing (1.9) is [22, 19, 17]

$$I_1^*(F_N) = \int \int_{\pi_2} f^*(X) dX \quad (1.12)$$

with the Chebyshev radius, $\bar{\delta}(F_N)$, of the domain of undefinability of integral (1.9) in the form

$$\bar{\delta}(F_N) = \int \int_{\pi_2} \bar{\delta}(F_N, X) dX. \quad (1.13)$$

For a constructive solution of problems (1.10)–(1.11) and (1.12)–(1.13), as well as for the construction of efficient cubature formulae for computing integrals (1.1)–(1.3) we have to consider properties of majorants and minorants of the functional classes that are investigated.

2 Properties of Majorants and Minorants of Interpolational Classes $C_{2,L_1,L_2,N}^2$ and $C_{2,L,L,N}^2$.

From Definition 1.1 it follows that if there exists such a function $g_1(X)$ (or $g_2(X)$) with $g_1(X) = \max_{f \in F_N} f(X)$ (or $g_2(X) = \min_{f \in F_N} f(X)$) from class F_N , then it coincides with the majorant (or minorant) of the class of functions that is investigated. We also note that

$$A_{F_N}^+(X) = \sup_{f \in F_N} f(X) = \min_{v=1,\dots,N} (f_v + L\|X - X_v\|), \quad (2.1)$$

$$A_{F_N}^-(X) = \inf_{f \in F_N} f(X) = \max_{v=1,\dots,N} (f_v - L\|X - X_v\|), \quad (2.2)$$

where $X = (x_1, x_2)$, $D = \pi_2$. In the case $F_N = C_{2,L_1,L_2,N}^2$ we define

$$\|X\| = \|X\|_2 = |x_1| + \frac{L_2}{L_1}|x_2|, \quad (2.3)$$

and for the case $F_N = C_{2,L,L,N}^2$

$$\|X\| = \|X\|_2 = |x_1| + |x_2|. \quad (2.4)$$

On the example of class $F_N = C_{2,L_1,L_2,N}^2$ we will show that functions $A_{F_N}^+(X)$ and $A_{F_N}^-(X)$ defined by (2.1) and (2.2) indeed satisfy Definition 1.1.

First, let us show that function $A_{C_{2,L_1,L_2,N}^2}^+(X)$ satisfies condition (a). For any $f(X) \in C_{2,L_1,L_2,N}^2$ we have

$$\begin{aligned} f(X) - A_{C_{2,L_1,L_2,N}^2}^+(X) &= f(X) - \min_{v=1,\dots,N} (f_v + L_1\|X - X_v\|_2) = \\ f(X) - f_{v_0} - L_1\|X - X_{v_0}\|_2 &\leq 0. \end{aligned} \quad (2.5)$$

The last inequality in (2.5) follows from the fact that inequalities (1.4) and the relationship

$$|f(\bar{X}) - f(\tilde{X})| \leq L_1\|\bar{X} - \tilde{X}\|_2, \quad \bar{X} = (x_1, x_2), \quad \tilde{X} = (\bar{x}_1, \bar{x}_2) \quad (2.6)$$

are equivalent. Indeed,

$$\begin{aligned} |f(\bar{X}) - f(\tilde{X})| &= |f(\bar{x}_1, \bar{x}_2) - f(\bar{x}_1, \tilde{x}_2)| = |f(\bar{x}_1, \bar{x}_2) - f(\bar{x}_1, \tilde{x}_2) + f(\bar{x}_1, \tilde{x}_2) - \\ &f(\bar{x}_1, \tilde{x}_2)| \leq L_1|\bar{x}_1 - \tilde{x}_1| + L_2|\bar{x}_2 - \tilde{x}_2| = L_1\|\bar{X} - \tilde{X}\|_2. \end{aligned} \quad (2.7)$$

In other words, from the definition of class $C_{2,L_1,L_2,N}^2$ and the relationship (2.1) inequality (2.6) follows immediately. The inverse statement is also true. Let $\bar{X} = (\bar{x}_1, \bar{x}_2)$, $\tilde{X} = (\tilde{x}_1, \tilde{x}_2)$, $\bar{\tilde{X}} = (\bar{\tilde{x}}_1, \bar{\tilde{x}}_2)$. Then

$$|f(\bar{x}_1, \bar{x}_2) - f(\bar{x}_1, \tilde{x}_2)| \leq L_1\|\bar{X} - \tilde{X}\|_2 = L_1 \frac{L_2}{L_1} |\bar{x}_2 - \tilde{x}_2| = L_2 |\bar{x}_2 - \tilde{x}_2|, \quad (2.8)$$

and

$$|f(\bar{x}_1, \tilde{x}_2) - f(\bar{\tilde{x}}_1, \tilde{x}_2)| \leq L_1\|\bar{X} - \bar{\tilde{X}}\|_2 = L_1 |\bar{x}_1 - \bar{\tilde{x}}_1|. \quad (2.9)$$

Now let us show that function $A_{C_{2,L_1,L_2,N}}^+(X)$ belongs to class $C_{2,L_1,L_2,N}^2$, i.e. satisfies (1.4) or (2.6) and $A_{C_{2,L_1,L_2,N}}^+(X_v) = f_v$, $v = 1, \dots, N$. Indeed,

$$\begin{aligned} A_{C_{2,L_1,L_2,N}}^+(\bar{X}) - A_{C_{2,L_1,L_2,N}}^+(\tilde{\bar{X}}) &= f_{v_0} + L_1 \|X_{v_0} - \bar{X}\|_2 - f_{v_1} - L_1 \|X_{v_1} - \bar{X}\|_2 \leq \\ f_{v_1} + L_1 \|X_{v_1} - \bar{X}\|_2 - f_{v_1} - L_1 \|X_{v_1} - \tilde{\bar{X}}\|_2 &= L_1 \|X_{v_1} - \bar{X}\|_2 - L_1 \|X_{v_1} - \tilde{\bar{X}}\|_2 \leq \\ L_1 \|\bar{X} - \tilde{\bar{X}}\|_2. \end{aligned} \quad (2.10)$$

The second last inequality in (2.10) follows from

$$A_{C_{2,L_1,L_2,N}}^+(\bar{X}) = \min_{v=1,\dots,N} (f_v + L_1 \|X_v - \bar{X}\|_2). \quad (2.11)$$

Therefore,

$$A_{C_{2,L_1,L_2,N}}^+(\bar{X}) \leq f_{v_1} + L_1 \|X_{v_1} - \bar{X}\|_2. \quad (2.12)$$

The fact that $A_{C_{2,L_1,L_2,N}}^+(X_v) = f_v$, $v = 1, \dots, N$ follows from non-emptiness of class $C_{2,L_1,L_2,N}^2$ and relationship (1.4). Hence, we have shown that function

$$A_{C_{2,L_1,L_2,N}}^+(X) = \min_{v=1,\dots,N} (f_v + L_1 \|X - X_v\|_2) \quad (2.13)$$

is a majorant of class $C_{2,L_1,L_2,N}^2$. In a similar way, it can be shown that

$$A_{C_{2,L_1,L_2,N}}^-(X) = \max_{v=1,\dots,N} (f_v - L_1 \|X - X_v\|_2). \quad (2.14)$$

Let us define more precisely our set of nodes X_1, X_2, \dots, X_N .

- For class $C_{2,L,L,N}^2$ we denote this set as $\Delta_1 = \{X_s\}_{s=1,\dots,N}$, $N = (m+1)^2$. We assume that Δ_1 has the following structure

$$\begin{cases} X_s = (x_{1,i}; x_{2,j}), s = (i-1)(m+1) + j, \\ x_{1,i} = (i-1)\frac{1}{m}, x_{2,j} = (j-1)\frac{1}{m}, i, j = 1, m+1. \end{cases} \quad (2.15)$$

The grid Δ_1 splits the domain π_2 into m^2 equal squares K_p , $p = 1, \dots, m^2$ with sides $1/m$. We will call such squares elementary.

- For class $C_{2,L_1,L_2,N}^2$ we denote the set of nodes X_1, X_2, \dots, X_N by Δ_2 , where

$$\begin{cases} \Delta_2 = \{X_S\}_{S=1,\dots,N}, N = (m+1)(m_1+1), X_S = (x_{1,i}; x_{2,j}), \\ x_{1,i} = (i-1)\frac{1}{m}, x_{2,j} = (j-1)\frac{1}{m_1}, \\ s = (i-1)(m_1+1) + j, i = 1, \dots, m+1, j = 1, \dots, m_1+1, m_1 = \left[m \frac{L_1}{L_2}\right]. \end{cases} \quad (2.16)$$

In this case, grid Δ_2 splits square π_2 into mm_1 equal rectangles \bar{K}_p , $p = 1, \dots, mm_1$ with sides $1/m$, $L_2/(mL_1)$, and m equal rectangles \tilde{K}_p , $p = 1, \dots, m$ with sides $1/m$, $(1 - m_1 L_2 / (m L_1))$. Such rectangles will also be referred to as elementary.

Further, we introduce the following notation

1. $h = h_1 = 1/m$, $h_2 = L_2/(mL_1)$, $\bar{h}_2 = 1 - m_1 L_2/(mL_1)$;
2. $\sigma(K_p) = \{s_i\}_{i=1,\dots,4}$, where $s_1 = (i_1 - 1)(m + 1) + j_1$, $s_2 = (i_1 - 1)(m + 1) + j_1 + 1$, $s_3 = i_1(m + 1) + j_1 + 1$, $s_4 = i_1(m + 1) + j_1$ are numbers of nodes $X_{s_1}, X_{s_2}, X_{s_3}, X_{s_4}$, which correspond to vertices of the elementary square K_p , $p = 1, \dots, m^2$, $p = (i_1 - 1)m + j_1$, $i_1 = 1, \dots, m$, $j_1 = 1, \dots, m$;
3. $\sigma(\tilde{K}_p) = \{s_i\}_{i=1,\dots,4}$, where s_1, s_2, s_3, s_4 are defined as before, i.e. they give the numbers of nodes $X_{s_1}, X_{s_2}, X_{s_3}, X_{s_4}$, which correspond to vertices of the elementary rectangle \tilde{K}_p , $p = (i_1 - 1)m + j_1$, $i_1 = 1, \dots, m$, $j_1 = 1, \dots, m_1$;
4. $\sigma(\tilde{K}_p) = \{s_i\}_{i=1,2}$, where $s_1 = (p - 1)(m_1 + 1) + m_1 + 1$, $s_2 = p(m_1 + 1) + m_1 + 1$ are numbers of nodes X_{s_1}, X_{s_2} that correspond to two-out-of-four vertices of the elementary rectangle \tilde{K}_p , $p = 1, \dots, m$;
5. finally, let $f_{s_i} = f(X_{s_i})$, $i = 1, 2, 3, 4$.

Theorem 2.1 Let $F_N = C_{2,L_1,L_2,N}^2$ or $C_{2,L,L,N}^2$. Then in elementary region $K \in \pi_2$, the majorant and minorant of class F_N has the form

$$A_{F_N}^+(X) = \min_{s \in \sigma(K)} (f_s + L\|X - X_s\|), \quad A_{F_N}^-(X) = \max_{s \in \sigma(K)} (f_s - L\|X - X_s\|), \quad X \in K, \quad (2.17)$$

where for $F_N = C_{2,L_1,L_2,N}^2$ elementary region K is either \tilde{K}_p , $p = (i_1 - 1)m_1 + j_1$, $i_1 = 1, \dots, m$, $j_1 = 1, \dots, m_1$ or \tilde{K}_p , $p = 1, \dots, m$; and for $F_N = C_{2,L,L,N}^2$ elementary region K coincides with K_p , $p = 1, \dots, m^2$.

Proof. We will prove the theorem for the majorant $A_{F_N}^+(X)$. For the minorant $A_{F_N}^-(X)$ the proof is analogous.

Let $F_N = C_{2,L_1,L_2,N}^2$, $K = \tilde{K}_{\tilde{p}}$, $\tilde{p} = (\tilde{i} - 1)m_1 + \tilde{j}$, where \tilde{i} , \tilde{j} are certain fixed values of indices i_1 and j_1 . In this case $X_{s_1} = (x_{1,\tilde{i}}; x_{2,\tilde{j}})$, $X_{s_2} = (x_{1,\tilde{i}}; x_{2,\tilde{j}+1})$, $X_{s_3} = (x_{1,\tilde{i}+1}; x_{2,\tilde{j}+1})$, $X_{s_4} = (x_{1,\tilde{i}+1}; x_{2,\tilde{j}})$. Let us consider the following nodes of the grid

- nodes $X_{v_1} = (x_{1,\tilde{i}-k_1}; x_{2,\tilde{j}-k_2})$; $k_1 = 1, \dots, \tilde{i} - 1$, $k_2 = 1, \dots, \tilde{j} - 1$;
- nodes $X_{v_2} = (x_{1,\tilde{i}-k_1}; x_{2,\tilde{j}+k_2})$; $k_1 = 1, \dots, \tilde{i} - 1$, $k_2 = 2, \dots, m_1 + 1$;
- nodes $X_{v_3} = (x_{1,\tilde{i}+k_1}; x_{2,\tilde{j}+k_2})$; $k_1 = 2, \dots, m + 1 - \tilde{i}$, $k_2 = 2, \dots, m_1 + 1 - \tilde{j}$;
- nodes $X_{v_4} = (x_{1,\tilde{i}+k_1}; x_{2,\tilde{j}-k_2})$; $k_1 = 2, \dots, m + 1 - \tilde{i}$, $k_2 = 1, \dots, \tilde{j} - 1$.

We introduce the following functions

$$g_v(X) = f_v + L_1\|X - X_v\|_2, \quad v \in \sigma(\Delta_2). \quad (2.18)$$

It is easy to show that functions defined by (2.18) have the following property

$$g_{v_l}(X) \geq g_{s_l}(X) \quad \forall X \in \tilde{K}_{\tilde{p}}, \quad v_l \in \sigma(\Delta_2) \setminus \sigma(\tilde{K}_{\tilde{p}}), \quad l = 1, 2, 3, 4. \quad (2.19)$$

Indeed,

$$\begin{aligned} g_{v_1}(X) - g_{s_1}(X) &= f_{v_1} + L_1\|X - X_{v_1}\|_2 - f_{s_1} - L_1\|X - X_{s_1}\|_2 \geq -L_1 k_1 h_1 - L_2 k_2 h_2 + \\ &L_1(x_1 - (\tilde{i} - k_1)h_1) + L_2(x_2 - (\tilde{j} - k_2)h_2) - L_1(x_1 - \tilde{i}h_1) - L_2(x_2 - \tilde{j}h_2) = 0. \end{aligned} \quad (2.20)$$

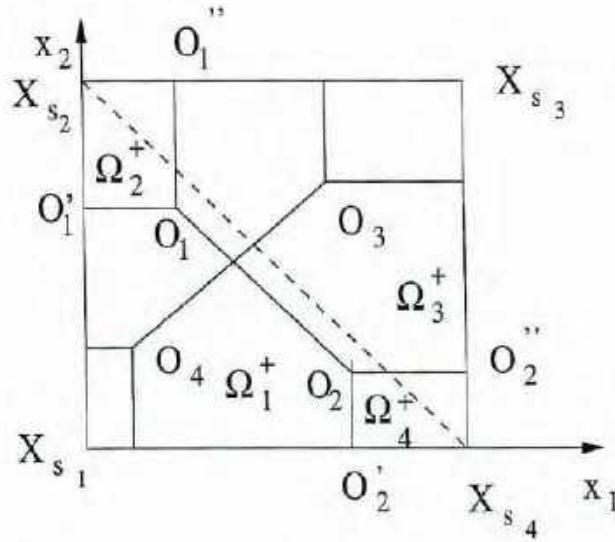


Figure 1: Splitting of elementary square K_p .

In a similar way it can be shown that $g_{v_l}(X) - g_{s_l}(X) \geq 0 \forall X \in \bar{K}_p$ and for $l = 2, 3, 4$. This means that for all functions $g_v(X)$, $v \in \sigma(\Delta_2) \setminus \sigma(\bar{K}_p)$ the inequality

$$g_v(X) \geq \min_{s \in \sigma(\bar{K}_p)} g_s(X) \quad (2.21)$$

holds $\forall X \in \bar{K}_p$. Analogously we reason for $K = \bar{K}_p$, $p = 1, \dots, m$. In this case we have to consider nodes X_{v_1} and X_{v_4} .

Therefore, we have proved that for an arbitrary function $g_v(X)$, $v \in \sigma(\Delta_2) \setminus \sigma(K)$ there exists function $g_s(X)$, $s \in \sigma(K)$ such that $g_v(X) \geq g_s(X)$, which confirms the statement of theorem 2.1. The case $F_N = C_{2,L,L,N}^2$ can be considered similarly. ■

Theorem 2.1 has several important consequences. Let us consider the class $C_{2,L,L,N}^2$. We place the origin of the plane (x_1, x_2) in the left lower vertex of elementary square K_p , i.e. at the node X_{s_1} . Then we split elementary square K_p into parts Ω_1^+ , Ω_2^+ , Ω_3^+ and Ω_4^+ as shown in Fig.1. The equations of five lines that split K_p into Ω_l^+ , $l = 1, 2, 3, 4$ have the following forms

$$\begin{cases} g_{s_1}(X) = g_{s_3}(X), f_{s_1} + L\|X - X_{s_1}\|_2 = f_{s_3} + L\|X - X_{s_3}\|_2, \\ L(x_1 + x_2) - L(h - x_1 + h - x_2) = f_{s_3} - f_{s_1}, x_2 = -x_1 + \frac{f_{s_3} - f_{s_1}}{2L} + h \end{cases} \quad (2.22)$$

for the line through O_1 , O_2 ;

$$g_{s_2}(X) = g_{s_3}(X), x_1 = \frac{f_{s_3} - f_{s_2}}{2L} + \frac{h}{2} \quad (2.23)$$

for O_1 , O_1'' :

$$g_{s_1}(X) = g_{s_2}(X), x_2 = \frac{f_{s_3} - f_{s_1}}{2L} + \frac{h}{2} \quad (2.24)$$

for O_1, O'_1 ;

$$g_{s_1}(X) = g_{s_4}(X), \quad x_1 = \frac{f_{s_4} - f_{s_1}}{2L} + \frac{h}{2} \quad (2.25)$$

for O'_2, O_2 ;

$$g_{s_4}(X) = g_{s_3}(X), \quad x_2 = \frac{f_{s_3} - f_{s_4}}{2L} + \frac{h}{2} \quad (2.26)$$

for O_2, O''_2 . Note that in Fig. 1 we present the case where $f_{s_2} + f_{s_4} \geq f_{s_3} + f_{s_1}$ and $f_{s_1} > f_{s_3}$.

Corollary 2.1 *The majorant of the class $C_{2,L,L,N}^2$ for $X \in K_p$ has the form*

$$A_{C_{2,L,L,N}^2}^+(X) = \bar{g}_{s_1}(X), \quad X \in \Omega_l^+, \quad l = 1, 2, 3, 4, \quad (2.27)$$

where $\bar{g}_{s_1}(X) = f_{s_1} + L\|X - X_{s_1}\|_2$, $\bigcup_{l=1}^4 \Omega_l^+ = K_p$, $p = 1, \dots, m^2$.

Proof. Let $X \in \Omega_1^+$. It is easy to show that

$$\bar{g}_{s_1}(X) \leq \bar{g}_{s_i}(X), \quad i = 2, 3, 4 \quad \forall X \in \Omega_1^+. \quad (2.28)$$

Indeed, let us prove this inequality, for example, for $i = 3$. We have

$$\begin{aligned} f_{s_1} + L\|X - X_{s_1}\|_2 - f_{s_3} - L\|X - X_{s_3}\|_2 &= f_{s_1} - f_{s_3} + L(x_1 + x_2) - \\ L(h - x_1 + h - x_2) &= f_{s_1} - f_{s_3} + 2L(x_1 + x_2 - h) \leq \\ f_{s_1} - f_{s_3} + (f_{s_3} - f_{s_1} + 2Lh) - 2Lh &= 0. \end{aligned} \quad (2.29)$$

The last inequality in (2.29) follows from the fact that in domain Ω_1^+ we have

$$x_2 \leq -x_1 + \frac{f_{s_3} - f_{s_1}}{2L} + h. \quad (2.30)$$

An analogous statement can also be formulated for the majorant of class $C_{2,L_1,L_2,N}^2$ when $X \in \tilde{K}_p$. In this case (which is similar to that in Fig.1) we have

$$X_{s_1} = (0; 0), \quad X_{s_2} = (0; h_2), \quad X_{s_3} = (h_1; h_2), \quad X_{s_4} = (h_1; 0). \quad (2.31)$$

Elementary rectangle \tilde{K}_p is split into sub-regions $\tilde{\Omega}_1^+, \tilde{\Omega}_2^+, \tilde{\Omega}_3^+, \tilde{\Omega}_4^+$ by the following five lines

$$x_2 = -\frac{L_1}{L_2}x_1 + \frac{f_{s_3} - f_{s_1}}{2L_2} + \frac{L_1h_1}{2L_2} + \frac{h_2}{2} \quad (2.32)$$

for the line through O_1, O_2 ;

$$x_1 = \frac{f_{s_3} - f_{s_2}}{2L_1} + \frac{h_1}{2} \quad (2.33)$$

for O_1, O''_1 ;

$$x_2 = \frac{f_{s_2} - f_{s_1}}{2L_2} + \frac{h_2}{2} \quad (2.34)$$

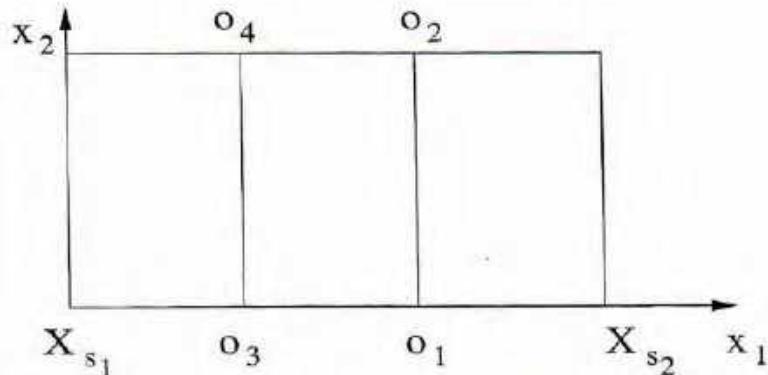


Figure 2: Splitting of elementary rectangle \tilde{K}_p .

for O'_1, O_1 :

$$x_1 = \frac{f_{s_4} - f_{s_1}}{2L_1} + \frac{h_1}{2} \quad (2.35)$$

for O'_2, O_2 :

$$x_2 = \frac{f_{s_3} - f_{s_4}}{2L_2} + \frac{h_2}{2} \quad (2.36)$$

for O_2, O''_2 .

Corollary 2.2 *The majorant of the class $C_{2,L_1,L_2,N}^2$, $X \in \tilde{K}_p$ has the form*

$$A_{C_{2,L_1,L_2,N}^2}^+(X) = \bar{g}_{s_l}(X), \quad X \in \tilde{\Omega}_l^+, \quad l = 1, 2, 3, 4, \quad (2.37)$$

where $\bar{g}_{s_l}(X) = f_{s_l} + L_1 \|X - X_{s_l}\|_2$, $\bigcup_{l=1}^4 \tilde{\Omega}_l^+ = \tilde{K}_p$, $p = 1, \dots, mm_1$.

Proof is analogous to the proof of Corollary 2.1. ■

Now, let us consider elementary rectangle \tilde{K}_p (see Fig. 2). As before, we place the origin in the lower left vertex of \tilde{K}_p . The line (O_1, O_2) , whose equation is

$$x_1 = \frac{f_{s_2} - f_{s_1}}{2L_1} + \frac{h_1}{2}, \quad (2.38)$$

splits elementary rectangle \tilde{K}_p into sub-regions $\tilde{\Omega}_1^+$ and $\tilde{\Omega}_2^+$. Therefore, the statement analogous to Corollary 2.2 also holds.

Corollary 2.3 *The majorant of the class $C_{2,L_1,L_2,N}^2$ for $X \in \tilde{K}_p$ has the form*

$$A_{C_{2,L_1,L_2,N}^2}^+(X) = \bar{g}_{s_l}(X), \quad X \in \tilde{\Omega}_l^+, \quad l = 1, 2, \quad (2.39)$$

where $\bar{g}_{s_l}(X) = f_{s_l} + L_1 \|X - X_{s_l}\|_2$, $\tilde{\Omega}_1^+ \cup \tilde{\Omega}_2^+ = \tilde{K}_p$, $p = 1, \dots, m$.

Therefore Corollaries 2.1 – 2.3 allow us to single out regions of linearity of the majorant $A_{F_N}^+(X)$ and the minorant $A_{F_N}^-(X)$ for classes $C_{2,L_1,L_2,N}^2$ and $C_{2,L,L,N}^2$, represent them in π_2 and, finally, solve constructively the problem of optimal-by-accuracy recovery of function $f(X)$ from these classes at point $X \in \pi_2$.

3 On Optimal Integration of Function Products in Classes $C_{2,L_1,L_2,N}^2$ and $C_{2,L,L,N}^2$.

In this section we deal with the general integral (1.1) assuming that $f(x_1, x_2) \in C_{2,L_1,L_2,N}^2$ or $f(x_1, x_2) \in C_{2,L,L,N}^2$ and that $\varphi_1(x_1), \varphi_2(x_2)$ are given integrable functions. We introduce the following notation

$$\begin{cases} \tilde{I}_p^* = \frac{1}{2} \int \int_{K_p} \left(A_{C_{2,L,L,N}^2}^+(X) + A_{C_{2,L,L,N}^2}^-(X) \right) \varphi_1(x_1) \varphi_2(x_2) dX, p = 1, \dots, m^2, \\ \tilde{I}_p^* = \frac{1}{2} \int \int_{\tilde{K}_p} \left(A_{C_{2,L_1,L_2,N}^2}^+(X) + A_{C_{2,L_1,L_2,N}^2}^-(X) \right) \varphi_1(x_1) \varphi_2(x_2) dX, p = 1, \dots, mm_1, \\ \tilde{I}_p^* = \frac{1}{2} \int \int_{\tilde{K}_p} \left(A_{C_{2,L_1,L_2,N}^2}^+(X) + A_{C_{2,L_1,L_2,N}^2}^-(X) \right) \varphi_1(x_1) \varphi_2(x_2) dX, p = 1, \dots, m. \end{cases} \quad (3.1)$$

We recall that in Section 2 we obtained explicit forms of functions $A_{F_N}^+(X), A_{F_N}^-(X)$ for $F_N = C_{2,L,L,N}^2$ and $F_N = C_{2,L_1,L_2,N}^2$. Each of the domains $K_p, \tilde{K}_p, \tilde{K}_p$ was split into sub-regions in which $A_{F_N}^+(X)$ and $A_{F_N}^-(X)$ were linear functions. Below we show how the results obtained in Section 2 can be applied to the constructive solution of the problem of computing integrals in (3.1).

Theorem 3.1 *If functions $\varphi_1(x_1)$ and $\varphi_2(x_2)$ do not change sign for $x_1, x_2 \in [0, 1]$, then optimal-by-accuracy cubature formulae for computing integrals (1.1) have the form*

$$\bar{I}^* = \sum_{p=1}^{m^2} \tilde{I}_p^* \text{ when } F_N = C_{2,L,L,N}^2 \quad (3.2)$$

and the form

$$\bar{I}^* = \sum_{p=1}^{mm_1} \tilde{I}_p^* + \sum_{l=1}^m \tilde{I}_l^* \text{ when } F_N = C_{2,L_1,L_2,N}^2. \quad (3.3)$$

The Chebyshev radius of the undefinability domain of integral values is defined by the formula

$$\delta(F_N) = \frac{1}{2} \int \int_{\pi_2} \left(A_{F_N}^+(X) - A_{F_N}^-(X) \right) |\varphi_1(x_1) \varphi_2(x_2)| dX. \quad (3.4)$$

Proof. If $\varphi_1(x_1) \varphi_2(x_2) > 0$, then $\forall X \in \pi_2, X = (x_1, x_2)$ we have

$$I^\pm(F_N) = \int \int_{\pi_2} A_{F_N}^\pm(X) \varphi_1(x_1) \varphi_2(x_2) dX. \quad (3.5)$$

Similarly, if $\varphi_1(x_1) \varphi_2(x_2) < 0$, then

$$I^\pm(F_N) = \int \int_{\pi_2} A_{F_N}^\mp(X) \varphi_1(x_1) \varphi_2(x_2) dX. \quad (3.6)$$

The statement of the theorem follows from the relationships (3.4) and (3.5) by taking into account (3.1), (1.6) and (1.7). ■

Remark 3.1 Let the domain π_2 be split into sub-regions $\Phi_q, q = 1, \dots, Q$ where the product of functions $\varphi_1(x_1)$ and $\varphi_2(x_2)$ preserves the sign. Then in each sub-region Φ_q cubature formulae (3.2), (3.3) are optimal-by-accuracy. However, in all domain π_2 Theorem 3.1 does not hold in general.

When Theorem 3.1 fails, the majorant of the class F_N is different from $A_{F_N}^+(X)$. Let, for example, $F_N = C_{2,L,L,N}^2$. We consider two neighbouring elementary squares K_p and K_{p+1} in which the sign of the product $\varphi_1(x_1)\varphi_2(x_2)$ changes from “+” to “-”.

In the transition from the region K_p to region K_{p+1} the function

$$\gamma^+(X) = \begin{cases} A_{F_N}^+(X), & X \in K_p, \\ A_{F_N}^-(X), & X \in K_{p+1} \end{cases} \quad (3.7)$$

has a discontinuity, hence the Lipschitz condition is violated and $\gamma^+(X) \notin F_N$.

Let us choose $\gamma^+(X)$ in the following form

$$\gamma^+(X) = \begin{cases} \min(A_{F_N}^+(X), l(X)), & X \in K_p, \\ \max(A_{F_N}^-(X), l(X)), & X \in K_{p+1}, \end{cases} \quad (3.8)$$

where function $l(X)$ performs “sewing” $A_{F_N}^+(X)$ and $A_{F_N}^-(X)$ in the transition from K_p to K_{p+1} .

In this case $\gamma^+(X)$ satisfies Definition 1.1 for the majorant of class F_N and the following relationship

$$\int \int_{K_p \cup K_{p+1}} \gamma^+(X) \varphi_1(x_1) \varphi_2(x_2) dX = \sup_{f \in F_N} \int \int_{K_p} f(X) \varphi_1(x_1) \varphi_2(x_2) dX + \\ \inf_{f \in F_N} \int \int_{K_{p+1}} f(X) \varphi_1(x_1) \varphi_2(x_2) dX. \quad (3.9)$$

takes place. Then

$$I^+(F_N) = \sup_{f \in F_N} \int \int_{\pi_2} f(X) \varphi_1(x_1) \varphi_2(x_2) dX = \int \int_{\pi_2} \gamma^+(X) \varphi_1(x_1) \times \\ \varphi_2(x_2) dX = \sum_{p=1}^{m^2} \int \int_{K_p} \gamma^+(X) \varphi_1(x_1) \varphi_2(x_2) dX = \sum_{p=1}^{m^2} I_p^+. \quad (3.10)$$

The choice of function $l(X)$ in (3.8) is determined by the condition

$$I_p^+ + I_{p+1}^+ = \sup_{f \in F_N} \int \int_{K_p \cup K_{p+1}} f(X) \varphi_1(x_1) \varphi_2(x_2) dX. \quad (3.11)$$

We also note that the need of “sewing” $A_{F_N}^+(X)$ and $A_{F_N}^-(X)$ directly follows from the fact that $\gamma^+(X) \in F_N$.

We define the function $l(X)$ in the following form

$$l(X) = -Lx_1 + B_1(x_2)x_2 + B_2(x_2), \quad (3.12)$$

where $B_1(x_2), B_2(x_2)$ are certain piecewise constant functions, values of which are defined by the relationship (3.8).

In the general case the problem of finding $B_1(x_2), B_2(x_2)$ in (3.12) is fairly difficult. Even in a simple case when zeros of functions $\varphi_1(x_1), \varphi_2(x_2)$ coincide with grid nodes, then the problem of construction of function $\gamma^+(X)$ is too difficult for this approach to be used in practice. Thus the need arises for a simpler close-to-optimal method.

Corollary 3.1 *Let functions $\varphi_1(x_1), \varphi_2(x_2)$ change sign when $x_1, x_2 \in [0, 1]$. Then the error of formulae (3.2), (3.3) will not be more than twice the optimal error.*

Proof. In Section 2 we constructed the majorants $A_{F_N}^+(X)$ and the minorants $A_{F_N}^-(X)$ in the cases when $F_N = C_{2,L_1,L_2,N}^2$ and $F_N = C_{2,L,L,N}^2$. We also recall that the function

$$f^*(X) = \frac{1}{2} (A_{F_N}^+(X) + A_{F_N}^-(X)) \quad (3.13)$$

is the optimal-by-accuracy approximation of function $f(X) \in F_N$.

For $F_N = C_{2,L,L,N}^2$ we denote

$$\hat{I}^*(f^*) = \int \int_{\pi_2} f^*(X) \varphi_1(x_1) \varphi_2(x_2) dX, X = (x_1, x_2). \quad (3.14)$$

In contrast to $\gamma^+(X)$ in the form (3.7), function $f^*(X) \in C_{2,L,L,N}^2$ is continuous as the sum of two continuous functions. By the same token, it satisfies the Lipschitz condition and passes through points $f_v, v = 1, \dots, N$. Hence, $\hat{I}^*(f^*) \in [I^-(F_N), I^+(F_N)]$, where $I^+(F_N)$ and $I^-(F_N)$ are limits of the values of integral (1.1) when $F_N = C_{2,L,L,N}^2$ (taking into account changes of the sign of functions $\varphi_1(x_1), \varphi_2(x_2)$). Moreover, the optimal-by-accuracy value of (1.1) for this class is

$$I^*(F_N) = \frac{1}{2} (I^+(F_N) + I^-(F_N)) \quad (3.15)$$

with the error determined as

$$\delta(F_N) = \frac{1}{2} (I^+(F_N) - I^-(F_N)). \quad (3.16)$$

It is easy to see that when functions $\varphi_1(x_1), \varphi_2(x_2)$ change sign for $x_1, x_2 \in [0, 1]$, $\hat{I}^*(f^*) \neq I^*(F_N)$ and the inequality

$$|\hat{I}^*(f^*) - I^*(F_N)| \leq \delta(F_N) \quad (3.17)$$

holds. Taking into account that

$$|I^2(f^*) - I^*(F_N)| \leq \delta(F_N), \quad (3.18)$$

and using the triangle inequality, from (3.17) and (3.18) we have

$$|\hat{I}^*(f^*) - I^2(F_N)| \leq 2\delta(F_N). \quad (3.19)$$

Inequality (3.19) leads to the statement of the theorem. The case $F_N = C_{2,L_1,L_2,N}^2$ is considered analogously. ■

It is worthwhile noting that relationships (3.8)–(3.18) hold not only for the function $f^*(X)$ but for any recovered function \tilde{f} from the class F_N . We also note that a spline-based approach proposed earlier in [17] can also be used for the construction of efficient cubature formulae for computing the integral $I^2(f)$ in classes $C_{2,L_1,L_2,N}^2$ and $C_{2,L,L,N}^2$. It can be shown, for example, that the class $C_{1,L,N \times M}^2$ is close to classes $C_{2,L_1,L_2,N}^2$ and $C_{2,L,L,N}^2$ and functions $\tilde{S}_1(x_1, x_2)$ and $\tilde{S}_2(x_1, x_2)$ constructed using a linear-spline approximation (see (3.1) in [17]) belong to classes $C_{2,L_1,L_2,N}^2$ and $C_{2,L,L,N}^2$ respectively. Moreover, in the general case, cubature formulae

$$Q_1(\tilde{S}_1) = \int_0^1 \int_0^1 \tilde{S}_1(x_1, x_2) \varphi_1(x_1) \varphi_2(x_2) dx_1 dx_2, \quad (3.20)$$

$$Q_2(\tilde{S}_2) = \int_0^1 \int_0^1 \tilde{S}_2(x_1, x_2) \varphi_1(x_1) \varphi_2(x_2) dx_1 dx_2, \quad (3.21)$$

have the same accuracy properties as formulae (3.2), (3.3). However, it is reasonable to apply formulae (3.20), (3.21) only in the case when we know not Lipschitz constants themselves but only their estimates. Although the method of integrand approximation by a linear spline proposed in [17] (see also [2, 3, 15, 16, 17, 21, 22, 23, 24]) allows us to construct optimal-by-order cubature formulae without knowledge of Lipschitz constants, that method is unable to constructively compute error estimates for such formulae.

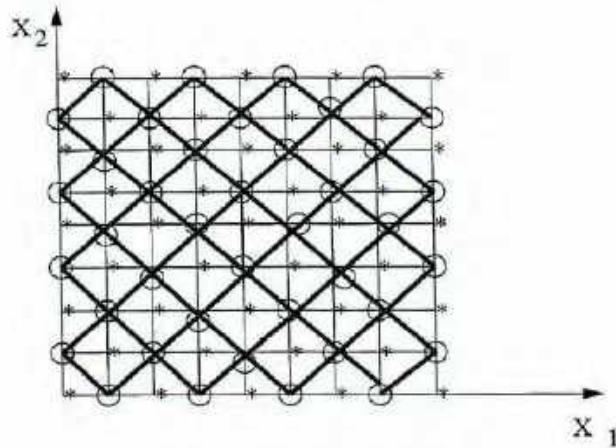
If *a priori* information about the problem is known exactly, then the approach proposed in this section has a number of advantages. First of all we admit that if zeros of functions $\varphi_1(x_1)$, $\varphi_2(x_2)$ are located relatively sparsely with respect to grid nodes (the weak oscillations case [16]), then in regions with constant sign of functions $\varphi_1(x_1)$, $\varphi_2(x_2)$ formulae (3.2) and (3.3) will be optimal-by-accuracy. Moreover, the proposed approach allows us to simultaneously construct an estimate of optimal error $v(F_N, \hat{I}^*, f)$ (see Corollary 3.1):

$$v(F_N, \hat{I}^*, f) \leq \frac{1}{2} \int \int_{\pi_2} (A_{F_N}^+(X) - A_{F_N}^-(X)) |\varphi_1(x_1) \varphi_2(x_2)| dX. \quad (3.22)$$

Therefore, in the case of strong oscillations of functions $\varphi_1(x_1)$, $\varphi_2(x_2)$ under exact *a priori* information, the application of cubature formulae (3.2) and (3.3) is more favorable.

4 The Choice of Grids in the Class $C_{2,L,L,N}^2$.

The passage from a functional class F to an interpolational class F_N is usually due to the desire to maximise the usage of *a priori* available information about the problem. However, in working with interpolational classes it is important to realise that in practice we often have to deal with functions with fairly complicated structures. Hence, for computing functional characteristics (such as function values) we may need an expensive physical or computational experiment. Such situations occur in automotive design problems, signal and image processing and many other applications [4, 5, 10, 8]. This leads to a dilemma. Indeed, on the one hand we have to in the most complete way obtain *a priori* information about

Figure 3: Splitting π_2 by the grid γ .

the problem. On the other hand, we have to decrease the number of expensive function evaluations.

In the construction of optimal-by-accuracy and optimal-by-order cubature formulae for computing integrals $I_i^2(f)$, $i = 1, 2, 3$ in the class $C_{2,L,L,N}^2$, the resolution of this dilemma requires the consideration of optimal (in a certain sense specified below) grids in π_2 that allow us to compute function values only at nodes of such a grid.

Let $F_N = C_{2,L,L,N}^2$. First, we consider the grid γ which splits π_2 into $4n^2$ equal elementary squares K_p with side $1/(2n)$, $p = 1, \dots, 4n^2$ (see Fig 3). Then, we split γ into two subsets: $\gamma_1 \cup \gamma_2 = \gamma$, $\gamma_1 \cap \gamma_2 = \emptyset$. Let us assume that the grid $\gamma_1 \subset \gamma$ consists of nodes $X_{v_1} = ((i-1)\frac{1}{2n}; (j-1)\frac{1}{2n})$, $v_1 = (i-1)(2n+1) + j$, $i = \{1, 3, \dots, 2n+1\}$, $j = \{2, 4, \dots, 2n\}$ and nodes $X_{v_2} = ((i-1)\frac{1}{2n}, \frac{j}{2n})$, $v_2 = (i-1)(2n+1) + j + 1$, $i = \{2, 4, \dots, 2n\}$, $j = \{1, 3, \dots, 2n+1\}$. Similarly, let the grid $\gamma_2 \subset \gamma$ consist of the nodes $X_{\mu_1} = ((i-1)\frac{1}{2n}, (j-1)\frac{1}{2n})$, $\mu_1 = (i-1)(2n+1) + j$, $i, j = \{1, 3, \dots, 2n+1\}$ and nodes $X_{\mu_2} = ((i-1)\frac{1}{2n}, \frac{j}{2n})$, $\mu_2 = (i-1)(2n+1) + j$, $i, j = \{2, 4, \dots, 2n\}$. The grid γ_1 splits the domain π_2 into certain elementary regions of rhombic forms and their parts. The nodes of γ_2 are centres of these rhombuses. On Fig. 3 the nodes of γ_1 are highlighted with circles, and the nodes of γ_2 are highlighted with stars. Further denote

- the set of nodes from γ_1 that lie on the sides of π_2 by $\tilde{\gamma}_1$;
- the set of nodes from γ_2 that lie on the sides of π_2 by $\tilde{\gamma}_2$;
- the set of nodes from $\tilde{\gamma}_2$ that consists of the vertices of π_2 by $\tilde{\gamma}_2^*$.

It is easy to see that $\tilde{\gamma}_2$ consists completely of nodes of the form X_{μ_1} .

We assume that function $f(X)$ may be given by its values not in all nodes of the grid γ , but only at nodes of the grid $\gamma_1 \subset \gamma$. Therefore, we allow the situation when we are given function values not in $N = (2n+1)^2$ nodes, but only at $\bar{N} = 2n(n+1)$ nodes, i.e. only at two vertices of the square K_p , $p = 1, \dots, 4n^2$. Hence, instead of the class $C_{2,L,L,N}^2$ it is more reasonable to consider the class $C_{2,L,L,\bar{N}}^2$, which is defined as follows. It is the class of such functions that are defined in the domain $\pi_2 = \{X = (x_1, x_2) : 0 \leq x_i \leq 1, i = 1, 2\}$, that satisfy the Lipschitz condition with constant L (in each variable) and that take fixed values

f_1, \dots, f_N at nodes X_1, \dots, X_N of grid γ_1 respectively.

Let us consider a certain node $X_{s_0} = (\bar{i}\bar{h}, \bar{j}\bar{h})$, $X_{s_0} \in \gamma_2 \setminus \tilde{\gamma}_2$, $\bar{h} = 1/(2n)$. From the set of all nodes we single out a subset of nodes that is defined as follows: $X_{s_1} = ((\bar{i}-1)\bar{h}, \bar{j}\bar{h})$, $X_{s_2} = (\bar{i}\bar{h}, (\bar{j}+1)\bar{h})$, $X_{s_3} = ((\bar{i}+1)\bar{h}, \bar{j}\bar{h})$, $X_{s_4} = (\bar{i}\bar{h}, (\bar{j}-1)\bar{h})$. Let $\sigma(\{X_{s_l}\}_{l=1,2,3,4}) = \{s_l\}_{l=1,2,3,4}$, $f(X_{s_l}) = f_{s_l}$, $l = 1, 2, 3, 4$. Then the following result holds.

Theorem 4.1 *Let $F_N = C_{2,L,L,\bar{N}}^2$. Then $\forall X_{s_0} \in \gamma_2 \setminus \tilde{\gamma}_2$ we have*

$$A_{C_{2,L,L,\bar{N}}}^+(X_{s_0}) = \min_{s \in \sigma(\{X_{s_l}\}_{l=1,2,3,4})} f_s + L\bar{h}, \quad A_{C_{2,L,L,\bar{N}}}^-(X_{s_0}) = \max_{s \in \sigma(\{X_{s_l}\}_{l=1,2,3,4})} f_s - L\bar{h}. \quad (4.1)$$

Proof. First we introduce the functions defined $\forall X_v \in \gamma_1$ $\{X_{s_l}\}_{l=1,2,3,4}$ as

$$g_{s_l}(X) = f_{s_l} + L\|X - X_{s_l}\|_2, \quad l = 1, 2, 3, 4, \quad g_{s_v}(X) = f_v + L\|X - X_v\|_2. \quad (4.2)$$

Let us show that

$$g_v(X_{s_0}) - g_{s_l}(X_{s_0}) \geq 0 \quad \forall v \in \sigma(\gamma_1 \setminus \{X_{s_l}\}_{l=1,2,3,4}), \quad \forall s_l, l = 1, 2, 3, 4. \quad (4.3)$$

As an example, we consider the node $X_{\bar{v}} = ((\bar{i}-1-k_1)\bar{h}, (\bar{j}+k_2)\bar{h})$, $X_{\bar{v}} \in \gamma_1 \setminus \{X_{s_l}\}_{l=1,2,3,4}$. We have

$$\begin{aligned} g_{\bar{v}}(X_{s_0}) - g_{s_1}(X_{s_0}) &= f_{\bar{v}} + L\|X_{s_0} - X_{\bar{v}}\|_2 - f_{s_1} - L\|X_{s_0} - X_{s_1}\|_2 \geq \\ &- L(k_1 + k_2)\bar{h} + L(k_1 + k_2 + 1)\bar{h} - L\bar{h} = 0. \end{aligned} \quad (4.4)$$

Similarly it can be proved that inequalities analogous to (4.3) also hold for $s_l, l = 2, 3, 4$. From the relationship (2.1) and inequalities (4.3) it follows that

$$A_{C_{2,L,L,\bar{N}}}^+(X_{s_0}) = \min_{i=1, \dots, N} (f_i + L\|X_{s_0} - X_{s_i}\|_2) = \min_{s \in \sigma(\{X_{s_l}\}_{l=1,2,3,4})} f_s + L\bar{h}. \quad (4.5)$$

Therefore, we have shown that the value of the majorant of the class $C_{2,L,L,\bar{N}}^2$ at any point $X_{s_0} \in \gamma_2 \setminus \tilde{\gamma}_2$ is determined by its value at four closest to X_{s_0} nodes of the grid γ_1 (i.e. by the nodes of the form X_{s_l} for which $\|X_{s_0} - X_{s_l}\|_2 = 1/(2n)$, $l = 1, 2, 3, 4$). For the minorant $A_{C_{2,L,L,\bar{N}}}^-$ the proof is analogous. ■

Corollary 4.1 *Let $X_{s_0} \in \tilde{\gamma}_2 \setminus \gamma_2^*$. Then the following relationships hold*

$$A_{C_{2,L,L,\bar{N}}}^+(X_{s_0}) = \min_{l=1,2,3} f_{s_l} + L\bar{h}, \quad A_{C_{2,L,L,\bar{N}}}^-(X_{s_0}) = \max_{l=1,2,3} f_{s_l} - L\bar{h}, \quad (4.6)$$

with s_l the number of the node X_{s_l} of the grid γ_1 such that $\|X_{s_0} - X_{s_l}\|_2 = 1/(2n)$, $l = 1, 2, 3$.

Corollary 4.2 *Let $X_{s_0} \in \gamma_2^*$. Then the following relationships hold*

$$A_{C_{2,L,L,\bar{N}}}^+(X_{s_0}) = \min(f_{s_1}, f_{s_2}) + L\bar{h}, \quad A_{C_{2,L,L,\bar{N}}}^-(X_{s_0}) = \max(f_{s_1}, f_{s_2}) - L\bar{h}, \quad (4.7)$$

with s_1, s_2 the numbers of the nodes X_{s_1}, X_{s_2} of the grid $\tilde{\gamma}_1$ such that $\|X_{s_0} - X_{s_i}\|_2 = 1/(2n)$, $i = 1, 2$.

Proofs of the Corollaries 4.1 and 4.2 are analogous to the proof of Theorem 4.1.

As we mentioned before, the difficulties in the realisation of approach (1.6)–(1.8) lie in the need for constructive computation of quantities $I_i^+(F_N), I_i^-(F_N), i = 1, 2, 3$. In Sections 2 and 3 we constructively solved this problem under the assumption that functions of the given class are known at the nodes of the grid γ . Hence, for computing $I_i^+(C_{2,L,L,\bar{N}}^2)$ we extend the definition of $f(X)$ as $f(X) = A_{C_{2,L,L,\bar{N}}^2}^+(X), X \in \gamma_2$, and for computing $I_i^-(C_{2,L,L,\bar{N}}^2)$ we extend the definition of $f(X)$ as $f(X) = A_{C_{2,L,L,\bar{N}}^2}^-(X), X \in \gamma_2$, where in both cases $i = 1, 2, 3$.

Lemma 4.1 *For the majorant of the class $C_{2,L,L,\bar{N}}^2$ the following relationship holds*

$$A_{C_{2,L,L,\bar{N}}^2}^+(X) = \min_{\mu=1,\dots,N} (f_\mu + L\|X - X_\mu\|_2), \quad (4.8)$$

with $f_\mu = f(X_\mu)$ for $X_\mu \in \gamma_1$ and with $f_\mu = A_{C_{2,L,L,\bar{N}}^2}^+(X_\mu)$ for $X_\mu \in \gamma_2$.

Proof. Let us consider the function $B^+(X) = \min_{\mu=1,\dots,N} (f_\mu + L\|X - X_\mu\|_2)$ with $f_\mu = f(X_\mu)$ for $X_\mu \in \gamma_1$ and with $f_\mu = A_{C_{2,L,L,\bar{N}}^2}^+(X_\mu)$ for $X_\mu \in \gamma_2$.

According to the definition of the majorant we have

$$A_{C_{2,L,L,\bar{N}}^2}^+(X) = \min_{v=1,\dots,\bar{N}} (f_v + L\|X - X_v\|_2). \quad (4.9)$$

Let us now show that functions $A_{C_{2,L,L,\bar{N}}^2}^+(X)$ and $B^+(X)$ coincide. Indeed,

$$\begin{aligned} A_{C_{2,L,L,\bar{N}}^2}^+(X) - B^+(X) &= \min_{v=1,\dots,\bar{N}} (f_v + L\|X - X_v\|_2) - \min_{\mu=1,\dots,N} (f_\mu + L\|X - X_\mu\|_2) = \\ &f_{v_0} + L\|X - X_{v_0}\|_2 - f_{\mu_0} - L\|X - X_{\mu_0}\|_2 \geq \\ &f_{v_0} + L\|X - X_{v_0}\|_2 - f_{v_0} - L\|X - X_{v_0}\|_2 = 0. \end{aligned} \quad (4.10)$$

The inequality in (4.10) follows from the fact that $\gamma_1 \subset \gamma$ and $B^+(X) \leq f_{v_0} + L\|X - X_{v_0}\|_2$.

If $X_{\mu_0} \in \gamma_1$, then

$$\begin{aligned} f_{v_0} + L\|X - X_{v_0}\|_2 - f_{\mu_0} - L\|X - X_{\mu_0}\|_2 &\leq \\ f_{\mu_0} + L\|X - X_{\mu_0}\|_2 - f_{\mu_0} - L\|X - X_{\mu_0}\|_2 &= 0. \end{aligned} \quad (4.11)$$

On the other hand, if $X_{\mu_0} \in \gamma_2$, then from Theorem 4.1 and its corollaries it follows that

$$\begin{aligned} f_{v_0} + L\|X - X_{v_0}\|_2 - f_{\mu_0} - L\|X - X_{\mu_0}\|_2 &= f_{v_0} + L\|X - X_{v_0}\|_2 - f_{v_1} - L\bar{h} - \\ L\|X - X_{\mu_0}\|_2 &\leq f_{v_1} + L\|X - X_{v_1}\|_2 - f_{v_1} - L(\bar{h} + |x_1 - x_{1,\mu_0}| + |x_2 - x_{2,\mu_0}|) \leq \\ L(\bar{h} + |x_1 - x_{1,\mu_0}| + |x_2 - x_{2,\mu_0}|) - L(\bar{h} + |x_1 - x_{1,\mu_0}| + |x_2 - x_{2,\mu_0}|) &= 0. \end{aligned} \quad (4.12)$$

Inequalities (4.10)–(4.12) lead to the statement of the lemma. ■

Analogously to the above proof it can be shown that

$$A_{C_{2,L,L,\bar{N}}^2}^-(X) = \max_{\mu=1,\dots,N} (f_\mu - L\|X - X_\mu\|_2), \quad (4.13)$$

with $f_\mu = f(X_\mu)$ for $X_\mu \in \gamma_1$ and with $f_\mu = A_{C_{2,L,L,\bar{N}}^2}^-(X_\mu)$ for $X_\mu \in \gamma_2$.

The results obtained in Sections 2–4 allow us to efficiently solve problems (1.10)–(1.11), (1.12)–(1.13) as well as construct cubature formulae for computing (1.1)–(1.3) with a substantial reduction of required *a priori* information. Indeed, we propose to use the information about values of function $f(X)$ only at those nodes of the grid γ_1 which are centers of such balls with radius $1/(2n)$ that cover π_2 in an optimal way.

The grids γ and γ_1 split the domain π_2 into elementary regions K_p and K'_d respectively. Therewith, diameters of K_p and K'_d coincide in the given norm. From this fact it follows that the accuracy estimates for the recovery of $f^*(X)$ in classes $C_{2,L,L,N}^2$ and $C_{2,L,L,N}^2$ coincide in spite of the fact that the number of nodes of the grid γ is two times larger than the number of nodes of the grid γ_1 . Finally we note that the use of the optimal-by-accuracy recovery of function $f^*(X) \in C_{2,L,L,N}^2$ in formulae (3.2)–(3.3) justifies the application of the proposed grids γ_1 for these cubature formulae. Advantages of the proposed approach can be easily observed when $\varphi_1(x_1) = 1, \varphi_2(x_2) = 1$.

We conclude this section with the following result.

Theorem 4.2 *The following estimate*

$$\bar{\delta}(C_{2,L,L,N}^2) \leq \bar{\delta}(C_{2,L,L,\bar{N}}^2) \quad (4.14)$$

holds.

Proof. It is easy to see that

$$A_{C_{2,L,L,N}^2}^+(X) \geq A_{C_{2,L,L,N}^2}^+(X), \quad A_{C_{2,L,L,N}^2}^-(X) \leq A_{C_{2,L,L,N}^2}^-(X), \quad X \in \pi_2. \quad (4.15)$$

In addition, when $f(X) = \text{const}, X \in \gamma$ we have

$$\delta(C_{2,L,L,N}^2) = \bar{\delta}(C_{2,L,L,N}^2) = \bar{\delta}(C_{2,L,L,\bar{N}}^2). \quad (4.16)$$

From (4.15) and (4.16) the statement of the theorem follows immediately. ■

5 Optimal-By-Accuracy Cubature Formulae for Functions From Classes $C_{2,L_1,L_2,N}^2$ and $C_{2,L,L,N}^2$.

Let $F_N = C_{2,L_1,L_2,N}^2$. Using Corollary 2.2 and relationships (2.32)–(2.36), it is easy to show that the splitting of \tilde{K}_p into regions $\tilde{\Omega}_l^+, l = 1, 2, 3, 4, p = 1, \dots, mm_1$ is determined by points $O_1(\bar{x}_{1,i}, \bar{x}_{2,j}), O_2 = (\bar{x}_{1,i}, \bar{x}_{2,j})$ (the situation is similar to that shown in Fig. 1) with

$$\begin{cases} \bar{x}_{1,i} = x_{1,i} + \frac{h_1}{2} + \frac{f_{i+1,j} - f_{i,j}}{2L_1} \delta_1 + \frac{f_{i+1,j+1} - f_{i,j+1}}{2L_1} \delta_2, \\ \bar{x}_{1,i} = x_{1,i} + \frac{h_1}{2} + \frac{f_{i+1,j+1} - f_{i,j+1}}{2L_1} \delta_1 + \frac{f_{i+1,j} - f_{i,j}}{2L_1} \delta_2, \end{cases} \quad (5.1)$$

$$\bar{x}_{2,j} = x_{2,j} + \frac{h_2}{2} + \frac{f_{i+1,j+1} - f_{i+1,j}}{2L_2}, \quad \bar{x}_{2,j} = x_{2,j} + \frac{h_2}{2} + \frac{f_{i,j+1} - f_{i,j}}{2L_2}, \quad (5.2)$$

$$\begin{cases} \delta_1 = \frac{1}{2} (1 - \text{sign}(f_{i,j} + f_{i+1,j+1} - f_{i,j+1} - f_{i+1,j})), \\ \delta_2 = \frac{1}{2} (1 + \text{sign}(f_{i,j} + f_{i+1,j+1} - f_{i,j+1} - f_{i+1,j})), \end{cases} \quad (5.3)$$

and $i = 1, \dots, m$, $j = 1, \dots, m_1$. In an analogous way we obtain that the splitting of \tilde{K}_p into regions $\tilde{\Omega}_l^-, l = 1, 2, 3, 4$, $p = 1, \dots, mm_1$ is determined by points $O_3 = (\tilde{x}_{1,i}, \tilde{x}_{2,j})$, $O_4 = (\tilde{\tilde{x}}_{1,i}, \tilde{\tilde{x}}_{2,j})$ (similar to Fig. 1) with

$$\begin{cases} \tilde{x}_{1,i} = x_{1,i} + \frac{h_1}{2} - \frac{f_{i+1,j+1} - f_{i,j+1}}{2L_1} \delta_1 - \frac{f_{i+1,j} - f_{i,j}}{2L_1} \delta_2, \\ \tilde{\tilde{x}}_{1,i} = x_{1,i} + \frac{h_1}{2} - \frac{f_{i+1,j} - f_{i,j}}{2L_1} \delta_1 - \frac{f_{i+1,j+1} - f_{i,j+1}}{2L_1} \delta_2, \end{cases} \quad (5.4)$$

$$\tilde{x}_{2,j} = x_{2,j} + \frac{h_2}{2} - \frac{f_{i+1,j+1} - f_{i+1,j}}{2L_2}, \quad \tilde{\tilde{x}}_{2,j} = x_{2,j} + \frac{h_2}{2} - \frac{f_{i,j+1} - f_{i,j}}{2L_2}, \quad (5.5)$$

and δ_1, δ_2 defined by relationships (5.3), $i = 1, \dots, m$, $j = 1, \dots, m_1$. Using Corollary 2.3 and the relationship (2.38) we split \tilde{K}_p into $\tilde{\Omega}_l^+$, $l = 1, 2$ by points $O_1 = (\hat{x}_{1,i}, x_{2,m_1+1})$, $O_2 = (\hat{\tilde{x}}_{1,i}, 1)$, and similarly, split of K_p into $\tilde{\Omega}_l^-$, $l = 1, 2$ by points $O_3 = (\hat{x}_{1,i}, x_{2,m_1+1})$, $O_4 = (\hat{\tilde{x}}_{1,i}, 1)$ (see Fig. 1), where

$$\hat{x}_{1,i} = x_{1,i} + \frac{h_1}{2} + \frac{f_{i+1,m_1+1} - f_{i,m_1+1}}{2L_1}, \quad \hat{\tilde{x}}_{1,i} = x_{1,i} + \frac{h_1}{2} - \frac{f_{i+1,m_1+1} - f_{i,m_1+1}}{2L_1} \quad (5.6)$$

and $i = 1, \dots, m$, $p = 1, \dots, m$.

Now let $F_N = C_{2,L,L,N}^2$. Using Corollary 2.1 and relationships (2.22)–(2.26), it is easy to show that the splitting of K_p into Ω_l^+ , $l = 1, \dots, 4$ is determined by points $O_1 = (\bar{x}_{1,i}, \bar{\tilde{x}}_{2,j})$, $O_2 = (\bar{x}_{1,i}, \bar{x}_{2,j})$ and the splitting K_p into Ω_l^- , $l = 1, 2, 3, 4$ is determined by points $O_3 = (\tilde{x}_{1,i}, \tilde{\bar{x}}_{2,j})$, $O_4 = (\tilde{\tilde{x}}_{1,i}, \tilde{\bar{x}}_{2,j})$ (see Fig. 1), where $\bar{x}_{1,i}$, $\bar{\tilde{x}}_{1,i}$, $\bar{x}_{1,i}$, $\tilde{\bar{x}}_{1,i}$ ($i = 1, \dots, m$) and $\bar{x}_{2,j}$, $\bar{\tilde{x}}_{2,j}$, $\tilde{\bar{x}}_{2,j}$, $\tilde{\tilde{x}}_{2,j}$ ($j = 1, \dots, m$, $p = 1, \dots, m^2$) are computed by formulae (5.1)–(5.5) respectively for $L_1 = L_2 = L$ and $h_1 = h_2 = h$.

We start by considering the problem of computing optimal-by-accuracy values of integral $I_1^2(f)$. Let us introduce the following notation

$$U_p^* = \frac{1}{2}(U_p^+ + U_p^-) = \frac{1}{2} \left(\int \int_{K_p} A_{C_{2,L,L,N}^2}^+(X) dX + \int \int_{K_p} A_{C_{2,L,L,N}^2}^-(X) dX \right), \quad (5.7)$$

where $p = 1, \dots, m^2$;

$$\bar{U}_p^* = \frac{1}{2}(\bar{U}_p^+ + \bar{U}_p^-) = \frac{1}{2} \left(\int \int_{\bar{K}_p} A_{C_{2,L_1,L_2,N}^2}^+(X) dX + \int \int_{\bar{K}_p} A_{C_{2,L_1,L_2,N}^2}^-(X) dX \right), \quad (5.8)$$

where $p = 1, \dots, mm_1$ and

$$\tilde{U}_p^* = \frac{1}{2}(\tilde{U}_p^+ + \tilde{U}_p^-) = \frac{1}{2} \left(\int \int_{\tilde{K}_p} A_{C_{2,L_1,L_2,N}^2}^+(X) dX + \int \int_{\tilde{K}_p} A_{C_{2,L_1,L_2,N}^2}^-(X) dX \right), \quad (5.9)$$

where $p = 1, \dots, m$. From Theorem 3.1 it follows that the optimal-by-accuracy cubature formula for computing integral $I_1^2(f)$ has the form

$$U^* = \sum_{p=1}^{m^2} U_p^* \text{ when } F_N = C_{2,L,L,N}^2 \quad (5.10)$$

and the form

$$\bar{U}^* = \sum_{p=1}^{mm_1} \bar{U}_p^* + \sum_{p=1}^m \tilde{U}_p^* \text{ when } F_N = C_{\bar{2},L_1,L_2,N}^2. \quad (5.11)$$

Moreover,

$$\bar{\delta}(C_{2,L,L,N}^2) = \frac{1}{2} \sum_{p=1}^{m^2} (U_p^+ - U_p^-), \bar{\delta}(C_{\bar{2},L_1,L_2,N}^2) = \frac{1}{2} \left(\sum_{p=1}^{mm_1} (\bar{U}_p^+ - \bar{U}_p^-) + \sum_{p=1}^m (\tilde{U}_p^+ - \tilde{U}_p^-) \right). \quad (5.12)$$

Let

$$\begin{cases} \bar{\kappa}_{1,i} = x_{1,i} + \frac{h_1}{2} + \delta \left(\frac{f_{i+1,j} - f_{i,j}}{2L_1} \mu_1 + \frac{f_{i+1,j+1} - f_{i,j+1}}{2L_1} \mu_2 \right), \\ \bar{\bar{\kappa}}_{1,i} = x_{1,i} + \frac{h_1}{2} + \delta \left(\frac{f_{i+1,j+1} - f_{i,j+1}}{2L_1} \mu_1 + \frac{f_{i+1,j} - f_{i,j}}{2L_1} \mu_2 \right), \end{cases} \quad (5.13)$$

$$\begin{cases} \bar{\kappa}_{2,j} = x_{2,j} + \frac{h_2}{2} + \delta \frac{f_{i+1,j+1} - f_{i+1,j}}{2L_2}, \\ \bar{\bar{\kappa}}_{2,j} = x_{2,j} + \frac{h_2}{2} + \delta \frac{f_{i,j+1} - f_{i,j}}{2L_2}, \end{cases} \quad (5.14)$$

$$\bar{\kappa}_{1,i} = x_{1,i} + \frac{h_1}{2} + \delta \frac{f_{i+1,m_1+1} - f_{i,m_1+1}}{2L_1}, \quad i = 1, \dots, m, j = 1, \dots, m_1, \quad (5.15)$$

where

$$\mu_1 = \frac{1}{2} ((1 - \delta)\delta_2 + (1 + \delta)\delta_1), \quad \mu_2 = \frac{1}{2} ((1 - \delta)\delta_1 + (1 + \delta)\delta_2), \quad (5.16)$$

$\delta \in \{-1, 1\}$, and δ_1, δ_2 defined by (5.3). Then the equation of the line that passes points $(\bar{\kappa}_{1,i}, \bar{\kappa}_{2,j}), (\bar{\bar{\kappa}}_{1,i}, \bar{\bar{\kappa}}_{2,j})$ has the form

$$\frac{x_1 - \bar{\kappa}_{1,i}}{\bar{\kappa}_{1,i} - \bar{\bar{\kappa}}_{1,i}} = \frac{x_2 - \bar{\kappa}_{2,j}}{\bar{\kappa}_{2,j} - \bar{\bar{\kappa}}_{2,j}}, \text{ or } (x_2 - \bar{\kappa}_{2,j})(\bar{\bar{\kappa}}_{1,i} - \bar{\kappa}_{1,i}) = (x_1 - \bar{\kappa}_{1,i})(\bar{\bar{\kappa}}_{2,j} - \bar{\kappa}_{2,j}), \quad (5.17)$$

that immediately leads to

$$x_2 = \bar{\kappa}_{2,j} + (x_1 - \bar{\kappa}_{1,i})(\bar{\bar{\kappa}}_{2,j} - \bar{\kappa}_{2,j}) / (\bar{\bar{\kappa}}_{1,i} - \bar{\kappa}_{1,i}). \quad (5.18)$$

Since

$$\bar{\bar{\kappa}}_{1,i} - \bar{\kappa}_{1,i} = \delta(\mu_1 - \mu_2) \frac{f_{i+1,j+1} - f_{i,j+1} - f_{i+1,j} + f_{i,j}}{2L_1}, \quad (5.19)$$

and

$$\bar{\kappa}_{2,j} - \tilde{\kappa}_{2,j} = \frac{\delta}{2L_2}(f_{i,j+1} + f_{i+1,j} - f_{i,j} - f_{i+1,j+1}), \quad (5.20)$$

we have

$$(\bar{\kappa}_{2,j} - \tilde{\kappa}_{2,j}) / (\bar{\kappa}_{1,i} - \tilde{\kappa}_{1,i}) = -\frac{L_1}{L_2}\mu_1 + \frac{L_1}{L_2}\mu_2. \quad (5.21)$$

From (5.21) we get

$$x_2 = \bar{\kappa}_{2,j} + \frac{L_1}{L_2}(\mu_1 - \mu_2)(\bar{\kappa}_{1,i} - x_1), \quad i = 1, \dots, m, \quad j = 1, \dots, m_1. \quad (5.22)$$

It is easy to see that equation (5.22) is the equation of the lines (O_1, O_2) and (O_3, O_4) with $\delta = 1$ and $\delta = -1$ respectively (similar to the situation shown in Fig. 1). By using Corollaries 2.2, 2.3 and by taking into account relationships (5.13)–(5.22) we get

$$\begin{aligned} \bar{U}_p^\pm &= \int_{x_{1,i}}^{\bar{\kappa}_{1,i}} \int_{x_{2,j}}^{\bar{\kappa}_{2,j}} (f_{i,j} + \delta(L_1(x_1 - x_{1,i}) + L_2(x_2 - x_{2,j}))) dx_2 dx_1 + \\ &\quad \int_{x_{1,i}}^{\bar{\kappa}_{1,i}} \int_{\bar{\kappa}_{2,j}}^{x_{2,j+1}} (f_{i,j+1} + \delta(L_1(x_1 - x_{1,i}) + L_2(x_{2,j+1} - x_2))) dx_2 dx_1 + \\ &\quad \int_{\bar{\kappa}_{1,i}}^{x_{1,i+1}} \int_{\bar{\kappa}_{2,j}}^{x_{2,j+1}} (f_{i+1,j+1} + \delta(L_1(x_{1,i+1} - x_1) + L_2(x_{2,j+1} - x_2))) dx_2 dx_1 + \\ &\quad \int_{\bar{\kappa}_{1,i}}^{x_{1,i+1}} \int_{x_{2,j}}^{\bar{\kappa}_{2,j}} (f_{i+1,j} + \delta(L_1(x_{1,i+1} - x_1) + L_2(x_2 - x_{2,j}))) dx_2 dx_1 + \\ &\quad \int_{\bar{\kappa}_{1,i}}^{\bar{\kappa}_{1,i}} \int_{x_{2,j}}^{\bar{\kappa}_{2,j} + \frac{L_1}{L_2}(\mu_1 - \mu_2)(\bar{\kappa}_{1,i} - x_1)} (\mu_1 f_{1,j} + \mu_2 f_{i+1,j} + \delta(L_1((\mu_1 - \mu_2)x_1 - \mu_1 x_{1,i} + \mu_2 x_{1,i+1}) + \\ &\quad L_2(x_2 - x_{2,j}))) dx_2 dx_1 + \int_{\bar{\kappa}_{1,i}}^{\bar{\kappa}_{1,i}} \int_{\bar{\kappa}_{2,j} + \frac{L_1}{L_2}(\mu_1 - \mu_2)(\bar{\kappa}_{1,i} - x_1)}^{x_{2,j+1}} (\mu_1 f_{i+1,j+1} + \mu_2 f_{i,j+1} + \delta(L_1((\mu_2 - \mu_1)x_1 - \mu_2 x_{1,i} - \mu_1 x_{1,i+1}) + L_2(x_{2,j+1} - x_2))) dx_2 dx_1, \end{aligned} \quad (5.23)$$

where $p = 1, \dots, mm_1, i = 1, \dots, m, j = 1, \dots, m_1$. Setting $\delta = 1$ and then $\delta = -1$ in (5.23), we get expressions for \bar{U}_p^+ and for \bar{U}_p^- respectively ($p = 1, \dots, mm_1$). We start with computing the sum $\bar{S}_p^\pm, p = 1, \dots, mm_1$ for the first four integrals in (5.23):

$$\begin{aligned} \bar{S}_p^\pm &= (f_{i,j} - \delta(L_1 x_{1,i} + L_2 x_{2,j})) (\bar{\kappa}_{1,i} - x_{1,i})(\bar{\kappa}_{2,j} - x_{2,j}) + (f_{i,j+1} + \delta(-L_1 x_{1,i} + L_2 x_{2,j+1})) \times \\ &\quad (\bar{\kappa}_{1,i} - x_{1,i})(x_{2,j+1} - \bar{\kappa}_{2,j}) + (f_{i+1,j+1} + \delta(L_1 x_{1,i+1} + L_2 x_{2,j+1})) (x_{1,i+1} - \bar{\kappa}_{1,i}) \times \\ &\quad (x_{2,j+1} - \bar{\kappa}_{2,j}) + (f_{i+1,j} + \delta(L_1 x_{1,i+1} - L_2 x_{2,j})) (x_{1,i+1} - \bar{\kappa}_{1,i})(\bar{\kappa}_{2,j} - x_{2,j}) + \end{aligned}$$

$$\begin{aligned}
 & \delta \left(L_1 \frac{(\bar{\kappa}_{1,i}^2 - x_{1,i}^2)}{2} (\bar{\kappa}_{2,j} - x_{2,j}) + L_2 \frac{(\bar{\kappa}_{2,j}^2 - x_{2,j}^2)}{2} (\bar{\kappa}_{1,i} - x_{1,i}) + L_1 \frac{(\bar{\kappa}_{1,i}^2 - x_{1,i}^2)}{2} \times \right. \\
 & (x_{2,j+1} - \bar{\kappa}_{2,j}) - L_2 \frac{(x_{2,j+1}^2 - \bar{\kappa}_{2,j}^2)}{2} (\bar{\kappa}_{1,i} - x_{1,i}) - L_1 \frac{(x_{1,i+1}^2 - \bar{\kappa}_{1,i}^2)}{2} (x_{2,j+1} - \bar{\kappa}_{2,j}) - \\
 & L_2 \frac{(x_{2,j+1}^2 - \bar{\kappa}_{2,j}^2)}{2} (x_{1,i+1} - \bar{\kappa}_{1,i}) - L_1 \frac{(x_{1,i+1}^2 - \bar{\kappa}_{1,i}^2)}{2} (\bar{\kappa}_{2,j} - x_{2,j}) + L_2 \frac{(\bar{\kappa}_{2,j}^2 - x_{2,j}^2)}{2} \\
 & (x_{1,i+1} - \bar{\kappa}_{1,i})) = (f_{i,j} + \delta(L_1 \bar{\kappa}_{1,i} + L_2 \bar{\kappa}_{2,j}))(\bar{\kappa}_{1,i} - x_{1,i}) \left(\frac{h_2}{2} + \frac{\delta(f_{i,j+1} - f_{i,j})}{2L_2} \right) + \\
 & (f_{i,j+1} + \delta(L_1 \bar{\kappa}_{1,i} - L_2 \bar{\kappa}_{2,j}))(\bar{\kappa}_{1,i} - x_{1,i}) \left(\frac{h_2}{2} - \frac{\delta(f_{i,j+1} - f_{i,j})}{2L_2} \right) + \\
 & (f_{i+1,j+1} - \delta(L_1 \bar{\kappa}_{1,i} + L_2 \bar{\kappa}_{2,j}))(x_{1,i+1} - \bar{\kappa}_{1,i}) \left(\frac{h_2}{2} - \frac{\delta(f_{i+1,j+1} - f_{i+1,j})}{2L_2} \right) + \\
 & (f_{i+1,j} - \delta(L_1 \bar{\kappa}_{1,i} - L_2 \bar{\kappa}_{2,j}))(x_{1,i+1} - \bar{\kappa}_{1,i}) \left(\frac{h_2}{2} + \frac{\delta(f_{i+1,j+1} - f_{i+1,j})}{2L_2} \right) = \frac{h_2}{2} \times \\
 & ((f_{i,j} + f_{i,j+1}) (\bar{\kappa}_{1,i} - x_{1,i}) + (f_{i+1,j+1} + f_{i+1,j})(x_{1,i+1} - \bar{\kappa}_{1,i}) + 2\delta L_1 \left(\bar{\kappa}_{1,i} (\bar{\kappa}_{1,i} - x_{1,i}) - \right. \\
 & \left. \bar{\kappa}_{1,i}(x_{1,i+1} - \bar{\kappa}_{1,i}) \right)) + \delta(\bar{\kappa}_{1,i} - x_{1,i}) (\delta(x_{2,j} + \frac{h_2}{2} + \delta \frac{(f_{i,j+1} - f_{i,j})}{2L_2})) (f_{i,j+1} - f_{i,j}) - \\
 & \frac{(f_{i,j+1} - f_{i,j})^2}{2L_2} + \delta(x_{1,i+1} - \bar{\kappa}_{1,i}) \left(\delta \left(x_{2,j} + \frac{h_2}{2} + \frac{\delta(f_{i+1,j+1} - f_{i+1,j})}{2L_2} \right) \right. \\
 & \left. (f_{i+1,j+1} - f_{i+1,j}) - \frac{(f_{i+1,j+1} - f_{i+1,j})^2}{2L_2} \right) = (\bar{\kappa}_{1,i} - x_{1,i}) (x_{2,j+1} f_{i,j+1} - x_{2,j} f_{i,j} + \\
 & \delta h_2 L_1 \bar{\kappa}_{1,i}) + (x_{1,i+1} - \bar{\kappa}_{1,i}) (x_{2,j+1} f_{i+1,j+1} - x_{2,j} f_{i+1,j} + \delta h_2 L_1 \bar{\kappa}_{1,i}). \tag{5.24}
 \end{aligned}$$

Then we compute the sum \bar{S}_p^\pm of the last two integrals in the relationship (5.23) for $p = 1, \dots, mm_1$:

$$\begin{aligned}
 \bar{S}_p^\pm &= (\mu_1 f_{i,j} + \mu_2 f_{i+1,j} - \delta(L_1(\mu_1 x_{1,i} - \mu_2 x_{1,i+1}) + L_2 x_{2,j})) \left(\left(\bar{\kappa}_{2,j} + \bar{\kappa}_{1,i} \frac{L_1}{L_2} (\mu_1 - \mu_2) \right) x_1 - \right. \\
 & \left. \frac{L_1}{L_2} (\mu_1 - \mu_2) \frac{x_1^2}{2} - x_{2,j} x_1 \right) \Big|_{\bar{\kappa}_{1,i}}^{\bar{\kappa}_{1,i}} + \delta(\mu_1 - \mu_2) L_1 \left((\bar{\kappa}_{2,j} + \frac{L_1}{L_2} (\mu_1 - \mu_2) \bar{\kappa}_{1,i}) \frac{x_1^2}{2} - \right. \\
 & \left. \frac{x_1^3}{3} \frac{L_1}{L_2} (\mu_1 - \mu_2) - x_{2,j} \frac{x_1^2}{2} \right) \Big|_{\bar{\kappa}_{1,i}}^{\bar{\kappa}_{1,i}} + \delta L_2 \left((\bar{\kappa}_{2,j} + \bar{\kappa}_{1,i} \frac{L_1}{L_2} (\mu_1 - \mu_2))^2 \frac{x_1}{2} - \right.
 \end{aligned}$$

$$\begin{aligned}
 & \left(\bar{\kappa}_{2,j} + \bar{\kappa}_{1,i} \frac{L_1}{L_2} (\mu_1 - \mu_2) \right) \frac{L_1}{L_2} (\mu_1 - \mu_2) \frac{x_1^2}{2} + \frac{L_1^2}{L_2^2} (\mu_1 - \mu_2)^2 \frac{x_1^3}{2} - x_{2,j+1}^2 \frac{x_1}{2} \Big|_{\bar{\kappa}_{1,i}}^{\bar{\kappa}_{1,i}} - \\
 & (\mu_1 f_{i+1,j+1} + \mu_2 f_{i,j+1} + \delta (L_2 x_{2,j+1} - L_1 (\mu_2 x_{1,i} - \mu_1 x_{1,i+1}))) \\
 & \left(\left(\bar{\kappa}_{2,j} + \bar{\kappa}_{1,i} \frac{L_1}{L_2} (\mu_1 - \mu_2) \right) x_1 - \frac{L_1}{L_2} (\mu_1 - \mu_2) \frac{x_1^2}{2} - x_{2,j+1} x_1 \right) \Big|_{\bar{\kappa}_{1,i}}^{\bar{\kappa}_{1,i}} - \delta (\mu_1 - \mu_2) \times \\
 & L_1 \left(x_{2,j+1} \frac{x_1^2}{2} - \left(\left(\bar{\kappa}_{2,j} + \bar{\kappa}_{1,i} \frac{L_1}{L_2} (\mu_1 - \mu_2) \right) \frac{x_1^2}{2} - \frac{L_1}{L_2} (\mu_1 - \mu_2) \frac{x_1^3}{3} \right) \right) \Big|_{\bar{\kappa}_{1,i}}^{\bar{\kappa}_{1,i}} - \\
 & \delta L_2 \left(x_{2,j+1}^2 \frac{x_1}{2} - \left(\left(\bar{\kappa}_{2,j} + \bar{\kappa}_{1,i} (\mu_1 - \mu_2) \frac{L_1}{L_2} \right)^2 \frac{x_1}{2} - \left(\bar{\kappa}_{2,j} + \bar{\kappa}_{1,i} \frac{L_1}{L_2} (\mu_1 - \mu_2) \right) \times \right. \right. \\
 & \left. \left. \frac{L_1}{L_2} (\mu_1 - \mu_2) \frac{x_1^2}{2} + \frac{L_1^2}{L_2^2} (\mu_1 - \mu_2) \frac{x_1^3}{6} \right) \right) \Big|_{\bar{\kappa}_{1,i}}^{\bar{\kappa}_{1,i}} = (\mu_1 (f_{i,j} - f_{i+1,j+1}) + \mu_2 (f_{i+1,j} - f_{i,j+1}) - \\
 & \delta (L_1 (\mu_1 - \mu_2) (x_{1,i} + x_{1,i+1}) + L_2 (x_{2,j} + x_{2,j+1}))) \left(\bar{\kappa}_{2,j} + \bar{\kappa}_{1,i} \frac{L_1}{L_2} (\mu_1 - \mu_2) \right) \times \\
 & (\bar{\kappa}_{1,i} - \bar{\bar{\kappa}}_{1,i}) - \frac{\bar{\kappa}_{1,i}^2 - \bar{\bar{\kappa}}_{1,i}^2}{2} \left(\frac{L_1}{L_2} ((\mu_1^2 - \mu_1 \mu_2) (f_{i,j} - f_{i+1,j+1}) + (\mu_1 \mu_2 - \mu_2^2) \times \right. \\
 & \left. (f_{i+1,j} - f_{i,j+1})) - \delta \frac{L_1^2}{L_2} (\mu_1 - \mu_2)^2 (x_{1,i} + x_{1,i+1}) - \delta L_1 (\mu_1 - \mu_2) (x_{2,j} + x_{2,j+1})) + \\
 & (\bar{\kappa}_{1,i} - \bar{\bar{\kappa}}_{1,i}) (x_{2,j+1} (\mu_1 f_{i+1,j+1} + \mu_2 f_{i+1,j} - \delta (L_1 (\mu_2 x_{1,i} - \mu_1 x_{1,i+1}) - L_2 x_{2,j+1})) - \\
 & x_{2,j} (\mu_1 f_{i,j} + \mu_2 f_{i+1,j} - \delta (L_1 (\mu_1 x_{1,i} - \mu_2 x_{1,i+1}) + L_2 x_{2,j})) - \delta L_1 (\mu_1 - \mu_2) \frac{(\bar{\kappa}_{1,i} - \bar{\bar{\kappa}}_{1,i})^2}{2} \times \\
 & (x_{2,j} + x_{2,j+1}) + \delta L_1 (\mu_1 - \mu_2) \bar{\kappa}_{2,j} x_1^2 \Big|_{\bar{\kappa}_{1,i}}^{\bar{\kappa}_{1,i}} + \frac{\delta L_1^2}{L_2} (\mu_1 - \mu_2)^2 \bar{\kappa}_{1,i} x_1^2 \Big|_{\bar{\kappa}_{1,i}}^{\bar{\kappa}_{1,i}} - \\
 & \frac{2}{3} \delta \frac{L_1^2}{L_2} (\mu_1 - \mu_2)^2 x_1^3 \Big|_{\bar{\kappa}_{1,i}}^{\bar{\kappa}_{1,i}} - \frac{\delta L_2}{2} (x_{2,j}^2 + x_{2,j+1}^2) (\bar{\kappa}_{1,i} - \bar{\bar{\kappa}}_{1,i}) + \delta L_2 \bar{\kappa}_{2,j}^2 x_1^2 \Big|_{\bar{\kappa}_{1,i}}^{\bar{\kappa}_{1,i}} + \\
 & \delta \bar{\kappa}_{1,i}^2 \frac{L_1^2}{L_2} (\mu_1 - \mu_2)^2 x_1 \Big|_{\bar{\kappa}_{1,i}}^{\bar{\kappa}_{1,i}} + 2 \delta L_1 (\mu_1 - \mu_2) \bar{\kappa}_{2,j} \bar{\kappa}_{1,i} x_1 \Big|_{\bar{\kappa}_{1,i}}^{\bar{\kappa}_{1,i}} - \delta L_1 (\mu_1 - \mu_2) \bar{\kappa}_{2,j} x_1^2 \Big|_{\bar{\kappa}_{1,i}}^{\bar{\kappa}_{1,i}} - \\
 & \delta \frac{L_1^2}{L_2} (\mu_1 - \mu_2)^2 \bar{\kappa}_{1,i} x_1^2 \Big|_{\bar{\kappa}_{1,i}}^{\bar{\kappa}_{1,i}} + \frac{\delta L_1^2}{L_2} \frac{x_1^3}{3} \Big|_{\bar{\kappa}_{1,i}}^{\bar{\kappa}_{1,i}} = (\bar{\kappa}_{1,i} - \bar{\bar{\kappa}}_{1,i}) ((\bar{\kappa}_{2,j} - x_{2,j}) (\mu_1 (f_{i,j} - \delta L_1 x_{1,i}) + \\
 & \mu_2 (f_{i+1,j} + \delta L_1 x_{1,i+1})) + (x_{2,j+1} - \bar{\kappa}_{2,j}) (\mu_1 (f_{i+1,j+1} + \delta L_1 x_{1,i+1}) + \mu_2 (f_{i,j+1} - \delta L_1 x_{1,i})) + \\
 & 2 \delta L_1 (\mu_1 - \mu_2) \bar{\kappa}_{1,i} \bar{\kappa}_{2,j} + \frac{\delta L_2}{2} (x_{2,j}^2 + x_{2,j+1}^2 + \bar{\kappa}_{2,j}^2) - \frac{\delta L_1^2}{L_2} \left(\frac{1}{3} (\bar{\kappa}_{1,i}^2 + \bar{\kappa}_{1,i} \bar{\bar{\kappa}}_{1,i}) - \bar{\bar{\kappa}}_{1,i}^2 \right). \quad (5.25)
 \end{aligned}$$

As a result of (5.24) and (5.25), from (5.23) we obtain that

$$\begin{aligned} \tilde{U}_p^\pm = \tilde{S}_p^\pm + \bar{\tilde{S}}_p^\pm &= (\bar{\kappa}_{1,i} - x_{1,i})(x_{2,j+1}f_{i,j+1} - x_{2,j}f_{i,j} + \delta h_2 L_1 \bar{\kappa}_{1,i}) + (x_{1,i+1} - \bar{\kappa}_{1,i}) \times \\ &(x_{2,j+1}f_{i+1,j+1} - x_{2,j}f_{i+1,j} + \delta h_2 L_1 \bar{\kappa}_{1,i}) + (\bar{\kappa}_{1,i} - \bar{\kappa}_{1,i}) ((\bar{\kappa}_{2,j} - x_{2,j})(\mu_1(f_{i,j} - \delta L_1 x_{1,i}) + \\ &\mu_2(f_{i+1,j} + \delta L_1 x_{1,i+1})) + (x_{2,j+1} - \bar{\kappa}_{2,j})(\mu_1(f_{i+1,j+1} + \delta L_1 x_{1,i+1}) + \mu_2(f_{i,j+1} - \delta L_1 x_{1,i})) + \\ &2\delta L_1(\mu_1 - \mu_2)\bar{\kappa}_{1,i}\bar{\kappa}_{2,j} + \frac{\delta L_2}{2}(x_{2,j}^2 + x_{2,j+1}^2 + \bar{\kappa}_{2,j}^2) - \frac{\delta L_1^2}{L_2}\left(\frac{1}{3}(\bar{\kappa}_{1,i}^2 + \bar{\kappa}_{1,i}\bar{\kappa}_{1,i}) - \bar{\kappa}_{1,i}^2\right)), \quad (5.26) \end{aligned}$$

where $p = 1, \dots, mm_1$. In the same way we can compute $\tilde{U}_p^\pm, p = 1, \dots, m$:

$$\begin{aligned} \tilde{U}_p^\pm &= \int_{x_{1,i}}^{\bar{\kappa}_{1,i}} \int_{x_{2,m_1+1}}^1 (f_{i,m_1+1} + \delta(L_1(x_1 - x_{1,i}) + L_2(x_2 - x_{2,m_1+1}))) dx_2 dx_1 + \\ &\int_{\bar{\kappa}_{1,i}}^{x_{1,i+1}} \int_{x_{2,m_1+1}}^1 (f_{i+1,m_1+1} + \delta(L_1(x_{1,i+1} - x_1) + L_2(x_2 - x_{2,m_1+1}))) dx_2 dx_1 = \\ &(f_{i,m_1+1} - \delta(L_1 x_{1,i} + L_2 x_{2,m_1+1}))(\bar{\kappa}_{1,i} - x_{1,i})(1 - x_{2,m_1+1}) + (f_{i+1,m_1+1} + \delta(L_1 x_{1,i+1} + \\ &L_2 x_{2,m_1+1}))(\bar{\kappa}_{1,i+1} - \bar{\kappa}_{1,i})(1 - x_{2,m_1+1}) + \delta\left(\frac{L_1}{2}((\bar{\kappa}_{1,i}^2 - x_{1,i}^2)(1 - x_{2,m_1+1}) - \right. \\ &\left.(x_{1,i+1}^2 - \bar{\kappa}_{1,i}^2)(1 - x_{2,m_1+1})) + \frac{L_2}{2}((1 - x_{2,m_1+1}^2)(\bar{\kappa}_{1,i} - x_{1,i}) + (1 - x_{2,m_1+1}^2) \right. \\ &\left.(x_{1,i+1} - \bar{\kappa}_{1,i}))\right) = (1 - x_{2,m_1+1})(f_{i,m_1+1}(\bar{\kappa}_{1,i} - x_{1,i}) + f_{i+1,m_1+1}(x_{1,i+1} - \bar{\kappa}_{1,i}) + \\ &\delta\left(L_1\left(\frac{1}{2}(x_{1,i}^2 + x_{1,i+1}^2) + \bar{\kappa}_{1,i}(\bar{\kappa}_{1,i} - x_{1,i} - x_{1,i+1})\right) + \frac{1}{2}L_2 h_1 \times (1 - x_{2,m_1+1})\right)). \quad (5.27) \end{aligned}$$

If we let $L_1 = L_2 = L, h_1 = h_2 = h$ in (5.26) we get an expression for $U_p^\pm, p = 1, \dots, m^2$. Therefore we have proved the following lemma.

Lemma 5.1 *Optimal-by-accuracy cubature formulae for computing the integral $I_1^2(f)$ in the classes $C_{2,L_1,L_2,N}^2$ and $C_{2,L,L,N}^2$ have the forms (5.11) and (5.10) respectively. The values of \tilde{U}_p^\pm ($p = 1, \dots, mm_1$) and \tilde{U}_p^\pm ($p = 1, \dots, m$) in (5.10), (5.11) are computed by formulae (5.26), (5.27) respectively and the value of U_p^\pm ($p = 1, \dots, m^2$) is computed by formula (5.26) for $L_1 = L_2 = L, h_1 = h_2 = h$. Error estimates of cubature formulae (5.10) and (5.11) are determined from the relationships (5.12).*

6 Optimal-By-Order Cubature Formulae for Computing Integrals with Fast Oscillatory Functions in Classes $C_{2,L_1,L_2,N}^2$ and $C_{2,L,L,N}^2$.

The approach described in Section 5 can be applied to the computation of integrals $I_2^2(f)$ and $I_3^2(f)$ in functional classes $C_{2,L_1,L_2,N}^2$ and $C_{2,L,L,N}^2$. In such cases cubature formulae can

be derived in explicit forms. In this section we consider a special case of the integral $I^2(f)$ when $\varphi_1(x_1) = \sin(\omega_1 x_1)$, $\varphi_2(x_2) = \sin(\omega_2 x_2)$, $|\omega_i| \geq 2\pi$, $i = 1, 2$.

Let $F_N = C_{2,L_1,L_2,N}^2$. As before, the splitting of the region \tilde{K}_p into sub-regions $\tilde{\Omega}_l^+, l = 1, 2, 3, 4, p = 1, \dots, mm_1$ is determined by points $O_1(\bar{\kappa}_{1,i}, \bar{\kappa}_{2,j})$, $O_2(\bar{\kappa}_{1,i}, \bar{\kappa}_{2,j})$ (similar to Fig. 1) with $\bar{\kappa}_{1,i}, \bar{\kappa}_{1,i}, \bar{\kappa}_{2,j}, \bar{\kappa}_{2,j}$ ($i = 1, \dots, m, j = 1, \dots, m_1$) computed by formulae (5.1)–(5.3). Analogously, the splitting of \tilde{K}_p into sub-region $\tilde{\Omega}_l^-, l = 1, 2, 3, 4, p = 1, \dots, mm_1$ is determined by points $O_3 = (\tilde{\kappa}_{1,i}, \tilde{\kappa}_{2,j})$, $O_4 = (\tilde{\kappa}_{1,i}, \tilde{\kappa}_{2,j})$ (similar to Fig. 1) with $\tilde{\kappa}_{1,i}, \tilde{\kappa}_{1,i}, \tilde{\kappa}_{2,j}, \tilde{\kappa}_{2,j}$ ($i = 1, \dots, m, j = 1, \dots, m_1$) computed by formulae (5.4)–(5.5). The splitting of the region \tilde{K}_p into subregions $\tilde{\Omega}_l^+, l = 1, 2$ is performed by the points $O_1(\hat{x}_{1,i}, x_{2,m_1+1}), O_2 = (\hat{x}_{1,i}, 1)$. Finally, the splitting of the region \tilde{K}_p into sub-regions $\tilde{\Omega}_l^-, l = 1, 2$ is performed by the points $O_3 = (\hat{x}_{1,i}, x_{2,m_1+1}), O_4 = (\hat{x}_{1,i}, 1)$ (see Fig. 2) with $\hat{x}_{1,i}, \hat{x}_{1,i}$ computed by formula (5.6). Let

$$\begin{aligned}\bar{T}_p^\pm &= \int \int_{\tilde{K}_p} A_{C_{2,L_1,L_2,N}^2}^\pm(X) \sin(\omega_1 x_1) \sin(\omega_2 x_2) dX, p = 1, \dots, mm_1, \\ \tilde{T}_p^\pm &= \int \int_{\tilde{K}_p} A_{C_{2,L_1,L_2,N}^2}^\pm(X) \sin(\omega_1 x_1) \sin(\omega_2 x_2) dX, p = 1, \dots, m.\end{aligned}\quad (6.1)$$

Taking into account Corollary 3.1 we conclude that the optimal-by-order cubature formula with constant not exceeding 2 (see also [20, 15]) for computing integral $I_2^2(f)$ in the class $C_{2,L_1,L_2,N}^2$ has the form

$$\bar{T}^* = \frac{1}{2} \left(\sum_{p=1}^{mm_1} (\bar{T}_p^+ + \bar{T}_p^-) + \sum_{p=1}^m (\tilde{T}_p^+ + \tilde{T}_p^-) \right), \quad (6.2)$$

therewith

$$\begin{aligned}v(C_{2,L_1,L_2,N}^2, \bar{T}^*, f) &\leq \frac{1}{2} \left(\sum_{p=1}^{mm_1} (\max(\bar{T}_p^+, \bar{T}_p^-) - \min(\bar{T}_p^+, \bar{T}_p^-)) + \sum_{p=1}^m (\max(\tilde{T}_p^+, \tilde{T}_p^-) - \min(\tilde{T}_p^+, \tilde{T}_p^-)) \right).\end{aligned}\quad (6.3)$$

By using Corollaries 2.2, 2.3 and by taking into account relationships (5.13)–(5.22) we obtain that $\bar{T}_p^\pm, p = 1, \dots, mm_1$ can be determined as follows

$$\begin{aligned}\bar{T}_p^\pm &= \int_{x_{1,i}}^{\bar{\kappa}_{1,i}} \int_{x_{2,j}}^{\bar{\kappa}_{2,j}} (f_{i,j} + \delta(L_1(x_1 - x_{1,i}) + L_2(x_2 - x_{2,j}))) \sin(\omega_1 x_1) \sin(\omega_2 x_2) dx_2 dx_1 + \\ &\quad \int_{x_{1,i}}^{\bar{\kappa}_{1,i}} \int_{\bar{\kappa}_{2,j}}^{x_{2,j+1}} (f_{i,j+1} + \delta(L_1(x_1 - x_{1,i}) + L_2(x_{2,j+1} - x_2))) \sin(\omega_1 x_1) \sin(\omega_2 x_2) dx_2 dx_1 + \\ &\quad \int_{\bar{\kappa}_{1,i}}^{x_{1,i+1}} \int_{\bar{\kappa}_{2,j}}^{x_{2,j+1}} (f_{i+1,j+1} + \delta(L_1(x_{1,i+1} - x_1) + L_2(x_{2,j+1} - x_2))) \sin(\omega_1 x_1) \sin(\omega_2 x_2) dx_2 dx_1 + \\ &\quad \int_{\bar{\kappa}_{1,i}}^{x_{1,i+1}} \int_{x_{2,j}}^{\bar{\kappa}_{2,j}} (f_{i+1,j} + \delta(L_1(x_{1,i+1} - x_1) + L_2(x_2 - x_{2,j}))) \sin(\omega_1 x_1) \sin(\omega_2 x_2) dx_2 dx_1 +\end{aligned}$$

$$\begin{aligned} & \int_{\tilde{\kappa}_{1,i}}^{\tilde{\kappa}_{1,i}} \int_{\tilde{\kappa}_{2,j}}^{\tilde{\kappa}_{2,j} + \frac{L_1}{L_2}(\mu_1 - \mu_2)(\tilde{\kappa}_{1,i} - x_1)} (\mu_1 f_{1,j} + \mu_2 f_{i+1,j} + \delta(L_1((\mu_1 - \mu_2)x_1 - \mu_1 x_{1,i} + \mu_2 x_{1,i+1}) + \\ & L_2(x_2 - x_{2,j}))) \sin(\omega_1 x_1) \sin(\omega_2 x_2) dx_2 dx_1 + \int_{\tilde{\kappa}_{1,i}}^{\tilde{\kappa}_{1,i}} \int_{\tilde{\kappa}_{2,j} + \frac{L_1}{L_2}(\mu_1 - \mu_2)(\tilde{\kappa}_{1,i} - x_1)}^{x_{2,j+1}} (\mu_1 f_{i+1,j+1} + \mu_2 f_{i,j+1} + \\ & \delta(L_1((\mu_2 - \mu_1)x_1 - \mu_2 x_{1,i} + \mu_1 x_{1,i+1}) + L_2(x_2 - x_{2,j}))) \sin(\omega_1 x_1) \sin(\omega_2 x_2) dx_2 dx_1. \quad (6.4) \end{aligned}$$

Setting $\delta = 1$ in (6.4) we get the expression for \bar{T}_p^+ . Similarly, setting $\delta = -1$ in (6.4) gives the expression for \bar{T}_p^- , $p = 1, \dots, mm_1$. The sum \bar{S}_p^\pm , $p = 1, \dots, mm_1$ of the first four integrals in (6.4) can be computed explicitly:

$$\begin{aligned}
\bar{S}_p^{\pm} = & \frac{\cos \omega_1 \bar{\kappa}_{1,i} - \cos \omega_1 x_{1,i}}{\omega_1 \omega_2} \left(f_{i,j} (\cos \omega_2 \bar{\kappa}_{2,j} - \cos \omega_2 x_{2,j}) + f_{i,j+1} (\cos \omega_2 x_{2,j+1} - \cos \omega_2 \bar{\kappa}_{2,j}) \right) + \delta L_1 x_{1,i} (\cos \omega_2 x_{2,j} - \cos \omega_2 x_{2,j+1}) + \delta L_2 \left(x_{2,j} (\cos \omega_2 x_{2,j} - \cos \omega_2 \bar{\kappa}_{2,j}) + x_{2,j+1} (\cos \omega_2 x_{2,j+1} - \cos \omega_2 \bar{\kappa}_{2,j}) \right) + \\
& \frac{\cos \omega_1 x_{1,i+1} - \cos \omega_1 \bar{\kappa}_{1,i}}{\omega_1 \omega_2} \left(f_{i+1,j} (\cos \omega_2 \bar{\kappa}_{2,j} - \cos \omega_2 x_{2,j}) + f_{i+1,j+1} (\cos \omega_2 x_{2,j+1} - \cos \omega_2 \bar{\kappa}_{2,j}) \right) + \delta L_1 x_{1,i+1} (\cos \omega_2 x_{2,j+1} - \cos \omega_2 x_{2,j}) + \delta L_2 \left(x_{2,j} (\cos \omega_2 x_{2,j} - \cos \omega_2 \bar{\kappa}_{2,j}) + x_{2,j+1} (\cos \omega_2 x_{2,j+1} - \cos \omega_2 \bar{\kappa}_{2,j}) \right) + \\
& \frac{\delta}{\omega_1 \omega_2} \left(L_1 (\cos \omega_2 x_{2,j} - \cos \omega_2 x_{2,j+1}) (x_{1,i} \cos \omega_1 x_{1,i} + x_{1,i+1} \cos \omega_1 x_{1,i+1} - \bar{\kappa}_{1,i} \cos \omega_1 \bar{\kappa}_{1,i}) + \right. \\
& \bar{\kappa}_{1,i} \cos \omega_1 \bar{\kappa}_{1,i} + \frac{1}{\omega_1} (\sin \omega_1 \bar{\kappa}_{1,i} + \sin \omega_1 \bar{\kappa}_{1,i} - \sin \omega_1 x_{1,i} - \sin \omega_1 x_{1,i+1}) \Big) + \\
& L_2 (\cos \omega_1 x_{1,i} - \cos \omega_1 \bar{\kappa}_{1,i}) \left(x_{2,j} \cos \omega_2 x_{2,j} + x_{2,j+1} \cos \omega_2 x_{2,j+1} - 2 \bar{\kappa}_{2,j} \cos \omega_2 \bar{\kappa}_{2,j} + \right. \\
& \frac{1}{\omega_2} (2 \sin \omega_2 \bar{\kappa}_{2,j} - \sin \omega_2 x_{2,j} - \sin \omega_2 x_{2,j+1}) \Big) + (\cos \omega_1 \bar{\kappa}_{1,i} - \cos \omega_1 x_{1,i+1}) \times \\
& (x_{2,j} \cos \omega_2 x_{2,j} + x_{2,j+1} \cos \omega_2 x_{2,j+1} - 2 \bar{\kappa}_{2,j} \cos \omega_2 \bar{\kappa}_{2,j} + \frac{1}{\omega_2} (2 \sin \omega_2 \bar{\kappa}_{2,j} - \sin \omega_2 x_{2,j} - \\
& \sin \omega_2 x_{2,j+1})) = \frac{1}{\omega_1 \omega_2} \left((\cos \omega_1 \bar{\kappa}_{1,i} - \cos \omega_1 x_{1,i}) (f_{i,j} (\cos \omega_2 \bar{\kappa}_{2,j} - \cos \omega_2 x_{2,j}) + \right. \\
& f_{i,j+1} (\cos \omega_2 x_{2,j+1} - \cos \omega_2 \bar{\kappa}_{2,j})) + (\cos \omega_1 x_{1,i+1} - \cos \omega_1 \bar{\kappa}_{1,i}) (f_{i+1,j} (\cos \omega_2 \bar{\kappa}_{2,j} - \cos \omega_2 x_{2,j}) + \\
& f_{i+1,j+1} (\cos \omega_2 x_{2,j+1} - \cos \omega_2 \bar{\kappa}_{2,j})) - \delta (L_1 ((\cos \omega_2 x_{2,j} - \cos \omega_2 x_{2,j+1}) \\
& (\cos \omega_1 \bar{\kappa}_{1,i} (x_{1,i+1} - \bar{\kappa}_{1,i}) + \cos \omega_1 \bar{\kappa}_{1,i} (x_{1,i} - \bar{\kappa}_{1,i}) + \frac{1}{\omega_1} (\sin \omega_1 \bar{\kappa}_{1,i} + \sin \omega_1 \bar{\kappa}_{1,i} - \\
& \sin \omega_1 x_{1,i} - \sin \omega_1 x_{1,i+1}))) + L_2 ((\cos \omega_1 x_{1,i} - \cos \omega_1 \bar{\kappa}_{1,i}) (\cos \omega_2 x_{2,j} (x_{2,j} + x_{2,j+1} - \right.
\end{aligned}$$

$$2 \bar{\bar{\kappa}}_{2,j}) + \frac{1}{\omega_2} (2 \sin \omega_2 \bar{\kappa}_{2,j} - \sin \omega_2 x_{2,j} - \sin \omega_2 x_{2,j+1}) + (\cos \omega_1 \bar{\kappa}_{1,i} - \cos \omega_1 x_{1,i+1}) \\ (\cos \omega_2 \bar{\kappa}_{2,j} (x_{2,j} + x_{2,j+1} - 2 \bar{\kappa}_{2,j}) + \frac{1}{\omega_2} (2 \sin \omega_2 \bar{\kappa}_{2,j} - \sin \omega_2 x_{2,j} - \sin \omega_2 x_{2,j+1})) \Big) \Big) \Big) \quad (6.5)$$

We can also explicitly compute the sum $\bar{\bar{S}}_p^{\pm}$, $p = 1, \dots, mm_1$ of the last two integrals in (6.4):

$$\begin{aligned} \bar{\bar{S}}_p^{\pm} = & (\mu_1 f_{i,j} + \mu_2 f_{i+1,j} + \delta(L_1(\mu_2 x_{1,i+1} - \mu_1 x_{1,i}) - L_2 x_{2,j})) \frac{1}{\omega_2} \int_{\bar{\kappa}_{1,i}}^{\bar{\kappa}_{1,i}} \sin \omega_1 x_1 \times \\ & \left(\cos \omega_2 x_{2,j} - \cos \omega_2 (\bar{\kappa}_{2,j} + \frac{L_1}{L_2}(\mu_1 - \mu_2)(\bar{\kappa}_{1,i} - x_1)) \right) dx_1 + (\mu_1 f_{i+1,j+1} + \mu_2 f_{i,j+1} + \\ & \delta(L_1(\mu_1 x_{1,i+1} - \mu_2 x_{1,i}) - L_2 x_{2,j+1})) \frac{1}{\omega_2} \int_{\bar{\kappa}_{1,i}}^{\bar{\kappa}_{1,i}} \sin \omega_1 x_1 \left(\cos \omega_2 \left(\bar{\kappa}_{2,j} + \frac{L_1}{L_2}(\mu_1 - \mu_2) \right. \right. \\ & \left. \left. (\bar{\kappa}_{1,i} - x_1) \right) - \cos \omega_2 x_{2,j+1} \right) dx_1 + \delta L_1(\mu_1 - \mu_2) \left(\int_{\bar{\kappa}_{1,i}}^{\bar{\kappa}_{1,i}} \int_{x_{2,j}}^{\bar{\kappa}_{2,j} + \frac{L_1}{L_2}(\mu_1 - \mu_2)(\bar{\kappa}_{1,i} - x_1)} x_1 \times \right. \\ & \left. \sin \omega_1 x_1 \sin \omega_2 x_2 dx_2 dx_1 - \int_{\bar{\kappa}_{1,i}}^{\bar{\kappa}_{1,i}} \int_{x_{2,j} + \frac{L_1}{L_2}(\mu_1 - \mu_2)(\bar{\kappa}_{1,i} - x_1)}^{x_{2,j+1}} x_1 \sin \omega_1 x_1 \sin \omega_2 x_2 dx_2 dx_1 \right) + \\ & \delta L_2 \left(\int_{\bar{\kappa}_{1,i}}^{\bar{\kappa}_{1,i}} \int_{x_{2,j}}^{\bar{\kappa}_{2,j} + \frac{L_1}{L_2}(\mu_1 - \mu_2)(\bar{\kappa}_{1,i} - x_1)} x_2 \sin \omega_1 x_1 \sin \omega_2 x_2 dx_2 dx_1 + \right. \\ & \left. \int_{\bar{\kappa}_{1,i}}^{\bar{\kappa}_{1,i}} \int_{x_{2,j} + \frac{L_1}{L_2}(\mu_1 - \mu_2)(\bar{\kappa}_{1,i} - x_1)}^{x_{2,j+1}} x_2 \sin \omega_1 x_1 \sin \omega_2 x_2 dx_2 dx_1 \right) = \frac{1}{\omega_1 \omega_2} \left(\cos \omega_1 \bar{\bar{\kappa}}_{1,i} - \right. \\ & \left. \cos \omega_1 \bar{\kappa}_{1,i} (\mu_1 f_{i,j} + \mu_2 f_{i+1,j} + \delta(L_1(\mu_2 x_{1,i+1} - \mu_1 x_{1,i}) - L_2 x_{2,j})) \cos \omega_2 x_{2,j} - \right. \\ & \left. (\mu_1 f_{i+1,j+1} + \mu_2 f_{i,j+1} + \delta(L_1(\mu_1 x_{1,i+1} - \mu_2 x_{1,i}) + L_2 x_{2,j+1})) \cos \omega_2 x_{2,j+1} \right) + \\ & \frac{1}{2\omega_2} \left(\frac{1}{\omega_1 + \omega_2 \frac{L_1}{L_2}(\mu_1 - \mu_2)} \times (\cos(\omega_2 \bar{\kappa}_{2,j} - \omega_1 \bar{\kappa}_{1,i}) - \cos(\omega_2 \bar{\bar{\kappa}}_{2,j} - \omega_1 \bar{\bar{\kappa}}_{1,i})) + \right. \\ & \left. \frac{1}{\omega_1 - \omega_2 \frac{L_1}{L_2}(\mu_1 - \mu_2)} (\cos(\omega_2 \bar{\kappa}_{2,j} + \omega_1 \bar{\kappa}_{1,i}) - \cos(\omega_2 \bar{\bar{\kappa}}_{2,j} + \omega_1 \bar{\bar{\kappa}}_{1,i})) \right) (\mu_1 (f_{1,j} - \\ & f_{i+1,j+1}) + \mu_2 (f_{i+1,j} - f_{i,j+1}) - \delta(L_1(\mu_1 - \mu_2)(x_{1,i} + x_{1,i+1}) + L_2(x_{2,j+1} + x_{2,j}) + \\ & \delta L_1(\mu_1 - \mu_2) \left(\frac{1}{\omega_1 \omega_2} (\cos \omega_2 x_{2,j+1} + \cos \omega_2 x_{2,j}) \left(\frac{1}{\omega_1} (\sin \omega_1 \bar{\kappa}_{1,i} - \sin \omega_1 \bar{\bar{\kappa}}_{1,i}) - \right. \right. \\ & \left. \left. \bar{\kappa}_{1,i} \cos \omega_1 \bar{\kappa}_{1,i} + \bar{\bar{\kappa}}_{1,i} \cos \omega_1 \bar{\bar{\kappa}}_{1,i} \right) - \frac{2}{\omega_2} \int_{\bar{\kappa}_{1,i}}^{\bar{\kappa}_{1,i}} x_1 \sin \omega_1 x_1 \cos \omega_2 \left(\bar{\kappa}_{2,j} + \frac{L_1}{L_2}(\mu_1 - \mu_2) \times \right. \right. \\ & \left. \left. \bar{\kappa}_{1,i} \cos \omega_1 \bar{\kappa}_{1,i} + \bar{\bar{\kappa}}_{1,i} \cos \omega_1 \bar{\bar{\kappa}}_{1,i} \right) \right) \right) \end{aligned}$$

$$\begin{aligned}
 & (\bar{\kappa}_{1,i} - x_1)) dx_1) + \delta L_2 \left(\frac{1}{\omega_1 \omega_2} (\cos \omega_1 \bar{\bar{\kappa}}_{1,i} - \cos \omega_1 \bar{\kappa}_{1,i}) (x_{2,j+1} \cos \omega_2 x_{2,j+1} + x_{2,j} \times \right. \\
 & \left. \cos \omega_2 x_{2,j} - \frac{1}{\omega_2} (\sin \omega_2 x_{2,j+1} + \sin \omega_2 x_{2,j})) \right) + \frac{2}{\omega_2} \left(\frac{L_1}{L_2} (\mu_1 - \mu_2) \int_{\bar{\kappa}_{1,i}}^{\bar{\kappa}_{1,i}} x_1 \sin \omega_1 x_1 \times \right. \\
 & \cos \omega_2 (\bar{\kappa}_{2,j} + \frac{L_1}{L_2} (\mu_1 - \mu_2) (\bar{\kappa}_{1,i} - x_1)) dx_1 - (\bar{\kappa}_{2,j} + \frac{L_1}{L_2} (\mu_1 - \mu_2) \bar{\kappa}_{1,i}) \times \\
 & \left. \int_{\bar{\kappa}_{1,i}}^{\bar{\kappa}_{1,i}} \sin \omega_1 x_1 \cos \omega_2 \left(\bar{\kappa}_{2,j} + \frac{L_1}{L_2} (\mu_1 - \mu_2) \times (\bar{\kappa}_{1,i} - x_1) \right) dx_1 \right) + \frac{1}{\omega_2} (\sin \omega_2 (\bar{\kappa}_{2,j} + \\
 & \left. \frac{L_1}{L_2} (\mu_1 - \mu_2) \bar{\kappa}_{1,i}) \int_{\bar{\kappa}_{1,i}}^{\bar{\kappa}_{1,i}} \sin \omega_1 x_1 \cos \omega_2 (\frac{L_1}{L_2} (\mu_1 - \mu_2) x_1) dx_1 - \cos \omega_2 (\bar{\kappa}_{2,j} + \right. \\
 & \left. \frac{L_1}{L_2} (\mu_1 - \mu_2) \bar{\kappa}_{1,i}) \int_{\bar{\kappa}_{1,i}}^{\bar{\kappa}_{1,i}} \sin \omega_1 x_1 \sin \omega_2 (\frac{L_1}{L_2} (\mu_1 - \mu_2) x_1) dx_1 \right) \left. \right) = \\
 & \frac{\cos \omega_1 \bar{\bar{\kappa}}_{1,i} - \cos \omega_1 \bar{\kappa}_{1,i}}{\omega_1 \omega_2} ((\mu_1 f_{i,j} + \mu_2 f_{i+1,j} + \delta L_1 (\mu_1 x_{1,i+1} - \mu_1 x_{1,i})) \cos \omega_2 x_{2,j} - \\
 & (\mu_1 f_{i+1,j+1} + \mu_2 f_{i,j+1} + \delta L_1 (\mu_1 x_{1,i+1} - \mu_2 x_{1,i})) \cos \omega_2 x_{2,j+1} - \frac{\delta L_2}{\omega_2} (\sin \omega_2 x_{2,j+1} + \\
 & \sin \omega_2 x_{2,j})) + \frac{1}{2\omega_2} \left(\frac{1}{\omega_1 + \omega_2 \frac{L_1}{L_2} (\mu_1 - \mu_2)} ((\cos(\omega_2 \bar{\kappa}_{2,j} - \omega_1 \bar{\kappa}_{1,i}) - \cos(\omega_2 \bar{\bar{\kappa}}_{2,j} - \right. \\
 & \left. \omega_1 \bar{\bar{\kappa}}_{1,i})) + \frac{1}{\omega_1 - \omega_2 \frac{L_1}{L_2} (\mu_1 - \mu_2)} (\cos(\omega_2 \bar{\kappa}_{2,j} + \omega_1 \bar{\kappa}_{1,i}) - \cos(\omega_2 \bar{\bar{\kappa}}_{2,j} + \omega_1 \bar{\bar{\kappa}}_{1,i})) \right) \times \\
 & (\mu_1 (f_{i,j} - f_{i+1,j+1}) + \mu_2 (f_{i+1,j} - f_{i,j+1}) + \delta (L_1 (\mu_1 - \mu_2) (2\bar{\kappa}_{1,i} - x_{1,i+1} - x_{1,i}) + \\
 & L_2 (2\bar{\kappa}_{2,j} - x_{2,j+1} - x_{2,j}))) + \frac{\delta L_1 (\mu_1 - \mu_2)}{\omega_1 \omega_2} (\cos \omega_2 x_{2,j+1} + \cos \omega_2 x_{2,j}) \times \\
 & \left(\frac{1}{\omega_1} (\sin \omega_1 \bar{\kappa}_{1,i} - \sin \omega_1 \bar{\bar{\kappa}}_{1,i}) - \bar{\kappa}_{1,i} \cos \omega_1 \bar{\kappa}_{1,i} + \bar{\bar{\kappa}}_{1,i} \cos \omega_1 \bar{\bar{\kappa}}_{1,i} \right) - \\
 & \frac{\delta L_2}{\omega_2^2} ((\sin(\omega_2 \bar{\kappa}_{2,j} - \omega_1 \bar{\kappa}_{1,j}) - \sin(\omega_2 \bar{\bar{\kappa}}_{2,j} - \omega_1 \bar{\bar{\kappa}}_{1,i})) \frac{1}{\omega_1 + \omega_2 \frac{L_1}{L_2} (\mu_1 - \mu_2)} + \\
 & \frac{1}{\omega_1 - \omega_2 \frac{L_1}{L_2} (\mu_1 - \mu_2)} (\sin(\omega_2 \bar{\kappa}_{2,j} + \omega_1 \bar{\kappa}_{1,i}) - \sin(\omega_2 \bar{\bar{\kappa}}_{2,j} + \omega_1 \bar{\bar{\kappa}}_{1,i})) \right). \tag{6.6}
 \end{aligned}$$

By combining the results obtained in (6.5) and (6.6), we get that

$$\bar{T}_p^{\pm} = \sum_{k=1}^4 \bar{W}_k^{\pm}, \quad p = 1, \dots, mm_1; \quad (6.7)$$

where

$$\begin{aligned} \bar{W}_1^{\pm} &= \frac{1}{\omega_1 \omega_2} \left((\cos \omega_1 \bar{\kappa}_{1,i} - \cos \omega_1 x_{1,i}) (f_{i,j} (\cos \omega_2 \bar{\kappa}_{2,j} - \cos \omega_2 x_{2,j}) + f_{i,j+1} \times \right. \\ &\quad \left. (\cos \omega_2 x_{2,j+1} - \cos \omega_2 \bar{\kappa}_{2,j})) + (\cos \omega_1 x_{1,i+1} - \cos \omega_1 \bar{\kappa}_{1,i}) (f_{i+1,j} (\cos \omega_2 \bar{\kappa}_{2,j} - \right. \\ &\quad \left. \cos \omega_2 x_{2,j}) + f_{i+1,j+1} (\cos \omega_2 x_{2,j+1} - \cos \omega_2 \bar{\kappa}_{2,j})) + (\cos \omega_1 \bar{\kappa}_{1,i} - \cos \omega_1 \bar{\kappa}_{1,i}) \times \right. \\ &\quad \left. ((\mu_1 f_{i,j} + \mu_2 f_{i+1,j}) \cos \omega_2 x_{2,j} - (\mu_1 f_{i+1,j+1} + \mu_2 f_{i,j+1}) \cos \omega_2 x_{2,j+1})) \right); \end{aligned} \quad (6.8)$$

$$\begin{aligned} \bar{W}_2^{\pm} &= \frac{\delta L_1}{\omega_1 \omega_2} ((\cos \omega_2 x_{2,j} - \cos \omega_2 x_{2,j+1}) (x_{1,i+1} \cos \omega_1 \bar{\kappa}_{1,i} + x_{1,i} \cos \omega_1 \bar{\kappa}_{1,i} - \\ &\quad \frac{1}{\omega_1} (\sin \omega_1 x_{1,i} + \sin \omega_1 x_{1,i+1})) + (\frac{\sin \omega_1 \bar{\kappa}_{1,i}}{\omega_1} - \bar{\kappa}_{1,i} \cos \omega_1 \bar{\kappa}_{1,i}) ((1 + (\mu_1 - \mu_2)) \times \\ &\quad \cos \omega_2 x_{2,j} - (1 - (\mu_1 - \mu_2)) \cos \omega_2 x_{2,j+1}) + (\frac{\sin \omega_1 \bar{\kappa}_{1,i}}{\omega_1} - \bar{\kappa}_{1,i} \cos \omega_1 \bar{\kappa}_{1,i}) ((1 - \\ &\quad (\mu_1 - \mu_2)) \cos \omega_2 x_{2,j} - (1 + (\mu_1 - \mu_2)) \cos \omega_2 x_{2,j+1})); \end{aligned} \quad (6.9)$$

$$\begin{aligned} \bar{W}_3^{\pm} &= \frac{\delta L_2}{\omega_1 \omega_2} ((\cos \omega_1 x_{1,i} - \cos \omega_1 \bar{\kappa}_{1,i}) ((x_{2,j} + x_{2,j+1} - 2 \bar{\kappa}_{2,j}) \cos \omega_2 \bar{\kappa}_{2,j} + \\ &\quad \frac{2}{\omega_2} \sin \omega_2 \bar{\kappa}_{2,j}) + (\cos \omega_1 \bar{\kappa}_{1,i} - \cos \omega_1 x_{1,i+1}) ((x_{2,j} + x_{2,j+1} - 2 \bar{\kappa}_{2,j}) \cos \omega_2 \bar{\kappa}_{2,j} + \\ &\quad \frac{2}{\omega_2} \sin \omega_2 \bar{\kappa}_{2,j}) + \frac{1}{\omega_2} (\cos \omega_1 x_{1,i+1} - \cos \omega_1 x_{1,i}) (\sin \omega_2 x_{2,j} + \sin \omega_2 x_{2,j+1})); \end{aligned} \quad (6.10)$$

$$\begin{aligned} \bar{W}_4^{\pm} &= \frac{1}{\omega_2} \left(\left(\frac{\cos(\omega_2 \bar{\kappa}_{2,j} - \omega_1 \bar{\kappa}_{1,i}) - \cos(\omega_2 \bar{\kappa}_{2,j} - \omega_1 \bar{\kappa}_{1,i})}{2(\omega_1 + \omega_2 \frac{L_1}{L_2}(\mu_1 - \mu_2))} \right. \right. \\ &\quad \left. \left. + \frac{\cos(\omega_2 \bar{\kappa}_{2,j} + \omega_1 \bar{\kappa}_{1,i}) - \cos(\omega_2 \bar{\kappa}_{2,j} + \omega_1 \bar{\kappa}_{1,i})}{2(\omega_1 - \omega_2 \frac{L_1}{L_2}(\mu_1 - \mu_2))} (\mu_1 (f_{i,j} - f_{i+1,j+1}) + \right. \right. \\ &\quad \left. \left. \mu_2 (f_{i+1,j} - f_{i,j+1}) + \delta(L_1(\mu_1 - \mu_2)(2\bar{\kappa}_{1,i} - x_{1,i+1} - x_{1,i}) + L_2(2\bar{\kappa}_{2,j} - \right. \right. \\ &\quad \left. \left. x_{2,j} - x_{2,j+1})) - \frac{\delta L_2}{\omega_2} \left(\frac{\sin(\omega_2 \bar{\kappa}_{2,j} - \omega_1 \bar{\kappa}_{1,i}) - \sin(\omega_2 \bar{\kappa}_{2,j} - \omega_1 \bar{\kappa}_{1,i})}{\omega_1 + \omega_2 \frac{L_1}{L_2}(\mu_1 - \mu_2)} + \right. \right. \\ &\quad \left. \left. \frac{\sin(\omega_2 \bar{\kappa}_{2,j} - \omega_1 \bar{\kappa}_{1,i}) - \sin(\omega_2 \bar{\kappa}_{2,j} - \omega_1 \bar{\kappa}_{1,i})}{\omega_1 - \omega_2 \frac{L_1}{L_2}(\mu_1 - \mu_2)} \right) \right). \end{aligned} \quad (6.11)$$

In a similar way we compute $\tilde{T}_p^\pm, p = 1, \dots, m$:

$$\begin{aligned}
 \tilde{T}_p^\pm &= \int_{x_{1,i}}^{\tilde{x}_{1,i}} \int_{x_{2,m_1+1}}^1 (f_{i,m_1+1} + \delta(L_1(x_1 - x_{1,i}) + L_2(x_2 - x_{2,m_1+1}))) \sin \omega_1 x_1 \times \\
 &\quad \sin \omega_2 x_2 dx_2 dx_1 + \int_{\tilde{x}_{1,i}}^{x_{1,i+1}} \int_{x_{2,m_1+1}}^1 (f_{i+1,m_1+1} + \delta(L_1(x_{1,i+1} - x_1) + L_2(x_2 - x_{2,m_1+1}))) \times \\
 &\quad \sin \omega_1 x_1 \sin \omega_2 x_2 dx_2 dx_1 = \frac{1}{\omega_1 \omega_2} ((f_{i,m_1+1} - \delta(L_1 x_{1,i} + L_2 x_{2,m_1+1})) (\cos \omega_1 \tilde{x}_{1,i} - \cos \omega_1 x_{1,i}) \\
 &\quad (\cos \omega_2 - \cos \omega_2 x_{2,m_1+1}) + (f_{i+1,m_1+1} + \delta(L_1 x_{1,i+1} - L_2 x_{2,m_1+1})) (\cos \omega_1 x_{1,i+1} - \cos \omega_1 x_{1,i}) \\
 &\quad (\cos \omega_2 - \cos \omega_2 x_{2,m_1+1})) + \delta \left(\frac{L_1}{\omega_1 \omega_2} (\cos \omega_2 x_{2,m_1+1} - \cos \omega_2) \left(\frac{1}{\omega_1} (\sin \omega_1 \tilde{x}_{1,i} - \sin \omega_1 x_{1,i}) - \right. \right. \\
 &\quad \left. \left. \tilde{x}_{1,i} \cos \omega_1 \tilde{x}_{1,i} + x_{1,i} \cos \omega_1 x_{1,i} - \frac{1}{\omega_1} (\sin \omega_1 x_{1,i+1} - \sin \omega_1 \tilde{x}_{1,i}) + x_{1,i+1} \cos \omega_1 x_{1,i+1} - \right. \right. \\
 &\quad \left. \left. \tilde{x}_{1,i} \cos \omega_1 \tilde{x}_{1,i} + \frac{L_2}{\omega_1 \omega_2} \left(\frac{1}{\omega_2} (\sin \omega_2 - \sin \omega_2 x_{2,m_1+1}) - \cos \omega_2 + x_{2,m_1+1} \cos \omega_2 x_{2,m_1+1} \right) \times \right. \right. \\
 &\quad \left. \left. (\cos \omega_1 x_{1,i} - \cos \omega_1 \tilde{x}_{1,i} + \cos \omega_1 \tilde{x}_{1,i} - \cos \omega_1 x_{1,i+1}) \right) = \frac{1}{\omega_1 \omega_2} ((\cos \omega_2 - \cos \omega_2 x_{2,m_1+1}) \times \right. \\
 &\quad \left. (f_{i,m_1+1} (\cos \omega_1 \tilde{x}_{1,i} - \cos \omega_1 x_{1,i}) + f_{i+1,m_1+1} (\cos \omega_1 x_{1,i+1} - \cos \omega_1 \tilde{x}_{1,i}) + \delta L_1 \times \right. \\
 &\quad \left. \left((2\tilde{x}_{1,i} - x_{1,i} - x_{1,i+1}) \cos \omega_1 \tilde{x}_{1,i} + \frac{1}{\omega_1} (\sin \omega_1 x_{1,i} + \sin \omega_1 x_{1,i+1} - 2 \sin \omega_1 \tilde{x}_{1,i}) \right) \right) + \delta L_2 \times \\
 &\quad \left. (\cos \omega_1 x_{1,i} - \cos \omega_1 x_{1,i+1}) ((x_{2,m_1+1} - 1) \cos \omega_2 + \frac{1}{\omega_2} (\sin \omega_2 - \sin \omega_2 x_{2,m_1+1})) \right). \quad (6.12)
 \end{aligned}$$

By setting $\delta = 1$ in (6.12) we get the explicit expression for \tilde{T}_p^+ , and by setting $\delta = -1$ we obtain the explicit expression for $\tilde{T}_p^-, p = 1, \dots, m$. Therefore we have proved the following lemma.

Lemma 6.1 *Optimal-by-order (with constant not exceeding 2) cubature formulae for computing the integral $I_2^2(f)$ in the class $C_{2,L_1,L_2,N}^2$ have the form (6.2) with \tilde{T}_p^\pm ($p = 1, \dots, mm_1$) and \tilde{T}_p^\pm ($p = 1, \dots, m$) computed by formulae (6.5)–(6.11) and (6.12) respectively. The error estimate of cubature formulae (6.2) is defined by the relationship (6.3).*

Now let $F_N = C_{2,L,L,N}^2$. As it was mentioned in Section 5, the splitting of K_p into regions $\Omega_l^+, l = 1, 2, 3, 4$ is determined by points $O_1(\bar{x}_{1,i}, \bar{x}_{2,j})$, $O_2 = (\bar{x}_{1,i}, \bar{x}_{2,j})$, and the splitting of K_p into regions $\Omega_l^-, l = 1, 2, 3, 4$ is determined by points $O_3 = (\tilde{x}_{1,i}, \tilde{x}_{2,j})$, $O_4 = (\tilde{x}_{1,i}, \tilde{\tilde{x}}_{2,j})$ (see Fig. 1), where $\bar{x}_{1,i}, \bar{\bar{x}}_{1,i}, \tilde{x}_{1,i}, \tilde{\tilde{x}}_{1,i}$ ($i = 1, \dots, m$) and $\bar{x}_{2,j}, \bar{\bar{x}}_{2,j}, \tilde{x}_{2,j}, \tilde{\tilde{x}}_{2,j}$ ($j = 1, \dots, m, p = 1, \dots, m^2$) are computed by formulae (5.1)–(5.5) respectively for $L_1 = L_2 = L$ and $h_1 = h_2 = h$. Let

$$T_p^\pm = \int \int_{K_p} A_{C_{2,L,L,N}^2}^\pm(X) \sin \omega_1 x_1 \sin \omega_2 x_2 dX, \quad p = 1, \dots, m^2. \quad (6.13)$$

By taking into account Corollary 3.1, we obtain that optimal-by-order (with constant not exceeding 2) cubature formulae for computing integrals $I_2^2(f)$ in class $C_{2,L,L,N}^2$ have the form

$$T^* = \frac{1}{2} \sum_{p=1}^{m^2} (T_p^+ + T_p^-), \quad (6.14)$$

therewith

$$v(C_{2,L,L,N}^2, T^*, f) \leq \frac{1}{2} \sum_{p=1}^{m^2} (\max(T_p^+, T_p^-) - \min(T_p^+, T_p^-)). \quad (6.15)$$

By setting $L_1 = L_2 = L$ and $h_1 = h_2 = h$ in (6.7) we obtain the expression for $T_p^\pm, p = 1, \dots, m^2$. This leads to the following result.

Lemma 6.2 *Optimal-by-order (with constant not exceeding 2) cubature formulae for computing the integral $I_2^2(f)$ in the class $C_{2,L,L,N}^2$ have the form (6.14) with $T_p^\pm (p = 1, \dots, m^2)$ are computed by formulae (6.7)–(6.11) for $L_1 = L_2 = L$ and $h_1 = h_2 = h$. The error estimate of cubature formulae (6.14) is defined by the relationship (6.15).*

Explicit forms of optimal-by-order cubature formulae for computing the integral $I_3^2(f)$ in the classes $C_{2,L_1,L_2,N}$ and $C_{2,L,L,N}^2$ can be derived analogously.

7 Acknowledgements

The authors are grateful to Prof. V. Zadiraka and Dr T. Sag, for fruitful discussions and the Australian Research Council for partial support (Grant 179406). We also thank Dr D. Smith for his helpful assistance at the final stage of preparation of this paper.

References

- [1] Alaylioglu, A., Numerical Evaluation of Finite Fourier Integrals, *J. Comput. Appl. Math.*, **9**, 1983, 305–313.
- [2] Berezovskii, A.I., Ivanov, V.V., On Optimal-By-Accuracy Uniform Spline Approximation, *S. Mathematics (Iz. VUZ)*, No. **10**, 1977, 14–24.
- [3] Berezovskii, A.I., Nechiporenko, N.E., Optimal Accuracy Approximation of Functions and Their Derivatives, *Journal of S. Mathematics*, **54**, 1991, 799–812.
- [4] Blahut, R.E., *Fast Algorithms for Digital Signal Processing*, Addison-Wesley, 1987.
- [5] Chan, S.C. and Ho, K.L., A New Two-Dimensional Fast Cosine Transform, *IEEE Trans. Signal Process.*, **39**, No. 2, 1991, 481–485.
- [6] Cools, R., Constructing Cubature Formulae: The Science Behind the Art, *Acta Numerica*, Cambridge University Press, 1997, 1–54.
- [7] Davis, P. and Rabinowitz, P., *Methods of Numerical Integration*, Academic Press, 1984.

- [8] Drachman, B. and Ross, J., Approximation of Certain Functions Given by Integrals with Highly Oscillatory Integrands, *IEEE Transactions on Antennas and Propagation*, **42**, No. 9, 1994, 1355–1356.
- [9] Einarson, B., Numerical Calculation of Fourier Integrals with Cubic Splines, *BIT*, **8**, No. 3, 1968, 279–286.
- [10] Ersoy, O.K., *Fourier-Related Transforms, Fast Algorithms, and Applications*, Prentice Hall, 1997.
- [11] Filon, L.N., On a Quadrature Formula for Trigonometric Integrals, *Proc. Roy. Soc. Edinburg*, **49**, 1928, 38–47.
- [12] Haider, Q. and Liu, L.C., Fourier and Bessel Transformations of Highly Oscillatory Functions, *J. Phys. A: Math. Gen.*, **25**, 1992, 6755–6760.
- [13] Hopkins, H.H., Numerical Evaluation of a Class of Double Integrals of Oscillatory Functions, *IMA J. of Numerical Analysis*, **9**, 1989, 61–80.
- [14] Levin, D., Fast Integration of Rapidly Oscillatory Functions, *J. Comput. Appl. Math.*, **67**, 1996, 95–101.
- [15] Melnik, K.N. and Melnik, R.V.N., On Computational Aspects of Certain Optimal Digital Signal Processing Algorithms, *Proc. of Computational Technique and Applications Conference: CTAC97*, Eds. J. Noye, M. Teubner and A. Gill, World Scientific, 1998, 433–440.
- [16] Melnik, K.N. and Melnik, R.V.N., Optimal-By-Order Quadrature Formulae for Fast Oscillatory Functions with Inaccurately Given A Priori Information, *Technical Report SC-MC-9810, Department of Mathematics and Computing, University of Southern Queensland*, 1998, submitted.
- [17] Melnik, K.N. and Melnik, R.V.N., A Note on Optimal-By-Order Cubature Formulae for Fast Oscillatory Functions in Lipschitz Classes, *Technical Report SC-MC-9811, Department of Mathematics and Computing, University of Southern Queensland*, 1998, submitted.
- [18] Mysovskih, I.P., *Interpolatorische Kubaturformulen*, Institut fur Geometrie und Praktische Matematik der RWTH Aachen, Bericht 74, 1992.
- [19] Sucharev, A., *Minimax Models in the Theory of Numerical Methods*, Kluwer Academic Publishers, 1992.
- [20] Traub, J.F. and Wozniakowski, H., *A General Theory of Optimal Algorithms*, New York: Academic Press, 1980.
- [21] Zadiraka, V.K. and Kasenov, S. Z., Optimal-By-Accuracy Quadrature Formulae for Computing Fourier Transform of Finite Functions from $C_{L,N}$, *Ukr. Mathematical Journal*, **38**, No. 2, 1986, 233–237.
- [22] Zadiraka, V.K., Abatov, N.T., Optimally Exact Algorithms for Solutions of a Certain Numerical Integration Problem, *Ukr. Mathematical Journal*, **43**, 1991, 43–54.
- [23] Zheludev, V.A., Periodic Splines and the Fast Fourier Transform, *Computational Mathematics and Mathematical Physics*, **32**, No. 2, 1992, 149.

- [24] Zheludev, V.A., Processing of Periodic Signals Using Spline-Wavelets, *Radioelectronics and Communications Systems*, **38**, No. 3, 1995, 1.
- [25] Zhileikin, Ya.M. and Kukarkin, A.B., A Fast Fourier-Bessel Transformation Algorithm, *Computational Mathematics and Mathematical Physics*, **35**, No. 7, 1995, 901.

USQ



TOOWOOMBA

**A NOTE ON OPTIMAL-BY-ORDER
CUBATURE FORMULAE FOR FAST
OSCILLATORY FUNCTIONS IN
LIPSCHITZ CLASSES**

K N Melnik

Department of Computer Science, Flinders University

R V N Melnik

Department of Mathematics & Computing, USQ

Faculty of Sciences Working Paper Series

THE UNIVERSITY OF
SOUTHERN
QUEENSLAND

Working Papers

FACULTY
OF
SCIENCES

**A NOTE ON OPTIMAL-BY-ORDER
CUBATURE FORMULAE FOR FAST
OSCILLATORY FUNCTIONS IN
LIPSCHITZ CLASSES**

K N Melnik

Department of Computer Science, Flinders University

R V N Melnik

Department of Mathematics & Computing, USQ

Faculty of Sciences Working Paper Series

SC-MC-9811

23 April 1998

A NOTE ON OPTIMAL-BY-ORDER CUBATURE FORMULAE FOR FAST OSCILLATORY FUNCTIONS IN LIPSCHITZ CLASSES

K. N. Melnik *

Department of Computer Science,
Flinders University, Adelaide, SA 5001, Australia

R. V. N. Melnik †

Department of Mathematics and Computing,
University of Southern Queensland, QLD 4350, Australia

Abstract

In this article, we consider problems of numerical integration of fast oscillatory functions of two variables when an accurate value of the Lipschitz constant is not available. Using spline approximations, we propose optimal-by-order (with a constant not exceeding two) cubature formulae that are applicable for a wide range of oscillatory patterns.

Key words: optimal-by-order cubature formulae, fast oscillatory integrands, spline approximations, interpolational classes.

AMS Subject Classification: 65D32, 65D30, 65D07.

1 Introduction

In this paper we construct optimal-by-order cubature formulae for numerical integration of fast oscillatory functions. Such formulae are important for many applied problems in physics and chemistry where information about the integrand are taken from measurements. In order to explain the difficulties arising we start from the one dimensional case. Consider the product $f(x) \exp(-i\omega x)$ on an interval (a, b) , where $\omega(b - a) \gg 1$ and $f(x)$ is a smooth function. Both functions, $\Re(f(x) \exp(-i\omega x))$ and $\Im(f(x) \exp(-i\omega x))$, have on the interval (a, b) approximately $\omega(b - a)/\pi$ zeros. Therefore, if we would like to approximate such functions by

*Currently with the Electronic Data Systems, 60 Waymouth Street, Adelaide, 5000, Australia

†Corresponding author, E-mail: melnik@usq.edu.au

polynomials we need polynomials of degree $n \gg \omega(b-a)/\pi$ (indeed, a polynomial of degree n has no more than n zeros). This is not only impractical in many applications but also may lead to instability of computation [12]. In many practical situations it is more reasonable to treat the oscillating factor $\exp(-\omega ix)$ (the same can be said about $\sin(\omega x)$ or $\cos(\omega x)$) as a weight function. The idea of taking an oscillating factor into account in quadrature formula coefficients has been extensively developed since Filon's work (see, for example, [11, 2, 15] and references therein). Problems of the construction of optimal-by-order formulae for numerical integration of fast oscillatory functions in one dimension in the case when *a priori* information about the integrand is given inaccurately were recently discussed in [16, 17] (see also references therein).

In the two dimensional case the difficulties in numerical integration of fast oscillatory functions essentially increase. Different ways to overcome such difficulties were discussed by many authors (see, for example, [13, 30, 19]). A recent survey on the construction of cubature formulae to approximate multivariate integrals can be found in [7]. However, results on optimal-by-order cubature formulae when *a priori* information is given inaccurately are still lacking in the literature. Such results are important for many problems in signal processing, image recognition and processing, diffraction problems, scattering theory, crystallography and many other applications [5, 10, 8, 6, 13], where we have to deal with the computation of integrals

$$I^n(f) = \int_{a_1}^{b_1} \dots \int_{a_n}^{b_n} f(x_1, \dots, x_n) \varphi_1(x_1) \dots \varphi_n(x_n) dx_1 \dots dx_n. \quad (1.1)$$

In (1.1) it is assumed that $\varphi_k(x_k)$, $k = 1, \dots, n$ are known integrable functions and $f(x_1, x_2, \dots, x_n)$ belongs to some predefined functional class F . In this paper we propose optimal-by-order (with constant not exceeding 2) cubature formulae for computing integrals (1.1) when $n = 2$. We consider the integral

$$I^2(f) = \int_a^b \int_c^d f(x_1, x_2) \varphi_1(x_1) \varphi_2(x_2) dx_1 dx_2, \quad (1.2)$$

where $\varphi_1(x_1)$, $\varphi_2(x_2)$ are known integrable functions, and $f(x_1, x_2)$ belongs to a given functional class F in the case of approximately given *a priori* information. More precisely, we assume that function $f(x_1, x_2)$ is given by a table of its values at fixed grid points. Therefore, instead of a wider functional class F we have to consider its narrowing, an interpolational class F_N . This consideration approaches a situation typical in the majority of applications when information on integrands is taken from measurements. In this paper we consider the case when $F_N \equiv C_{1,L,N \times M}^2$. This interpolational class is defined as a class of two-variable functions defined on $\Omega = [a, b] \times [c, d]$, given by their fixed values f_{ij} at nodes $(x_{1,i}; x_{2,j})$ of an arbitrary grid on Ω with $i = 1, \dots, N$, $j = 1, \dots, M$ and such that

$$\sup_{x_1, x_2} \max(|f'_{x_1}|, |f'_{x_2}|) \leq L. \quad (1.3)$$

We assume only that a certain estimate of the Lipschitz constant L is available *a priori*.

The paper is organized as follows.

- In Section 2 we give a minimax definition of optimality for problems of numerical integration.
- In Section 3 we derive optimal-by-order cubature formulae using spline-approximations.

- In Section 4 we consider algorithmic aspects of computing estimates for Fourier transforms of two-variable functions. Numerical results are also presented in this section.
- Conclusions and future directions are discussed in Section 5.

2 Minimax Optimality in Numerical Integration

We aim at the construction of cubature formulae for numerical integration of (1.2) that give the best result for the worst function in the class. This minimax concept of optimality in the theory of numerical algorithms goes back to Chebyshev's work and is the most natural when all *a priori* information about the problem is contained in the fact that $f \in F_N$. If we denote by $r(F_N, A, f)$ the result of application of algorithm A to function f , then the error of integration of this function by A is

$$v(F_N, A, f) = |I^n(f) - r(F_N, A, f)|. \quad (2.1)$$

As the worst function in class F_N we take a function that provides $\sup_{f \in F_N} v(F_N, A, f)$ for the given algorithm A on the set of all cubature formulae which use the information from the definition of class F_N . We consider the following characteristic

$$\delta(F_N) = \inf_{A \in M} \sup_{f \in F_N} v(F_N, A, f). \quad (2.2)$$

Now we are in position to give the formal definition of optimal algorithms for numerical integration of (1.2) or, more generally, (1.1).

Definition 2.1 *A cubature formula on which $\delta(F_N)$ is achieved (provided it exists) is called optimal-by-accuracy for the given class F_N . If for a cubature formula A^0*

$$v(F_N, A^0, f) \leq \delta(F_N) + \eta, \quad \eta \geq 0, \quad (2.3)$$

then A^0 is called an optimal cubature formula on the class F_N with accuracy up to η . If $\eta = o(\delta(F_N))$ or $\eta = O(\delta(F_N))$, then A^0 is called an asymptotically optimal or optimal-by-order cubature formula respectively.

In a quite general setting, the construction of numerical integration algorithm for (1.2) can be thought as the solution of a problem $P(I, S)$ with input dataset $I \in \mathcal{I}$ and output dataset $S \in \mathcal{S}$, where \mathcal{I} and \mathcal{S} are certain metric spaces with metrics ρ_1 and ρ_2 respectively. Whenever instead of I an approximate input dataset \tilde{I} is given (which is typical in the majority of applications), we have to deal with an uncertainty domain \mathcal{D}_1 induced by the approximate nature of input data. It is important to underline that the uncertainty domain $\mathcal{D}_1 \subset \mathcal{I}$ always gives rise to an uncertainty domain of output dataset $\mathcal{D}_2 \subset \mathcal{S}$. This set \mathcal{D}_2 becomes the most complete characteristic of the problem solution. In principle any element of the uncertainty domain \mathcal{D}_2 can be considered as the solution of the problem (1.2). Under inaccurately given *a priori* information, as an optimal solution of the problem, it is reasonable to choose a point for which the maximum distance along \mathcal{D}_2 is the minimal among all points in \mathcal{S} . If \mathcal{D}_2 is a bounded set in the metric space (\mathcal{S}, ρ_2) , then an element $x_0 \in \mathcal{S}$ for which

$$\sup_{y \in \mathcal{D}_2} \rho_2(x_0, y) = \inf_{x \in \mathcal{S}} \sup_{y \in \mathcal{D}_2} \rho_2(x, y) \quad (2.4)$$

is known to be the Chebyshev center of \mathcal{D}_2 and the quantity (2.4) is the Chebyshev radius of this set. If \mathcal{D}_2 is unbounded further *a priori* information on the location of the solution set in \mathcal{S} has to be sought (see [17]).

For the functional class $C_{1,L,N \times M}^2$, considered in this paper, the quantity (see [16, 22] and references therein)

$$I^*(F_N) = \frac{1}{2} (I^+(F_N) + I^-(F_N)) \quad (2.5)$$

is taken as optimal-by-accuracy value of integral $I^2(f)$, where

$$I^+(F_N) = \sup_{f \in F_N} I^2(f), \quad I^-(F_N) = \inf_{f \in F_N} I^2(f) \quad (2.6)$$

are respectively upper and lower limits of the set of possible values of integrals (1.2) in the domain of integration on the functions of class F_N . Then

$$\delta(F_N) = \frac{1}{2} (I^+(F_N) - I^-(F_N)). \quad (2.7)$$

In this case $I^*(F_N)$ is the Chebyshev center of uncertainty domain of values $I^2(f)$ on class F_N . The Chybyshhev radius of this domain coincides with $\delta(F_N)$.

The given definition of optimality corresponds to what is known in the literature as the *pure strategy of decision making*. The *sequential strategy of decision making* can be introduced in a similar way [22]. Concepts of optimality based on probabilistic approaches (see, for example, [25, 22]) are not discussed in this paper.

3 Spline Approximations in Numerical Integration of Fast Oscillatory Functions.

Compared to the one-dimensional case, considered in detail in [17], the two-dimensional case is more difficult to investigate [13]. In this section we propose cubature formulae that do not require the explicit value of the Lipschitz constant L from (1.3). The construction of such formulae is based on the replacement of integrand $f(x_1, x_2) \in F_N$ in (1.2) by a linear spline of a special form. The close connection of spline approximations with the problems of numerical integration can be traced back to classical works of Kolmogorov, Nikolskii and others (see, for example, [20, 1, 14, 21]). In the context of numerical integration of fast oscillatory functions, one of the first works was due to Einarson [9]. Optimal-by-accuracy and optimal-by-order quadrature and cubature formulae for some special functional classes were developed in [3, 4, 27, 26]. In what follows we assume that the accuracy of the definition of function $f(x_1, x_2)$ in an $N \times M$ grid of nodes is known (for example, as a result of measurements). Our main results are presented for the case when $f(x_1, x_2) \in C_{1,L,N \times M}^2$ and only an estimate of the Lipschitz constant L is available *a priori*. In this case it is reasonable to apply integration formulae obtained by the residual minimization method [18, 23, 24] where the value of L is not involved. The results presented bellow generalise our results obtained for the one dimensional case in [16, 17].

Let us replace integrand $f(x_1, x_2) \in C_{1,L,N \times M}^2$ in (1.2) by the linear spline $S(x_1, x_2)$ of the following form

$$S(x_1, x_2) = (1-u)\left((1-t)f_{i,j} + tf_{i+1,j}\right) + u\left((1-t)f_{i,j+1} + tf_{i+1,j+1}\right), \quad (3.1)$$

where

$$(x_1, x_2) \in \Omega_{i,j}, \quad \Omega_{i,j} = [x_{1,i}, x_{1,i+1}] \times [x_{2,j}, x_{2,j+1}] \subset \Omega, \quad (3.2)$$

and

$$t = \frac{x_1 - x_{1,i}}{x_{1,i+1} - x_{1,i}}, \quad u = \frac{x_2 - x_{2,j}}{x_{2,j+1} - x_{2,j}}, \quad i = 1, \dots, N-1, \quad j = 1, \dots, M-1. \quad (3.3)$$

Since $S(x_1, x_2) \in C_{1,L,N \times M}^2$ (see, for example, [3, 4, 17] and references therein), the cubature formula

$$\tilde{R}(S) = \int_a^b \int_c^d S(x_1, x_2) \varphi_1(x_1) \varphi_2(x_2) dx_1 dx_2 \quad (3.4)$$

is optimal-by-order with a constant not exceeding two [25, 16, 17]. This result is independent of the mutual arrangement of nodes of an arbitrary grid on Ω and points where functions $\varphi_1(x_1)$, $\varphi_2(x_2)$ change their signs on Ω .

Important special cases of integrals (1.2) include those of the following form

$$I_2^2(f) = \int_a^b \int_c^d f(x_1, x_2) \sin(\omega x_1) \sin(\omega x_2) dx_1 dx_2, \quad (3.5)$$

and

$$I_3^2(f) = \int_a^b \int_c^d f(x_1, x_2) \cos(\omega x_1) \cos(\omega x_2) dx_1 dx_2. \quad (3.6)$$

For integral (3.5), the cubature formula (3.4), constructed using a linear spline (3.1)–(3.3), has the form

$$\begin{aligned} \tilde{R}_2(\omega, S) = & \sum_{i=1}^{N-1} \sum_{j=1}^{M-1} \frac{(f_{i,j} + f_{i+1,j+1} - f_{i,j+1} - f_{i+1,j})}{\omega^2(x_{1,i+1} - x_{1,i})(x_{2,j+1} - x_{2,j})} (\sin(\omega x_{1,i+1}) - \sin(\omega x_{1,i})) \times \\ & (\sin(\omega x_{2,j+1}) - \sin(\omega x_{2,j})) + \sum_{i=1}^{N-1} \frac{(\sin(\omega x_{1,i+1}) - \sin(\omega x_{1,i}))}{\omega^3(x_{1,i+1} - x_{1,i})} ((f_{i+1,1} - f_{i,1}) \cos(\omega c) - \\ & (f_{i+1,M} - f_{i,M}) \cos(\omega d)) + \sum_{j=1}^{M-1} \frac{(\sin(\omega x_{2,j+1}) - \sin(\omega x_{2,j}))}{\omega^2(x_{2,j+1} - x_{2,j})} ((f_{1,j+1} - f_{1,j}) \cos(\omega a) - \\ & (f_{N,j+1} - f_{N,j}) \cos(\omega b)) + \frac{1}{\omega^2} (f_{1,1} \cos(\omega a) \cos(\omega c) + f_{N,M} \cos(\omega b) \cos(\omega d) - \\ & f_{N,1} \cos(\omega b) \cos(\omega c) - f_{1,M} \cos(\omega a) \cos(\omega d)). \end{aligned} \quad (3.7)$$

Similarly, for integral (3.6), we have

$$\begin{aligned}
\tilde{R}_3(\omega, S) = & \sum_{i=1}^{N-1} \sum_{j=1}^{M-1} \frac{(f_{i,j} + f_{i+1,j+1} - f_{i,j+1} - f_{i+1,j})}{\omega^4(x_{1,i+1} - x_{1,i})(x_{2,j+1} - x_{2,j})} (\cos(\omega x_{1,i+1}) - \cos(\omega x_{1,i})) \times \\
& (\cos(\omega x_{2,j+1}) - \cos(\omega x_{2,j})) + \sum_{i=1}^{N-1} \frac{(\cos(\omega x_{1,i+1}) - \cos(\omega x_{1,i}))}{\omega^3(x_{1,i+1} - x_{1,i})} ((f_{i,1} - f_{i+1,1}) \sin(\omega c) - \\
& (f_{i,M} - f_{i+1,M}) \sin(\omega d)) + \sum_{j=1}^{M-1} \frac{(\cos(\omega x_{2,j+1}) - \cos(\omega x_{2,j}))}{\omega^2(x_{2,j+1} - x_{2,j})} ((f_{1,j} - f_{1,j+1}) \sin(\omega a) - \\
& (f_{N,j} - f_{N,j+1}) \sin(\omega b)) + \frac{1}{\omega^2} (f_{1,1} \sin(\omega a) \sin(\omega c) + f_{N,M} \sin(\omega b) \sin(\omega d) - \\
& f_{N,1} \sin(\omega b) \sin(\omega c) - f_{1,M} \sin(\omega a) \sin(\omega d)). \tag{3.8}
\end{aligned}$$

The formulae (3.7) and (3.8) hold for any arbitrary nonuniform grid constructed on Ω . In the special case of a uniform grid with steps h_i in the x_i -direction ($i = 1, 2$), cubature formulae (3.7) and (3.8) take the form

$$\begin{aligned}
\tilde{R}_2(\omega, S) = & \frac{4 \sin(\omega h_1/2) \sin(\omega h_2/2)}{\omega^4 h_1 h_2} \sum_{i=1}^{N-1} \sum_{j=1}^{M-1} (f_{i,j} + f_{i+1,j+1} - f_{i+1,j} - f_{i,j+1}) \times \\
& \cos\left(\omega(i + \frac{1}{2})h_1\right) \cos\left(\omega(j + \frac{1}{2})h_2\right) + \frac{2 \sin(\omega h_1/2)}{\omega^3 h_1} \sum_{i=1}^{N-1} \cos\left(\omega(i + \frac{1}{2})h_1\right) ((f_{i+1,1} - f_{i,1}) \times \\
& \cos(\omega c) - (f_{i+1,M} - f_{i,M}) \cos(\omega d)) + \frac{2 \sin(\omega h_2/2)}{\omega^3 h_2} \sum_{j=1}^{M-1} \cos\left(\omega(j + \frac{1}{2})h_2\right) ((f_{1,j+1} - f_{1,j}) \times \\
& \cos(\omega a) - (f_{N,j+1} - f_{N,j}) \cos(\omega b)) + \frac{1}{\omega^2} (f_{1,1} \cos(\omega a) \cos(\omega c) + f_{N,M} \cos(\omega b) \cos(\omega d) - \\
& f_{N,1} \cos(\omega b) \cos(\omega c) - f_{1,M} \cos(\omega a) \cos(\omega d)) \tag{3.9}
\end{aligned}$$

and

$$\begin{aligned}
\tilde{R}_3(\omega, S) = & \frac{4 \sin(\omega h_1/2) \sin(\omega h_2/2)}{\omega^4 h_1 h_2} \sum_{i=1}^{N-1} \sum_{j=1}^{M-1} (f_{i,j} + f_{i+1,j+1} - f_{i+1,j} - f_{i,j+1}) \times \\
& \sin\left(\omega(i + \frac{1}{2})h_1\right) \sin\left(\omega(j + \frac{1}{2})h_2\right) + \frac{2 \sin(\omega h_1/2)}{\omega^3 h_1} \sum_{i=1}^{N-1} \sin\left(\omega(i + \frac{1}{2})h_1\right) ((f_{i+1,1} - f_{i,1}) \times \\
& \sin(\omega c) - (f_{i+1,M} - f_{i,M}) \sin(\omega d)) + \frac{2 \sin(\omega h_2/2)}{\omega^3 h_2} \sum_{j=1}^{M-1} \sin\left(\omega(j + \frac{1}{2})h_2\right) ((f_{1,j+1} - f_{1,j}) \times \\
& \sin(\omega a) - (f_{N,j+1} - f_{N,j}) \sin(\omega b)) + \frac{1}{\omega^2} (f_{1,1} \sin(\omega a) \sin(\omega c) + f_{N,M} \sin(\omega b) \sin(\omega d) - \\
& f_{N,1} \sin(\omega b) \sin(\omega c) - f_{1,M} \sin(\omega a) \sin(\omega d))
\end{aligned}$$

$$f_{N,1} \sin(\omega b) \sin(\omega c) - f_{1,M} \sin(\omega a) \sin(\omega d)) \quad (3.10)$$

respectively.

We note that formulae (3.7), (3.10) contain all available *a priori* information about the problem. In the next section we use these formulae for computing Fourier transform estimates.

4 Computing Estimates of Fourier Transforms in the Two-Dimensional Case

In this section we consider functions that have bounded first derivatives and that are given by their exact values in $K = M \times N$ nodes of a uniform grid. The need in such a consideration arises by a number of applications in signal processing and image recognition, where it is important to compute different values, $\tilde{I}_2(\omega_{k_1}, \omega_{k_2})$, $\tilde{I}_3(\omega_{k_1}, \omega_{k_2})$, of estimates for sin- and cos-Fourier transformations:

$$\tilde{I}_2(\omega_{k_1}, \omega_{k_2}) = \int_a^b \int_c^d f(x, y) \sin(\omega_{k_1} x) \sin(\omega_{k_2} y) dx dy, \quad (4.1)$$

$$\tilde{I}_3(\omega_{k_1}, \omega_{k_2}) = \int_a^b \int_c^d f(x, y) \cos(\omega_{k_1} x) \cos(\omega_{k_2} y) dx dy, \quad (4.2)$$

where $\omega_{k_1} = 2\pi k_1/(b-a)$, $\omega_{k_2} = 2\pi k_2/(d-c)$, $k_i = 1, \dots, M_i - 1$, $M_i = 2^{m_i} + 1$ ($M_1 \equiv N$, $M_2 \equiv M$), and $m_i \geq 3$ ($i = 1, 2$) are integer numbers. As in Section 3, we assume that function $f(x, y)$ is finite in the domain $\Omega = [a, b] \times [c, d]$ and that $f(x_i, y_j) = f_{ij}$ where f_{ij} are given real numbers, (x_i, y_j) are nodes of the uniform grid $\Delta = \Delta_x \times \Delta_y$, $i = 1, \dots, M_1$, $j = 1, \dots, M_2$ with

$$\Delta_x = \{a = x_1 < x_2 < \dots < x_{M_1} = b\}, \quad \Delta_y = \{c = y_1 < y_2 < \dots < y_{M_2} = d\}, \quad (4.3)$$

The following algorithm for computing K values of estimates (4.1) and (4.2) is based on the results of section 3.

Algorithm 4.1

- Input $M_1, M_2, a, b, c, d, \{x_i\}, \{y_i\}, \{f_{ij}\}$, $i = 1, \dots, M_1$, $j = 1, \dots, M_2$;
- Compute values of frequencies $\{\omega_{k_1}\}$, $\{\omega_{k_2}\}$ by formulae:

$$\omega_{k_1} = 2\pi k_1/(b-a), \quad \omega_{k_2} = 2\pi k_2/(d-c), \quad k_i = 1, \dots, M_i - 1 (i = 1, 2); \quad (4.4)$$

- Compute steps $h_1 = (b-a)/(M_1 - 1)$, $h_2 = (d-c)/(M_2 - 1)$;
- Compute values $\{\sin(\omega_{k_1} x_i)\}$, $\{\cos(\omega_{k_1} x_i)\}$ and $\{\sin(\omega_{k_2} y_j)\}$, $\{\cos(\omega_{k_2} y_j)\}$ for $i = 1, \dots, M_1$, $j = 1, \dots, M_2$, $k_i = 1, \dots, M_i - 1 (i = 1, 2)$;
- Compute values $\tilde{S}_1(\omega_{k_1}, \omega_{k_2})$ by the formula

$$\begin{aligned} \tilde{S}_1(\omega_{k_1}, \omega_{k_2}) = & \frac{1}{\omega_{k_1}^2 \omega_{k_2} h_1} \sum_{i=1}^{M_1-1} (\sin(\omega_{k_1} x_{i+1}) - \sin(\omega_{k_1} x_i)) \times \\ & ((f_{i+1,1} - f_{i,1}) \cos(\omega_{k_2} c) - (f_{i+1,M_2} - f_{i,M_2}) \cos(\omega_{k_2} d)), \end{aligned} \quad (4.5)$$

for $k_i = 1, \dots, M_i - 1$, $i = 1, 2$;

- Compute values $\bar{S}_2(\omega_{k_1}, \omega_{k_2})$ by the formula

$$\begin{aligned} \bar{S}_2(\omega_{k_1}, \omega_{k_2}) = & \frac{1}{\omega_{k_1} \omega_{k_2}^2 h_2} \sum_{j=1}^{M_2-1} (\sin(\omega_{k_2} y_{j+1}) - \sin(\omega_{k_2} y_j)) \times \\ & ((f_{1,j+1} - f_{1,j}) \cos(\omega_{k_1} a) - (f_{M_1,j+1} - f_{M_1,j}) \cos(\omega_{k_1} b)), \end{aligned} \quad (4.6)$$

for $k_i = 1, \dots, M_i - 1$, $i = 1, 2$;

- Compute values $\bar{S}_3(\omega_{k_1}, \omega_{k_2})$ by the formula

$$\begin{aligned} \bar{S}_3(\omega_{k_1}, \omega_{k_2}) = & \frac{1}{\omega_{k_1} \omega_{k_2}^2 h_1 h_2} \sum_{i=1}^{M_1-1} \sum_{j=1}^{M_2-1} (f_{ij} + f_{i+1,j+1} - f_{i,j+1} - f_{i+1,j}) \times \\ & (\sin(\omega_{k_1} x_{i+1}) - \sin(\omega_{k_1} x_i)) (\sin(\omega_{k_2} y_{j+1}) - \sin(\omega_{k_2} y_j)); \end{aligned} \quad (4.7)$$

for $k_i = 1, \dots, M_i - 1$, $i = 1, 2$;

- Compute estimates $\bar{R}(\omega_{k_1}, \omega_{k_2})$ for the sin Fourier transform by the formula

$$\begin{aligned} \bar{R}(\omega_{k_1}, \omega_{k_2}) = & \bar{S}_3(\omega_{k_1}, \omega_{k_2}) + \bar{S}_1(\omega_{k_1}, \omega_{k_2}) + \bar{S}_2(\omega_{k_1}, \omega_{k_2}) + \\ & \frac{1}{\omega_{k_1} \omega_{k_2}} (f_{1,1} \cos(\omega_{k_1} a) \cos(\omega_{k_2} c) + f_{M_1,M_2} \cos(\omega_{k_1} b) \cos(\omega_{k_2} d) - \\ & f_{M_1,1} \cos(\omega_{k_1} b) \cos(\omega_{k_2} c) - f_{1,M_2} \cos(\omega_{k_1} a) \cos(\omega_{k_2} d)); \end{aligned} \quad (4.8)$$

for $k_i = 1, \dots, M_i - 1$, $i = 1, 2$;

- Compute values $\bar{\bar{S}}_1(\omega_{k_1}, \omega_{k_2})$ by the formula

$$\begin{aligned} \bar{\bar{S}}_1(\omega_{k_1}, \omega_{k_2}) = & \frac{1}{\omega_{k_1}^2 \omega_{k_2} h_1} \sum_{i=1}^{M_1-1} (\cos(\omega_{k_1} x_{i+1}) - \cos(\omega_{k_1} x_i)) \times \\ & ((f_{i,1} - f_{i+1,1}) \sin(\omega_{k_2} c) - (f_{i,M_2} - f_{i+1,M_2}) \sin(\omega_{k_2} d)) \end{aligned} \quad (4.9)$$

for $k_i = 1, \dots, M_1 - 1$, $i = 1, 2$;

- Compute values $\bar{\bar{S}}_2(\omega_{k_1}, \omega_{k_2})$ by the formula

$$\begin{aligned} \bar{\bar{S}}_2(\omega_{k_1}, \omega_{k_2}) = & \frac{1}{\omega_{k_1} \omega_{k_2}^2 h_2} \sum_{j=1}^{M_2-1} (\cos(\omega_{k_2} y_{j+1}) - \cos(\omega_{k_2} y_j)) \times \\ & ((f_{1,j} - f_{1,j+1}) \sin(\omega_{k_1} a) - (f_{M_1,j} - f_{M_1,j+1}) \sin(\omega_{k_1} b)) \end{aligned} \quad (4.10)$$

for $k_i = 1, \dots, M_i - 1$, $i = 1, 2$;

- Compute values $\bar{\bar{S}}_3(\omega_{k_1}, \omega_{k_2})$ by the formula

$$\begin{aligned} \bar{\bar{S}}_3(\omega_{k_1}, \omega_{k_2}) = & \frac{1}{\omega_{k_1} \omega_{k_2}^2 h_1 h_2} \sum_{i=1}^{M_1-1} \sum_{j=1}^{M_2-1} (f_{ij} + f_{i+1,j+1} - f_{i,j+1} - f_{i+1,j}) \times \\ & (\cos(\omega_{k_1} x_{i+1}) - \cos(\omega_{k_1} x_i)) (\cos(\omega_{k_2} y_{j+1}) - \cos(\omega_{k_2} y_j)) \end{aligned} \quad (4.11)$$

for $k_i = 1, \dots, M_i - 1$, $i = 1, 2$;

- Compute estimates $\bar{\bar{R}}(\omega_{k_1}, \omega_{k_2})$ for the sin Fourier transform by the formula

$$\begin{aligned} \bar{\bar{R}}(\omega_{k_1}, \omega_{k_2}) = & \bar{\bar{S}}_3(\omega_{k_1}, \omega_{k_2}) + \bar{\bar{S}}_1(\omega_{k_1}, \omega_{k_2}) + \bar{\bar{S}}_2(\omega_{k_1}, \omega_{k_2}) + \\ & \frac{1}{\omega_{k_1} \omega_{k_2}} (f_{1,1} \sin(\omega_{k_1} a) \sin(\omega_{k_2} c) + f_{M_1, M_2} \sin(\omega_{k_1} b) \sin(\omega_{k_2} d) - \\ & f_{M_1, 1} \sin(\omega_{k_1} b) \sin(\omega_{k_2} c) - f_{1, M_2} \sin(\omega_{k_1} a) \sin(\omega_{k_2} d)); \end{aligned} \quad (4.12)$$

for $k_i = 1, \dots, M_i - 1$, $i = 1, 2$;

- Output data $\{\omega_{k_1}\}$, $\{\omega_{k_2}\}$, $\bar{R}(\omega_{k_1}, \omega_{k_2})$, $\bar{\bar{R}}(\omega_{k_1}, \omega_{k_2})$, $k_1 = 1, \dots, M_1 - 1$, $k_2 = 1, \dots, M_2 - 1$.

For the given problem the hereditary error is zero. We assume that calculations are perform in a floating-point regime with round-off of the results of arithmetic operations using the standard rule up to τ binary digits in normalised mantissae of numbers. Then, for computing $\bar{R}(\omega_{k_1}, \omega_{k_2})$ the estimate of round-off error has the form (see details in [26, 27] and references therein)

$$\begin{aligned} \epsilon \leq & \frac{2^\tau}{\omega_{k_1}^2 \omega_{k_2}^2 h_1 h_2} \left((14 + 1.06(M_1 - 1)) \omega_{k_2} h_2 \max_{1 \leq i \leq M_1 - 1} |(\sin(\omega_{k_1} x_{i+1}) - \sin(\omega_{k_1} x_i)) \times \right. \\ & ((f_{i+1,1} - f_{i,1}) \cos(\omega_{k_2} c) - (f_{i+1,M_2} - f_{i,M_2}) \cos(\omega_{k_2} d)) \Big| + (14 + 1.06(M_2 - 1)) \omega_{k_1} h_1 \times \\ & \left. \max_{1 \leq i \leq M_2 - 1} |(\sin(\omega_{k_2} y_{j+1}) - \sin(\omega_{k_2} y_j)) ((f_{1,j+1} - f_{1,j}) \cos(\omega_{k_1} a) - (f_{M_1,j+1} - f_{M_1,j}) \right. \\ & \left. \cos(\omega_{k_1} b))| (17 + 1.06(M_2 - 1) + 1.06(M_1 - 1)) \max_{1 \leq i \leq M_1 - 1, 1 \leq j \leq M_2 - 1} |(f_{i,j} + f_{i+1,j+1} - \right. \\ & f_{i,j+1} - f_{i+1,j}) (\sin(\omega_{k_1} x_{i+1}) - \sin(\omega_{k_1} x_i)) (\sin(\omega_{k_2} y_{j+1}) - \sin(\omega_{k_2} y_j))| + 19 \omega_{k_1} \omega_{k_2} h_1 h_2 \times \\ & |f_{1,1} \cos(\omega_{k_1} a) \cos(\omega_{k_2} c) + f_{M_1,M_2} \cos(\omega_{k_1} b) \cos(\omega_{k_2} d) - f_{M_1,1} \cos(\omega_{k_1} b) \cos(\omega_{k_2} c) - \\ & \left. f_{1,M_2} \cos(\omega_{k_1} a) \cos(\omega_{k_2} d)| \right). \end{aligned} \quad (4.13)$$

The estimate of the round-off error in computing $\bar{\bar{R}}(\omega_{k_1}, \omega_{k_2})$ can be obtained analogously.

Algorithm 4.1 was applied to computing estimates for the sin- and cos- Fourier transforms on a unit square

$$\int_0^1 \int_0^1 f(x, y) \sin(\omega_k x) \sin(\omega_k y) dx dy, \text{ and } \int_0^1 \int_0^1 f(x, y) \cos(\omega_k x) \cos(\omega_k y) dx dy, \quad (4.14)$$

in a wide range of frequencies ω_k .

Example 1. Let $f(x, y) = 5x + 2y$, $M_1 = 2^5 + 1$, $M_2 = 2^5 + 1$. Let us denote by RS and RC the estimates of sin- and cos- Fourier transforms, and by ST and CT the values of $I_2(\omega_k)$ and $I_3(\omega_k)$ obtained analytically. The results of computations, given in Table 1, confirm

Frequency	RS	RC	ST	CT
7.0685830	-.024910500	.065944380	-.024910520	.065944400
159.1740000	.000209473	.000204954	.000209473	.000204954
516.0066000	-.000005434	.000013134	-.000005434	.000013134
864.9852000	.000018757	.000000000	.000018757	.000000000
4741.7110000	.000000233	.000000234	.000000233	.000000234

Table 1: $f(x, y) = 5x + 2y$.

that the cubature formulae (3.9), (3.10) are exact on linear functions, taking into account round-off error.

The next two examples demonstrate the efficiency of formulae (3.9) and (3.10) for non-linear functions.

Example 2. Let $f(x, y) = 7x^3 + 5y^2$, $M_1 = 2^5 + 1$, $M_2 = 2^5 + 1$. Computational results are

Frequency	RS	RC	ST	CT
7.0685830	-.028925090	.154089900	-.028914290	.154049500
159.1740000	.000365483	.000352090	.000365134	.000351826
516.0066000	-.000009287	.000022648	-.000009287	.000022647
864.9852000	.000032153	.000000000	.000032154	.000000000
4741.7110000	.000000400	.000000401	.000000400	.000000401

Table 2: $f(x, y) = 7x^3 + 5y^2$.

given in Table 2.

Example 3. Let $f(x, y) = \exp(x) + \exp(y)$, $M_1 = 2^7 + 1$, $M_2 = 2^7 + 1$. The results of

Frequency	RS	RC	ST	CT
7.0685830	-.007473294	.056956570	-.007473259	.056956130
159.1740000	.000281086	.000159910	.000281086	.000159909
516.0066000	-.000002021	.000010219	-.000002021	.000010218
864.9852000	.000019926	.000000000	.000019927	.000000000
4741.7110000	.000000315	.000000181	.000000315	.000000181

Table 3: $f(x, y) = \exp(x) + \exp(y)$.

computations for this example are presented in Table 3.

Example 4. Finally, we compute estimates of sin- and cos- Fourier transforms

$$I_2(\omega_{k_1}, \omega_{k_2}) = \int_0^1 \int_0^1 (\exp(x) + \exp(y)) \sin(\omega_{k_1}x) \sin(\omega_{k_2}y) dx dy, \quad (4.15)$$

$$I_3(\omega_{k_1}, \omega_{k_2}) = \int_0^1 \int_0^1 (\exp(x) + \exp(y)) \cos(\omega_{k_1}x) \cos(\omega_{k_2}y) dx dy \quad (4.16)$$

for frequencies $\omega_{k_1} \geq 2\pi$, $\omega_{k_2} \geq 2\pi$ and $M_1 = 2^5 + 1$, $M_2 = 2^4 + 1$. Computations for different

Frequency ω_{k_1}	Frequency ω_{k_2}	RS	RC	ST	CT
7.0685830	6.8067850	-.008656369	.043334930	-.008653957	.043325920
7.0685830	228.2891000	-.000162559	.002106953	-.000162465	.002106964
7.0685830	4092.7100000	-.000008027	.000096140	-.000008025	.000096136
159.1740000	6.8067850	-.001271415	.002299137	-.001270950	.002298555
159.1740000	228.2891000	.000195806	.000111610	.000195800	.000111606
159.1740000	4092.7100000	.000012952	.000005093	.000012951	.000005093
516.0066000	6.8067850	-.000129295	.000581024	-.000129263	.000580921
516.0066000	228.2891000	-.000005804	.000028211	-.000005803	.000028213
516.0066000	4092.7100000	-.000000337	.000001287	-.000000337	.000001287
4741.7110000	6.8067850	-.000042736	-.000077432	-.000042718	-.000077418
4741.7110000	228.2891000	.000006549	-.000003760	.000006549	-.000003760
4741.7110000	4092.7100000	.000000433	-.000000172	.000000433	-.000000172

Table 4: $f(x, y) = \exp(x) + \exp(y)$.

values of M_1 and M_2 in a wide range of frequencies $\omega_{k_1}, \omega_{k_2}$ showed that with increasing frequencies, the accuracy of computation increases for cubature formulae (3.9), (3.10). For the same set of frequencies $(\omega_{k_1}, \omega_{k_2})$, a variation of steps h_1 and h_2 does not substantially influence the accuracy of computations. Hence, the accuracy of cubature formulae (3.9), (3.10) is practically independent of the mutual arrangement of grid nodes and zeros of oscillating functions ($\sin(\omega_{k_1}x)\sin(\omega_{k_2}y)$ and $\cos(\omega_{k_1}x)\cos(\omega_{k_2}y)$ respectively). In conclusion to this section, we note that for computing $\bar{S}_1(\omega_{k_1}, \omega_{k_2})$, $\bar{S}_2(\omega_{k_1}, \omega_{k_2})$, $\bar{\bar{S}}_1(\omega_{k_1}, \omega_{k_2})$, $\bar{\bar{S}}_2(\omega_{k_1}, \omega_{k_2})$ we used the Fast Fourier Transform (FFT) algorithm in its discrete version (see, for example, [5, 10, 6]. The results obtained were used for computing values of $\bar{S}_3(\omega_{k_1}, \omega_{k_2})$, $\bar{\bar{S}}_3(\omega_{k_1}, \omega_{k_2})$, $k_i = 1, \dots, M_i - 1, i = 1, 2$. Asymptotically, for large K , the number of arithmetic operations for our algorithm using the FFT constitutes 1% of the number of arithmetical operations of the method that does not use FFT.

5 Conclusions and Future Directions

The design and implementation of optimal algorithms (and algorithms close to optimal with respect to their computational characteristics) is comparable in importance with that of new computer hardware. In this paper we constructed and tested cubature formulae for numerical integration of fast oscillatory functions from interpolational class $C_{1,L,N \times M}^2$ when only an estimate of the Lipschitz constant is known *a priori*. Our cubature formulae are optimal-by-order (with a constant not exceeding two) and can be applied for practically arbitrary oscillation patterns of integrand. We proposed an efficient algorithm for computing estimates of Fourier transformations. A simple computational implementation of this algorithm is an important feature in the design of efficient software application packages.

Issues in the construction of optimal-by-order cubature formulae for Lipschitz interpolational classes under assumption that L is *a priori* known will be the subject of a separate publication. Finally, we note that our results can be generalise to functional classes given by quasi-metrics.

6 Acknowledgements

The authors are grateful to Professors V. Zadiraka and T. Sag for fruitful discussions and the Australian Research Council for a partial support (Grant 179406). We thank Dr Christopher Vance for his kind attention to our work at the final stage of its preparation.

References

- [1] Ahlberg, J.H., Nilson, E.N. and Walsh, J.L., *The Theory of Splines and Their Applications*, NY, Academic Press, 1967.
- [2] Alaylioglu, A., Numerical Evaluation of Finite Fourier Integrals, *J. Comput. Appl. Math.*, **9**, 1983, 305–313.
- [3] Berezovskii, A.I., Ivanov, V.V., On Optimal-By-Accuracy Uniform Spline Approximation, *S. Mathematics (Iz. VUZ)*, No. 10, 1977, 14–24.
- [4] Berezovskii, A.I., Nechiporenko, N.E., Optimal Accuracy Approximation of Functions and Their Derivatives, *Journal of S. Mathematics*, **54**, 1991, 799–812.
- [5] Blahut, R.E., *Fast Algorithms for Digital Signal Processing*, Addison-Wesley, 1987.
- [6] Chan, S.C. and Ho, K.L., A New Two-Dimensional Fast Cosine Transform, *IEEE Trans. Signal Process.*, **39**, No. 2, 1991, 481–485.
- [7] Cools, R., Constructing Cubature Formulae: The Science Behind the Art, *Acta Numerica*, Cambridge University Press, 1997, 1–54.
- [8] Drachman, B. and Ross, J., Approximation of Certain Functions Given by Integrals with Highly Oscillatory Integrands, *IEEE Transactions on Antennas and Propagation*, **42**, No. 9, 1994, 1355–1356.
- [9] Einarson, B., Numerical Calculation of Fourier Integrals with Cubic Splines, *BIT*, **8**, No. 3, 1968, 279–286.
- [10] Ersoy, O.K., *Fourier-Related Transforms, Fast Algorithms, and Applications*, Prentice Hall, 1997.
- [11] Filon, L.N., On a Quadrature Formula for Trigonometric Integrals, *Proc. Roy. Soc. Edinburgh*, **49**, 1928, 38–47.
- [12] Haider, Q. and Liu, L.C., Fourier and Bessel Transformations of Highly Oscillatory Functions, *J. Phys. A: Math. Gen.*, **25**, 1992, 6755–6760.
- [13] Hopkins, H.H., Numerical Evaluation of a Class of Double Integrals of Oscillatory Functions, *IMA J. of Numerical Analysis*, **9**, 1989, 61–80.
- [14] Korneichuk, N.P., Ligun, A.A. and Babenko, V.F., *Extremal Properties of Polynomials and Splines*, NY, Nova Science Publishers, 1996.
- [15] Levin, D., Fast Integration of Rapidly Oscillatory Functions, *J. Comput. Appl. Math.*, **67**, 1996, 95–101.
- [16] Melnik, K.N. and Melnik, R.V.N., On Computational Aspects of Certain Optimal Digital Signal Processing Algorithms, *Proc. of Computational Technique and Applications Conference: CTAC97*, Eds. J. Noye, M. Teubner and A. Gill, World Scientific, 1998, 433–440.

- [17] Melnik, K.N. and Melnik, R.V.N., Optimal-By-Order Quadrature Formulae for Fast Oscillatory Functions with Inaccurately Given A Priori Information, *Technical Report SC-MC-9810, Department of Mathematics and Computing, University of Southern Queensland, 1998*, submitted.
- [18] Morozov, V.A., *Methods of Solving Incorrectly Posed Problems*, Springer-Verlag, 1984.
- [19] Mysovskih, I.P., *Interpolatorische Kubaturformulen*, Institut fur Geometrie und Praktische Matematik der RWTH Aachen, Bericht 74, 1992.
- [20] Nikolskii, S.M., *Quadrature Formulae*, Delhi, Hindustan Pub. Corp, 1964, International Monographs on Advanced Mathematics and Physics, 29.
- [21] Nurnberger, G., *Approximation by Spline Functions*, Springer-Verlag, 1989.
- [22] Sucharev, A., *Minimax Models in the Theory of Numerical Methods*, Kluwer Academic Publishers, 1992.
- [23] Tikhonov, A.N. et al. *Numerical Methods for the Solution of Ill-Posed Problems*, Dordrecht, Kluwer Academic, 1995.
- [24] Tikhonov, A.N., Leonov, A.S. and Yagola, A.G., *Nonlinear Ill-Posed Problems*, London, Chapman & Hall, 1996.
- [25] Traub, J.F. and Wozniakowski, H., *A General Theory of Optimal Algorithms*, New York: Academic Press, 1980.
- [26] Zadiraka, V.K. and Kasenov, S. Z., Optimal-By-Accuracy Quadrature Formulae for Computing Fourier Transform of Finite Functions from $C_{L,N}$, *Ukr. Mathematical Journal*, **38**, No. 2, 1986, 233–237.
- [27] Zadiraka, V.K., Abatov, N.T., Optimally Exact Algorithms for Solutions of a Certain Numerical Integration Problem, *Ukr. Mathematical Journal*, **43**, 1991, 43-54.
- [28] Zheludev, V.A., Periodic Splines and the Fast Fourier Transform, *Computational Mathematics and Mathematical Physics*, **32**, No. 2, 1992, 149.
- [29] Zheludev, V.A., Processing of Periodic Signals Using Spline-Wavelets, *Radioelectronics and Communications Systems*, **38**, No. 3, 1995, 1.
- [30] Zhileikin, Ya.M. and Kukarkin, A.B., A Fast Fourier-Bessel Transformation Algorithm, *Computational Mathematics and Mathematical Physics*, **35**, No. 7, 1995, 901.

USQ



TOOWOOMBA

**OPTIMAL-BY-ORDER QUADRATURE
FORMULAE FOR FAST OSCILLATORY
FUNCTIONS WITH INACCURATELY
GIVEN A PRIORI INFORMATION**

K N Melnik

Department of Computer Science, Flinders University

R V N Melnik

Department of Mathematics & Computing, USQ

Faculty of Sciences Working Paper Series

SC MC 9810

THE UNIVERSITY OF
S O U T H E R N
Q U E E N S L A N D

Working Papers

FACULTY
OF
SCIENCES

**OPTIMAL-BY-ORDER QUADRATURE
FORMULAE FOR FAST OSCILLATORY
FUNCTIONS WITH INACCURATELY
GIVEN A PRIORI INFORMATION**

K N Melnik

Department of Computer Science, Flinders University

R V N Melnik

Department of Mathematics & Computing, USQ

Faculty of Sciences Working Paper Series

SC-MC-9810

23 April 1998

OPTIMAL-BY-ORDER QUADRATURE FORMULAE FOR FAST OSCILLATORY FUNCTIONS WITH INACCURATELY GIVEN A PRIORI INFORMATION

K. N. Melnik *

Department of Computer Science,
Flinders University, Adelaide, SA 5001, Australia

R. V. N. Melnik †

Department of Mathematics and Computing,
University of Southern Queensland, QLD 4350, Australia

Abstract

In this article, the authors construct optimal-by-order quadrature formulae for integration of fast oscillatory functions in interpolational classes $C_{1,L,N}^1$ and $C_{1,L,N,\epsilon}^1$. The construction of efficient formulae for numerical integration of fast oscillatory functions is based on the application of the residual method and the method of quasi-solutions. Both cases, weak and strong oscillations, are considered. Results of numerical examples are presented.

Key words: Chebyshev center, interpolational classes, method of quasi-solutions, residual method, optimal-by-order quadrature formulae.

AMS Subject Classification: 65D32, 65D30, 65D07.

1 Introduction

The problem of computing finite integrals with oscillatory functions arises in many areas of mathematics. In mathematical literature some of the most frequently cited examples of this problem are connected with the computation of Fourier transformations and the solution of boundary value problems for partial differential equations. In applications we often come to the above problem when modelling optical and automated control systems, constructing

*Currently with the Electronic Data Systems, 60 Waymouth Street, Adelaide, 5000, Australia

†Corresponding author, E-mail: melnik@usq.edu.au

direction diagrammes of antennas, solving problems in radioastronomy, crystallography, signal processing and image recognition and when statistically processing experimental data [8, 12].

From the mathematical point of view, many of these applied problems can be reduced to the computation of integrals

$$I^n(f) = \int_{a_1}^{b_1} \dots \int_{a_n}^{b_n} f(x_1, \dots, x_n) \varphi_1(x_1) \dots \varphi_n(x_n) dx_1 \dots dx_n, \quad (1.1)$$

where $\varphi_k(x_k), k = 1, \dots, n$ are known integrable oscillatory functions, and $f(x_1, \dots, x_n)$ belongs to a predefined functional class F . In this paper we consider a special case of (1.1) ($n = 1$)

$$I^1(f) = \int_a^b f(x) \varphi(x) dx, \quad (1.2)$$

with a given oscillatory function $\varphi(x)$ such as $\sin(\omega x)$, $\cos(\omega x)$, $\exp(-\omega x)$ and with possibly very large values of ω . Large values of ω that are typical, for example, when processing high frequency signals, lead to major problems in computing (1.2). Indeed, using standard approaches such as the Gaussian method, even if $f(x)$ is a smooth function, one has to choose a very high degree polynomial approximating $f(x)$. The degree of such a polynomial has to substantially exceed $[\omega[b-a]/\pi]$. This may result in computational instability [13].

In order to overcome the above difficulties, coefficients of quadrature formulae have to include the dependence on ω . The classical formulae with such a dependence is the Filon formula and its modifications [10, 11, 19]. The further development of the Filon method has been connected with an approximation of $f(x)$ by an interpolating polynomial and the consequent integration, wherein $\varphi(x)$ has been treated as a weight function. Substantial contributions to the topics of numerical integration of fast oscillatory functions were made by Collatz, Erugin, Sobolev, Krylov, Nikolskii and many other outstanding mathematicians (some references can be found, for example, in [9, 2, 5, 34, 31]). Due to its practical importance, much efforts in this field have been concentrated on the development of algorithms for computing specific type integrals such as Fourier and Bessel integrals [24, 25, 26, 13, 33]. However, the constructive procedures for obtaining optimal-by-order, rather than optimal-by-accuracy, quadrature formulae (or formulae close to them in a certain sense) for the integration of fast oscillatory functions are still lacking in the literature. It is especially true in the case when the problem is considered from the point of view of the total error, taking into account inaccuracy of *a priori* available information [4, 32, 20, 30]. Such a consideration brings us closer to the real situation when information about integrands are taken from measurements. This case is in the main focus of our paper.

In the construction of methods for numerical integration, there is a natural contradiction between the desire to choose a wider functional class and the desire to better describe the problem and take into account many properties of the problem (that, in turn, leads to narrowing of the functional class). However, if the functional class is chosen, then all *a priori* information available for the construction of an algorithm for numerical integration is contained in the inclusion $f \in F$ [27]. In this paper we assume that function $f(x) \in F$ is given by a fixed table of its values $f_i, i = 1, \dots, N$ in N fixed points $\{x_i\}_{i=1}^N$ from its domain of definition. Although such a definition leads to a considerable narrowing of the corresponding class F to an interpolational class F_N , it allows us the maximal use of available information

about the function and, as a result, it allows us to improve the quality of quadrature formulae. Since in practice instead of exact input data $\{f_i\}_{i=1}^N$ we often know only approximate values $\{\tilde{f}_i\}_{i=1}^N$, we also consider the case of approximate definition of input data where values $\{f_i\}_{i=1}^N$ are taken from the domain defined by inequalities $|\tilde{f}_i - f_i| \leq \epsilon_i$, $i = 1, \dots, N$. In this case we shall say that $f(x) \in F_{N,\epsilon}$ with a fixed vector $\epsilon = (\epsilon_1, \dots, \epsilon_N)$. We specify two main interpolational classes, considered in this paper, as follows.

- $C_{1,L,N}^1$ is the class of continuous functions defined on the interval $[a, b]$, which have bounded (by constant L) first derivatives and take fixed values $f(x_1) = f_1, \dots, f(x_N) = f_N$ at fixed nodes of arbitrary grid, x_1, \dots, x_N ;
- $C_{1,L,N,\epsilon}^1$ is the class of continuous on $[a, b]$ functions which have bounded (by L) first derivatives and take values from the intervals $[f_i - \epsilon_i, f_i + \epsilon_i]$, $\epsilon_i \geq 0$, $i = 1, \dots, N$ at fixed nodes of arbitrary grid, x_1, \dots, x_N .

In what follows we assume that these functional classes are non-empty. The interest to such interpolational classes and other functional classes that satisfy different forms of the Lipschitz condition (or more generally defined by their quasi-metrics [27]) has recently dramatically increased in the context of optimisation problems.

Many problems of applied and computational mathematics, including those of numerical integration, can be described in the following generic form. We have to solve a problem $P(I, S)$ with a set of initial data $I \in (\mathcal{M}_1, \rho_1)$ and a solution (or a set of solutions) $S \in (\mathcal{M}_2, \rho_2)$ where \mathcal{M}_1 and \mathcal{M}_2 are certain metric spaces with metrics ρ_1 and ρ_2 respectively. Since exact initial data are typically unavailable, instead of I we are typically given another set \tilde{I} of approximate data. This set is the origin of an *uncertainty domain of input data* (see also [27]), denoted further by $U_1 \in \mathcal{M}_1$. In turn, the uncertainty domain U_1 gives rise to an *uncertainty domain of the solution* denoted further by $U_2 \in \mathcal{M}_2$. It is this set U_2 that determines properties of our solution in a sense that any element $R \in U_2$ can be formally considered as a solution of the problem.

If U_2 is a bounded set, then the Chebyshev center of U_2 is taken as the *optimal solution* of the problem. In other words we choose an element $x_0 \in \mathcal{M}_2$ such that

$$\sup_{y \in E} \rho_2(x_0, y) = \inf_{x \in \mathcal{M}_2} \sup_{y \in E} \rho_2(x, y). \quad (1.3)$$

The quantity defined by (1.3) gives the least possible error of the problem solution under given data, and is called the Chebyshev radius of the set U_2 . If \mathcal{M}_2 is a Banach space that is uniformly convex in every direction, then U_2 has at least one Chebyshev center.

If U_2 is unbounded (which is the case, for example, for many ill-posed problems), then more *a priori* information have to be used to locate the solution in S . Assume, for example, that the solution belongs to a subdomain G of S ($G \subset S$). Then, as the optimal solution of the problem we take the Chebyshev center of the set $F = G \cap U_2$. This case is much more difficult for investigation compared to the case of bounded U_2 . However, instead of finding the Chebyshev center of F , it is often more efficient to use other elements of F . Indeed, it is known that an arbitrary element of F represents this set with the accuracy not exceeding two times of the accuracy of representation of F by its Chebyshev radius [30, 31]. In this paper, we use two elements of F that have advantages over the Chebyshev center in the case when *a priori* information is given inaccurately. These elements can be defined as follows.

- We consider a point S_1 for which the distance from the given point of S (say zero point) is minimal compared to the distance from other points from F . The method of finding the point S_1 is known as *the residual method* (see [21, 28, 29] and references therein).
- The other our choice, which is especially efficient if F is compact, is the point S_2 , for which the corresponding point of the uncertainty domain U_1 is the least remote from a given point. The method of finding the point S_2 is known as *the method of quasi-solutions* (see [14, 15, 28, 29] and references therein).

The use of these two methods is the basis for our constructions of optimal (and close to them) quadrature formulae in interpolational classes $C_{1,L,N}^1$ and $C_{1,L,N,\epsilon}^1$.

The paper is organized as follows.

- In Section 2, using the residual method and the method of quasi-solutions, we obtain quadrature formulae for the numerical solution of problem (1.2).
- In Section 3 we derive error estimates for these formulae applied to computing (1.2) with $\varphi(x) = \sin(\omega x)$ and $\varphi(x) = \cos(\omega x)$ in interpolational class $C_{1,L,N}^1$. We consider two principally different cases, the case of weak oscillations and the case of strong oscillations of the integrand.
- In Section 4 we generalise the results obtained in Section 3 to class $C_{1,L,N,\epsilon}^1$.
- In Section 5 we consider the problem of computing estimates of the Fourier transforms when *a priori* information is given approximately. Algorithms and some numerical examples are also presented in this section.
- Conclusions and future directions are discussed in Section 6.

2 Construction of Optimal Quadrature Formulae in Interpolational Classes

Let $f \in F_N$ with F_N be an interpolational class and let \mathcal{M} be a set of integration algorithms. Then the accuracy of integration of a certain function f by algorithm A has to be chosen according to a certain criterion by which we can estimate the quality of the algorithm in terms of its error function $v(F_N, A, f)$. For the construction of quadrature formulae for (1.2), we use *the method of limit functions* which consists of the construction of the best algorithm for the worst function in the class [31, 27]. This minimax approach to the solution of problems in theory of numerical methods goes back to Chebyshev's works and with respect to the problems of optimal quadratures was first formulated by A.N. Kolmogorov (see [22] and references therein). Since that time, one may trace a close connection of efficient quadrature formulae with spline-approximations [1, 16, 23]. The idea of spline approximation is used also in this paper. As the worst function in class F_N we take a function that provides $\sup_{f \in F_N} v(F_N, A, f)$ for the given algorithm.

We introduce the following characteristic

$$\delta(F_N) = \inf_{A \in \mathcal{M}} \sup_{f \in F_N} v(F_N, A, f), \quad (2.1)$$

where

$$v(F_N, A, f) = |I^n(f) - r(F_N, A, f)|, \quad (2.2)$$

$r(F_N, A, f)$ is the result of application of algorithm A to function f , \mathcal{M} is the set of all quadrature formulae that use information consisting of the definition of class F_N . For the general integral (1.1) the above characteristic is introduced in the same way.

Definition 2.1 A quadrature formula on which $\delta(F_N)$ is achieved (assuming that such a limit exists) is called optimal-by-accuracy for the given class. If there exists a quadrature formulae A^0 such that

$$v(F_N, A^0, f) \leq \delta(F_N) + \eta, \quad \eta \geq 0, \quad (2.3)$$

then A^0 is called optimal quadrature formula on the class F_N with accuracy up to η . If $\eta = o(\delta(F_N))$ or $\eta = O(\delta(F_N))$, then A^0 is called asymptotically optimal or optimal-by-order quadrature formula.

For the construction of optimal-by-order and close to them quadrature formulae in interpolational classes F_N (say, $C_{1,L,N}^1$) and $F_{N,\epsilon}$ (say, $C_{1,L,N,\epsilon}^1$), we follow a general procedure, known in literature as the method of limit functions (see [31] and references therein).

Let consider a functional class F_N of functions defined in a given domain.

Definition 2.2 Function $f^\pm(x)$ is called majorant (minorant) of class F_N if both of the following two conditions

- $f^+(x) \geq f(x) (f^-(x) \leq f(x)) \forall f \in F_N$ and
- $f^+ \in F_N (f^- \in F_N)$.

are satisfied.

We construct upper and lower limits of the set of all possible values of integrals (1.2) in the domain of integration on functions of class F_N as follows

$$I^+(F_N) = \sup_{f \in F_N} I^1(f), \quad I^-(F_N) = \inf_{f \in F_N} I^1(f). \quad (2.4)$$

Quantities $I^\pm(F_N)$ in (2.4) are achieved on $f^\pm(x) \in F_N$, which are referred to as majorant and minorant of class F_N respectively. Then, the Chebyshev center $I^*(F_N)$ and the Chebyshev radius $\delta^*(F_N)$ of the uncertainty domain of values $I^1(f)$ on class F_N are determined as follows (see also [20, 27] and references therein)

$$I^*(F_N) = \frac{1}{2} (I^+(F_N) + I^-(F_N)), \quad \delta^*(F_N) = \frac{1}{2} (I^+(F_N) - I^-(F_N)). \quad (2.5)$$

As follows from the above discussion and Definition 2.1, a quadrature formula that computes $I^*(F_N)$ is called optimal-by-accuracy. If the domain of values of integral $I^1(f)$ is D , then the quantity $\delta^*(F_N)$ gives the error of representation of D by $I^*(F_N)$. Furthermore, a quadrature formula for computing $\bar{I}^1(f)$ such that

$$\sup_{f \in F_N} |\bar{I}^1(f) - I^1(f)| \leq \delta^* + \eta, \quad \eta > 0 \quad (2.6)$$

with $\eta = o(\delta^*), O(\delta^*), \delta^* \rightarrow 0$ is called respectively asymptotically optimal and optimal-by-order.

We note that under the given information about the problem any other quadrature formula do not give accuracy less than δ^* . We also note that for interpolational class F_N Chebyshev radius δ^* coincides with the optimal estimate. When $F_N = C_{1,L,N}^1$ or $F_{N,\epsilon} = C_{1,L,N,\epsilon}^1$ optimal-by-accuracy algorithms for the numerical evaluation of integral (1.2) with $\varphi(x) = \sin(\omega x)$ and $\varphi(x) = \cos(\omega x)$ (where ω is a real number that determines an oscillatory factor of the integrand such that $|\omega| \geq 2\pi/(b-a)$ and $f(x)$) were investigated in [3, 31] (see also references therein). Although such algorithms are suitable for a wide range of oscillatory patterns (with the assumption that the values of L and ϵ used in these algorithms are given accurately), in many practical cases it is not possible to apply these algorithms. Indeed, in practice it is typical that numerical *a priori* information (used for the definition of functional classes) is inaccurate. Hence, instead of exact values of L and ϵ , we rather have some estimations of these values. Below we propose efficient algorithms for such situations. We construct optimal-by-order (with a constant not exceeding 2), rather than optimal-by-accuracy, algorithms that are based on methods of quasi-solutions and the residual minimization.

In numerical integration algorithms, constructed on the basis of quasi-solutions, integrand $f(x)$ is approximated by a function which is the solution of the following problem [3, 14, 15, 28, 29]:

$$\min_{f \in F} \max_{i=1,\dots,N} \epsilon_i \text{ where } \epsilon_i = |f(x_i) - \tilde{f}_i|. \quad (2.7)$$

The method of quasi-solutions consists of the determination of such a function that least deviates from the given set of points (x_i, \tilde{f}_i) , $i = 1, \dots, N$. It is known [3], that the solution of (2.7) in such classes as $C_{1,L,N,\epsilon}^1$ (and as a special case in $C_{1,L,N}^1$) is a linear spline $S(x, L)$ for which maximal deviation from the given points (x_i, \tilde{f}_i) , $i = 1, \dots, N$ is minimal, i.e.

$$S(x, L) = \tilde{f}_i + \frac{x - x_i}{x_{i+1} - x_i} (\tilde{f}_{i+1} - \tilde{f}_i), \quad x \in [x_i, x_{i+1}], \quad i = 1, \dots, N-1, \quad (2.8)$$

where

$$\hat{f}_i = (\tilde{f}_i^+ + \tilde{f}_i^-)/2, \quad \tilde{f}_i^\pm = \pm \max_{1 \leq j \leq N} (\pm(\tilde{f}_j \mp L|x_j - x_i|)), \quad i = 1, \dots, N. \quad (2.9)$$

It is often the case that *a priori* information that defines the class F is given in the form of certain constraints on a functional $\Phi(f)$. For functional classes $C_{1,L,N}^1$ and $C_{1,L,N,\epsilon}^1$ this functional is the uniform norm of the derivative. In quadrature formulae constructed on the basis of the residual method [21, 28, 29], integrand $f(x)$ is approximated by a function which is the solution of the following problem

$$\min_{f \in F} \Phi(f). \quad (2.10)$$

The solution of (2.10) is a linear spline $S(x, M)$ which is defined by (2.8), (2.9) with constant L changed for constant M where

$$M = \max_{1 \leq i \leq N} \left(0, \max_{j > i} \frac{|\tilde{f}_j - \tilde{f}_i| - \epsilon_j - \epsilon_i}{x_j - x_i} \right). \quad (2.11)$$

The quadrature formulae, constructed on the basis of the method of quasi-solutions and the residual method, have the following forms respectively

$$\bar{R}(\varphi, S) = \int_a^b S(x, L)\varphi(x)dx, \quad (2.12)$$

$$\bar{\bar{R}}(\varphi, S) = \int_a^b S(x, M)\varphi(x)dx. \quad (2.13)$$

Formulae (2.12) and (2.13), used for computing (1.2), are optimal-by-order with a constant that does not exceed 2 (even compared with the case of exactly given L and ϵ) [30]. Typically, both the residual method and the method of quasi-solutions, are directed to a more precise recovery of available *a priori* information. Therefore, the application of formulae (2.12), (2.13) are the most appropriate for the case of inaccurately given *a priori* information.

3 Error Estimates for Optimal-By-Order Quadrature Formulae in Class $C_{1,L,N}^1$.

For the integrals

$$I_2^1(\omega, f) = \int_a^b f(x) \sin(\omega x)dx, \quad (3.1)$$

$$I_3^1(\omega, f) = \int_a^b f(x) \cos(\omega x)dx, \quad (3.2)$$

with a real number ω such that $|\omega| \geq 2\pi/(b-a)$, quadrature formulae (2.12) and (2.13) have the following forms

$$R_2(\omega, S) = \sum_{i=1}^{N-1} \left(\frac{\hat{f}'_i}{\omega^2} (\sin(\omega x_{i+1}) - \sin(\omega x_i)) \right) - \frac{1}{\omega} (\hat{f}_N \cos(\omega x_N) - \hat{f}_1 \cos(\omega x_1)), \quad (3.3)$$

$$R_3(\omega, S) = \sum_{i=1}^{N-1} \left(\frac{\hat{f}'_i}{\omega^2} (\cos(\omega x_{i+1}) - \cos(\omega x_i)) \right) + \frac{1}{\omega} (\hat{f}_N \sin(\omega x_N) - \hat{f}_1 \sin(\omega x_1)), \quad (3.4)$$

where $\hat{f}'_i = (\hat{f}_{i+1} - \hat{f}_i)/(x_{i+1} - x_i)$ (\hat{f}_i , $i = 1, \dots, N$ are defined by (2.9)).

In order to obtain error estimates for optimal quadrature formulae, constructed for computing (3.1), (3.2), we have to require a special mutual arrangement of points x_i , $i = 1, \dots, N$ and zeros of $\sin(\omega x)$ (or $\cos(\omega x)$) on $[a, b]$ [20]. We note that in class $C_{1,L,N}^1$ the solutions of problems (2.7) and (2.10) coincide and the linear spline $S(x, L)$ is a broken line that joins points (x_i, f_i) , $i = 1, \dots, N$

$$S(x, L) = f_i + \frac{x - x_i}{x_{i+1} - x_i} (f_{i+1} - f_i), \quad x \in [x_i, x_{i+1}], \quad i = 1, \dots, N-1. \quad (3.5)$$

Indeed, if $f(x) \in C_{1,L,N}^1$ then $\hat{f}_i = f_i$. Let us consider the second formula in (2.9). Since $\tilde{f}_i = f_i$, $i = 1, \dots, N$, it may be rewritten in the following form

$$f_i^\pm = \pm \max_{1 \leq j \leq N} (\pm(f_j \mp L|x_j - x_i|)), \quad i = 1, \dots, N. \quad (3.6)$$

It is obvious that for any function $f(x) \in C_{1,L,N}^1$

$$|f_j - f_i| \leq L|x_j - x_i|, \text{ i.e. } -L|x_j - x_i| \leq f_j - f_i \leq L|x_j - x_i| \quad (3.7)$$

and

$$-f_j - L|x_j - x_i| \leq -f_i \leq -f_j + L|x_j - x_i|, \quad i, j = 1, \dots, N. \quad (3.8)$$

Since

$$f_i \geq f_j - L|x_j - x_i| \quad \text{and} \quad -f_i \geq -f_j - L|x_j - x_i|, \quad (3.9)$$

using (3.6) and the chain of inequalities (3.7), (3.8) we have that

$$\max_{1 \leq j \leq N} (f_j - L|x_j - x_i|) = f_i \quad \text{and} \quad \max_{1 \leq j \leq N} (-f_j - L|x_j - x_i|) = -f_i \quad (3.10)$$

from which formula (3.5) immediately follows.

Let us first consider the case when $N \geq |\omega|$. Computing (3.1), we assume that $\left[\frac{|\omega|}{\pi}(b-a)\right] + 1$ zeros of function $\sin \omega x$ (or $\cos \omega x$ for computing (3.2)) are included in the number of nodes $x_i, i = 1, \dots, N$. We refer to this condition as *Condition C1*. Then the following result holds.

Theorem 3.1 *Let $f(x) \in C_{1,L,N}^1$, $N \geq |\omega|$, condition C1 satisfied and $\{f_i\}_{i=1,\dots,N}$ are given on $[a, b]$ in the nodes $x_i, i = 1, \dots, N$ of the grid with fixed ends $x_1 = a, x_{N+1} = b$. Then quadrature formula (3.3) for the computation of integral (3.1) is optimal-by-order with a constant that does not exceed 2. This result holds with the following error estimate*

$$\begin{aligned} v(C_{1,L,N}^1, R_2(\omega, S), f) \leq & \\ & \frac{L}{\omega} \left(\sum_{i=1}^{N-1} \left| \sin\left(\frac{\omega}{2}(x_{i+1} + x_i)\right) \right| \left(\frac{4}{\omega} \left(\sin^2 \frac{\omega \Delta x_i}{4} - \sin^2 \frac{\omega |\Delta f_i|}{4L} \right) \right) + \right. \\ & \frac{2}{\omega} \left| \sin \frac{\omega \Delta x_n}{2} \cos\left(\omega(b - \frac{\Delta x_N}{2})\right) - \Delta x_N \cos(\omega b) \right| + \\ & \left. \frac{2}{\omega} \left| \sum_{i=1}^{N-1} \operatorname{sign}(\Delta f_i) \cos\left(\frac{\omega}{2}(x_{i+1} + x_i)\right) \left(\sin \frac{\omega |\Delta f_i|}{2L} - \frac{|\Delta f_i|}{L \Delta x_i} \sin \frac{\omega \Delta x_i}{2} \right) \right| \right), \end{aligned} \quad (3.11)$$

where $\Delta f_i = f_{i+1} - f_i$, $\Delta x_i = x_{i+1} - x_i$, $i = 1, \dots, N$.

Proof. It is obvious that $S(x_i, L) = f_i$, $i = 1, \dots, N$, $|S'(x, L)| \leq L$. Hence, $S(x, L) \in C_{1,L,N}^1$. Since

$$\inf_{\psi \in C_{1,L,N}^1} I_2^1(\psi) \leq R_2(\omega, S) \leq \sup_{\psi \in C_{1,L,N}^1} I_2^1(\psi), \quad (3.12)$$

then the error estimate for (3.3) can be obtained on the basis of the following inequality

$$|R_2(\omega, S) - I_2^1(f)| \leq \max(I_2^+(f) - R_2(\omega, S), R_2(\omega, S) - I_2^-(f)), \quad (3.13)$$

where $I_2^-(f)$ is the lower and $I_2^+(f)$ is the upper limits of the set of all possible values of integral $I_2^1(f)$ on $C_{1,L,N}^1$.

Taking into account oscillations of $\sin(\omega x)$, the limit functions of the class $C_{1,L,N}^1$, $f^+(x)$ and $f^-(x)$, on each elementary segment $[x_i, x_{i+1}]$ can be written in the explicit form.

(a) For $x \in [x_i, \bar{x}_i]$, $\bar{x}_i = (x_i + x_{i+1})/2 - |\Delta f_i|/(2L)$, $\Delta f_i = f_{i+1} - f_i$ ($\Delta x_i = x_{i+1} - x_i$) we have

$$f^+(x) = f_i + L(x - x_i)\text{sign}(\sin(\omega x_i)), \quad f^-(x) = f_i - L(x - x_i)\text{sign}(\sin(\omega x_i)). \quad (3.14)$$

(b) For $x \in [\bar{x}_i, \bar{\bar{x}}_i]$, $\bar{\bar{x}}_i = (x_i + x_{i+1})/2 + |\Delta f_i|/(2L)$ we have

$$\begin{aligned} f^+(x) &= \frac{1}{2} (1 + \text{sign}(\Delta f_i)\text{sign}(\sin(\omega x_i))) (f_i + L(x - x_i) \text{sign}(\sin(\omega x_i))) \\ &\quad + \frac{1}{2} (1 - \text{sign}(\Delta f_i)\text{sign}(\sin(\omega x_i))) (f_{i+1} + L(x_{i+1} - x) \text{sign}(\sin(\omega x_i))), \\ f^-(x) &= \frac{1}{2} (1 - \text{sign}(\Delta f_i)\text{sign}(\sin(\omega x_i))) (f_i - L(x - x_i) \text{sign}(\sin(\omega x_i))) + \\ &\quad \frac{1}{2} (1 + \text{sign}(\Delta f_i)\text{sign}(\sin(\omega x_i))) (f_{i+1} - L(x_{i+1} - x) \text{sign}(\sin(\omega x_i))). \end{aligned} \quad (3.15)$$

(c) For $x \in [\bar{\bar{x}}_i, x_{i+1}]$ we have

$$f^+(x) = f_{i+1} + L(x_{i+1} - x)\text{sign}(\sin(\omega x_i)), \quad f^-(x) = f_{i+1} - L(x_{i+1} - x)\text{sign}(\sin(\omega x_i)) \quad (3.16)$$

As usual, in (3.14)–(3.16) the signum of function $\sin(\omega x)$ on $[x_i, x_{i+1}]$ ($i = 1, \dots, N - 1$) is denoted by $\text{sign}(\sin(\omega x_i))$. We choose an approximation $f^*(x)$ for integrand $f(x)$ in the following form

$$f^*(x) = \begin{cases} S(x, L), & x \in [x_i, x_{i+1}], i = 1, \dots, N - 1, \\ f_N, & x \in [x_N, x_{N+1}]. \end{cases} \quad (3.17)$$

Let us obtain an estimate from above for $v(C_{1,L,N}^1, R_2(\omega, S), f)$. We have

$$\begin{aligned} v(C_{1,L,N}^1, R_2(\omega, S), f) &\leq \max \left(\sum_{i=1}^N \int_{x_i}^{x_{i+1}} (f^+(x) - f^*(x)) \times \right. \\ &\quad \left. \sin(\omega x) dx, \sum_{i=1}^N \int_{x_i}^{x_{i+1}} (f^*(x) - f^-(x)) \sin(\omega x) dx \right). \end{aligned} \quad (3.18)$$

The next step is to calculate explicitly integrals in (3.18). Taking into account the explicit forms of the majorant of class $C_{1,L,N}^1$ (see (3.14)–(3.16)) we get

$$\begin{aligned} \sum_{i=1}^N \int_{x_i}^{x_{i+1}} (f^+(x) - f^*(x)) \sin(\omega x) dx &= \sum_{i=1}^{N-1} \left(\left(L \text{sign}(\sin(\omega x_i)) - \frac{\Delta f_i}{\Delta x_i} \right) \int_{x_i}^{\bar{x}_i} (x - x_i) \times \right. \\ &\quad \left. \sin(\omega x) dx + \frac{1}{2} \left(L - \frac{|\Delta f_i|}{\Delta x_i} \right) \text{sign}(\sin(\omega x_i)) \int_{\bar{x}_i}^{\bar{\bar{x}}_i} ((1 + \text{sign}(\Delta f_i)\text{sign}(\sin(\omega x_i))) \times \right. \\ &\quad \left. \left. \sin(\omega x) dx + \frac{1}{2} \left(L - \frac{|\Delta f_{i+1}|}{\Delta x_{i+1}} \right) \text{sign}(\sin(\omega x_{i+1})) \int_{\bar{\bar{x}}_i}^{x_{i+1}} ((1 - \text{sign}(\Delta f_{i+1})\text{sign}(\sin(\omega x_{i+1}))) \times \right. \right. \\ &\quad \left. \left. \sin(\omega x) dx \right) \right) \end{aligned}$$

$$\begin{aligned}
& (x - x_i) + (1 - \text{sign}(\Delta f_i) \text{sign}(\sin(\omega x_i))) (x_{i+1} - x) \sin(\omega x) dx + (L \text{sign}(\sin(\omega x_i)) + \\
& \frac{|\Delta f_i|}{\Delta x_i}) \int_{\bar{x}_i}^{x_{i+1}} (x_{i+1} - x) \sin(\omega x) dx + L \text{sign}(\sin \omega x_N) \int_{x_N}^{x_{N+1}} (x - x_N) \sin(\omega x) dx = \\
& \frac{L}{\omega} \left(\sum_{i=1}^{N-1} \left| \sin \left(\frac{\omega}{2} (x_{i+1} + x_i) \right) \right| \times \left(\frac{4}{\omega} (\sin^2 \frac{\omega \Delta x_i}{4} - \sin^2 \frac{\omega |\Delta f_i|}{4L}) \right) + \right. \\
& \left. \frac{2}{\omega} \sum_{i=1}^{N-1} \text{sign}(\Delta f_i) \cos \left(\frac{\omega}{2} (x_{i+1} + x_i) \right) \times \left(\sin \frac{\omega |\Delta f_i|}{2L} - \frac{|\Delta f_i|}{L \Delta x_i} \times \right. \right. \\
& \left. \left. \sin \frac{\omega \Delta x_i}{2} \right) + \frac{2}{\omega} \left| \sin \frac{\omega \Delta x_{N-1}}{2} \cos \left(\omega \left(b - \frac{\Delta x_{N-1}}{2} \right) \right) - \Delta x_{N-1} \cos(\omega b) \right| \right). \quad (3.19)
\end{aligned}$$

In a similar way, taking into account the explicit representation of the minorant of class $C_{1,L,N}^1$, we obtain

$$\begin{aligned}
& \sum_{i=1}^N \int_{x_i}^{x_{i+1}} (f^*(x) - f^-(x)) \sin(\omega x) dx = \sum_{i=1}^{N-1} \left(\left(L \text{sign}(\sin(\omega x_i)) - \frac{|\Delta f_i|}{\Delta x_i} \right) \int_{x_i}^{\bar{x}_i} (x - x_i) \times \right. \\
& \left. \sin(\omega x) dx + \frac{1}{2} \left(L - \frac{|\Delta f_i|}{\Delta x_i} \right) \text{sign}(\sin(\omega x_i)) \int_{\bar{x}_i}^{x_{i+1}} ((1 - \text{sign}(\Delta f_i) \text{sign}(\sin(\omega x_i))) \times \right. \\
& \left. (x - x_i) + (1 + \text{sign}(\Delta f_i) \text{sign}(\sin(\omega x_i))) (x_{i+1} - x)) \sin(\omega x) dx + (L \text{sign}(\sin(\omega x_i)) - \right. \\
& \left. \frac{|\Delta f_i|}{\Delta x_i}) \int_{\bar{x}_i}^{x_{i+1}} (x_{i+1} - x) \sin(\omega x) dx \right) + L \text{sign}(\sin(\omega x_N)) \int_{x_N}^{x_{N+1}} (x - x_N) \sin(\omega x) dx = \\
& \frac{L}{\omega} \left(\sum_{i=1}^{N-1} \left| \sin \left(\frac{\omega}{2} (x_{i+1} + x_i) \right) \right| \times \left(\frac{4}{\omega} (\sin^2 \frac{\omega \Delta x_i}{4} - \sin^2 \frac{\omega |\Delta f_i|}{4L}) \right) - \right. \\
& \left. \frac{2}{\omega} \sum_{i=1}^{N-1} \text{sign}(\Delta f_i) \cos \left(\frac{\omega}{2} (x_{i+1} + x_i) \right) \times \left(\sin \frac{\omega |\Delta f_i|}{2L} - \frac{|\Delta f_i|}{L \Delta x_i} \times \right. \right. \\
& \left. \left. \sin \frac{\omega \Delta x_i}{2} \right) + \frac{2}{\omega} \left| \sin \frac{\omega \Delta x_{N-1}}{2} \cos \left(\omega \left(b - \frac{\Delta x_{N-1}}{2} \right) \right) - \Delta x_N \cos(\omega b) \right| \right). \quad (3.20)
\end{aligned}$$

Therefore, substituting (3.19) and (3.20) into (3.18) we derive that

$$\begin{aligned}
v(C_{1,L,N}^1, R_2(\omega, S), f) & \leq \frac{L}{\omega} \left(\sum_{i=1}^{N-1} \left| \sin \left(\frac{\omega}{2} (x_{i+1} + x_i) \right) \right| \times \right. \\
& \left. \left(\frac{4}{\omega} (\sin^2 \frac{\omega \Delta x_i}{4} - \sin^2 \frac{\omega |\Delta f_i|}{4L}) \right) + \frac{2}{\omega} \left(\sin \frac{\omega \Delta x_N}{2} \cos \left(\omega \left(b - \frac{\Delta x_N}{2} \right) \right) - \right. \right. \\
& \left. \left. \Delta x_N \cos(\omega b) \right) + \frac{2}{\omega} \left| \sum_{i=1}^{N-1} \text{sign}(\Delta f_i) \cos \left(\frac{\omega}{2} (x_{i+1} + x_i) \right) \left(\sin \frac{\omega |\Delta f_i|}{2L} - \right. \right. \right. \\
& \left. \left. \left. \sin \frac{\omega \Delta x_i}{2} \right) + \frac{2}{\omega} \left| \sin \frac{\omega \Delta x_{N-1}}{2} \cos \left(\omega \left(b - \frac{\Delta x_{N-1}}{2} \right) \right) - \Delta x_N \cos(\omega b) \right| \right|
\end{aligned}$$

$$\left| \frac{|\Delta f_i|}{L\Delta x_i} \sin \frac{\omega \Delta x_i}{2} \right) \Bigg| \Bigg), \quad (3.21)$$

which completes the proof. ■

Remark 3.1 An analogous result takes place for quadrature formula (3.4) when computing (3.2) in class $C_{1,L,N}^1$, provided condition C1 is satisfied. More precisely we have the following estimate

$$\begin{aligned} v(C_{1,L,N}^1, R_3(\omega, S), f) &\leq \\ &\frac{L}{\omega} \left(\sum_{i=1}^{N-1} \left| \cos\left(\frac{\omega}{2}(x_{i+1} + x_i)\right) \right| \left(\frac{4}{\omega} \left(\sin^2 \frac{\omega \Delta x_i}{4} - \sin^2 \frac{\omega |\Delta f_i|}{4L} \right) \right) + \right. \\ &\left. \frac{2}{\omega} \left| \Delta x_N \sin(\omega b) - \sin \frac{\omega \Delta x_N}{2} \times \cos\left(\omega(b - \frac{\Delta x_N}{2})\right) \right| \right) + \\ &\frac{2L}{\omega^2} \left| \sum_{i=1}^{N-1} \text{sign}(\Delta f_i) \sin\left(\frac{\omega}{2}(x_{i+1} + x_i)\right) \left(\sin \frac{\omega |\Delta f_i|}{2L} - \frac{|\Delta f_i|}{L\Delta x_i} \sin \frac{\omega \Delta x_i}{2} \right) \right|. \quad (3.22) \end{aligned}$$

The above results were obtained in the case of weak oscillations (see Condition C1). Now let us consider the case of strong oscillations. Let $f(x) \in C_{1,L,N}^1$ and let us assume that N nodes x_i , $i = 1, \dots, N$ are included in the number of zeros of $\sin(\omega x)$ ($\cos(\omega x)$), i.e. $N \ll \left(\left[\frac{\omega}{\pi}(b-a)\right] + 1\right)$. We refer to this condition as Condition C2. We also assume that there are k_i oscillations of function $\sin(\omega x)$ ($\cos(\omega x)$), $i = 1, \dots, N$ on segments $[x_i, x_{i+1}]$. Then the following result holds.

Theorem 3.2 Let $f(x) \in C_{1,L,N}^1$, $N < |\omega|$, condition C2 satisfied and $\{f_i\}_{i=1,\dots,N}$ are given on $[a, b]$ at nodes x_i , $i = 1, \dots, N$ of the grid with fixed ends $x_1 = a$, $x_{N+1} = b$. Then quadrature formula (3.3) for computing integral (3.1) is optimal-by-order with a constant not exceeding 2. This result holds with the following estimate

$$\begin{aligned} \bar{v}(C_{1,L,N}^1, R_2(\omega, S), f) &\leq \frac{2L}{\omega^2} \left(\left[\frac{|\omega|}{\pi} \right] - \sum_{i=1}^{N-1} \left(\left[\frac{|\omega|}{\pi} x_{i+1} \right] - \left[\frac{|\omega|}{\pi} x_i \right] \right) \times \right. \\ &\left. \sin^2 \frac{\pi |\Delta f_i|}{4L\Delta x_i} \right) + \frac{L}{|\omega|} \left| \frac{2}{\omega} \sin\left(\frac{\omega}{2}(b - \left[\frac{|\omega|}{\pi} \right] \frac{\pi}{|\omega|})\right) \cos\left(\frac{\omega}{2}(2 + \left[\frac{|\omega|}{\pi} \right]) \times \right. \right. \\ &\left. \left. \frac{\pi}{|\omega|} - b\right) - (1 - \left[\frac{|\omega|}{\pi} \right] \frac{\pi}{|\omega|}) \cos(\omega b) \right| + \frac{2L}{\omega^2} \left| \sum_{i=1}^{N-1} \text{sign}(\Delta f_i) \times \right. \\ &\left. \sum_{k=0}^{k_i-1} \cos\left(\frac{\omega}{2}(x_{i,k+1} + x_{i,k})\right) \left(\sin\left(\frac{\omega}{2} \frac{\pi |\Delta f_i|}{L\Delta x_i |\omega|}\right) - \frac{|\Delta f_i|}{L\Delta x_i} \sin \frac{\omega \pi}{2|\omega|} \right) \right|, \quad (3.23) \end{aligned}$$

where $\Delta f_i = f_{i+1} - f_i$, $\Delta x_i = x_{i+1} - x_i$, $k_i = [\lfloor \omega |x_{i+1}|/\pi \rfloor - \lfloor \omega |x_i|/\pi \rfloor]$, $x_{i,k} = \frac{\pi}{2|\omega|} (2[\frac{\omega}{\pi} x_i + \frac{1}{2}] + 2k + 1)$ are zeros of function $\sin(\omega x)$ on $[x_i, x_{i+1}]$, $x_{i,0} = x_i$, $x_{i,k_i} = x_{i+1}$, $k = 0, \dots, k_i$, $i = 1, \dots, N$.

Proof. We showed (see the proof of Theorem 2.1) that $S(x, L) \in C_{1,L,N}^1$. Therefore, spline $S(x, L)$ satisfies inequalities (3.12) and (3.13). Let $x_N = x_{N-1, [\omega/\pi]}$ be the last zero of function $\sin(\omega x)$ on $[x_{N-1}, x_{N+1}]$. Then, taking into account oscillations of function $\sin(\omega x)$, the limit functions of class $C_{1,L,N}^1$, $f^+(x)$ and $f^-(x)$, will have the form

$$f^+(x) = \bigcup_{i,k} f_{i,k}^+(x), \quad f^-(x) = \bigcup_{i,k} f_{i,k}^-(x), \quad (3.24)$$

where $f_{i,k}^+(x)$, $f_{i,k}^-(x)$ are defined as follows.

(a) For $x \in [x_{i,k}, \bar{x}_{i,k}]$, $\bar{x}_{i,k} = (x_{i,k} + x_{i,k+1})/2 - |\Delta f_i| \pi / (2L \Delta x_i |\omega|)$ we have

$$\begin{aligned} f_{i,k}^+(x) &= f_i + \frac{\Delta f_i}{\Delta x_i} (x_{i,k} - x_i) + L(x - x_{i,k}) \text{sign}(\sin(\omega x_{i,k})), \\ f_{i,k}^-(x) &= f_i + \frac{\Delta f_i}{\Delta x_i} (x_{i,k} - x_i) - L(x - x_{i,k}) \text{sign}(\sin(\omega x_{i,k})) \end{aligned} \quad (3.25)$$

(b) For $x \in [\bar{x}_{i,k}, \bar{\bar{x}}_{i,k}]$, $\bar{\bar{x}}_{i,k} = (x_{i,k} + x_{i,k+1})/2 + |\Delta f_i| \pi / (2L \Delta x_i |\omega|)$ we have

$$\begin{aligned} f_{i,k}^+(x) &= \frac{1}{2}(1 + \text{sign}(\Delta f_i) \text{sign}(\sin(\omega x_{i,k}))) \left(f_i + \frac{\Delta f_i}{\Delta x_i} (x_{i,k} - x_i) + \right. \\ &\quad L(x - x_{i,k}) \text{sign}(\sin(\omega x_{i,k})) \left. + \frac{1}{2}(1 - \text{sign}(\Delta f_i) \text{sign}(\sin(\omega x_{i,k}))) \times \right. \\ &\quad \left(f_i + \frac{\Delta f_i}{\Delta x_i} (x_{i,k+1} - x_i) + L(x_{i,k+1} - x) \text{sign}(\sin(\omega x_{i,k})) \right), \\ f_{i,k}^-(x) &= \frac{1}{2}(1 - \text{sign}(\Delta f_i) \text{sign}(\sin(\omega x_{i,k}))) \left(f_i + \frac{\Delta f_i}{\Delta x_i} (x_{i,k} - x_i) - \right. \\ &\quad L(x - x_{i,k}) \text{sign}(\sin(\omega x_{i,k})) \left. + \frac{1}{2}(1 + \text{sign}(\Delta f_i) \text{sign}(\sin(\omega x_{i,k}))) \times \right. \\ &\quad \left(f_i + \frac{\Delta f_i}{\Delta x_i} (x_{i,k+1} - x_i) - L(x_{i,k+1} - x) \text{sign}(\sin(\omega x_{i,k})) \right) \end{aligned} \quad (3.26)$$

(c) For $x \in [\bar{x}_{i,k}, x_{i,k}]$ we have

$$\begin{aligned} f_{i,k}^+(x) &= f_i + \frac{\Delta f_i}{\Delta x_i} (x_{i,k+1} - x_i) + L(x_{i,k+1} - x) \text{sign}(\sin(\omega x_{i,k})), \\ f_{i,k}^-(x) &= f_i + \frac{\Delta f_i}{\Delta x_i} (x_{i,k+1} - x_i) - L(x_{i,k+1} - x) \text{sign}(\sin(\omega x_{i,k})) \end{aligned} \quad (3.27)$$

In formulae (3.25)–(3.27) $\text{sign}(\sin(\omega x_{i,k}))$ denotes the signum of function $\sin(\omega x)$ on intervals $[x_{i,k}, x_{i,k+1}]$, $k = 0, \dots, k_i - 1$, $i = 1, \dots, N$. If further we chose an approximation $f_{i,k}^*(x)$ in the following form

$$f_{i,k}^*(x) = \begin{cases} S(x, L), & x \in [x_{i,k}, x_{i,k+1}], \quad k = 0, \dots, k_i - 1, \quad i = 1, \dots, N, \\ f_N, & x \in [x_{N-1, [\omega/\pi]}, x_{N+1}], \end{cases} \quad (3.28)$$

then in the case of strong oscillations, the error of quadrature formula (3.3) for computing integral (3.1) in class $C_{1,L,N}^1$ can be estimated as follows

$$\begin{aligned} \bar{v}(C_{1,L,N}^1, R_2(\omega, S), f) &\leq \max \left\{ \sum_{i=1}^{N+1} \sum_{k=0}^{k_i-1} \int_{x_{i,k}}^{x_{i,k+1}} (f_{i,k}^+(x) - f_{i,k}^*(x)) \sin(\omega x) dx, \right. \\ &\quad \left. \sum_{i=1}^{N+1} \sum_{k=0}^{k_i-1} \int_{x_{i,k}}^{x_{i,k+1}} (f_{i,k}^*(x) - f_{i,k}^-(x)) \sin(\omega x) dx \right\}. \end{aligned} \quad (3.29)$$

Using (3.25)–(3.28) integrals in (3.29) can be found in the explicit form. For the first component under the maximum sign in (3.29) we have

$$\begin{aligned} \sum_{i=1}^{N+1} \sum_{k=0}^{k_i-1} \int_{x_{i,k}}^{x_{i,k+1}} (f_{i,k}^+(x) - f_{i,k}^*(x)) \sin(\omega x) dx &= \sum_{i=1}^{N+1} \sum_{k=0}^{k_i-1} \left(\int_{x_{i,k}}^{\bar{x}_{i,k}} \left(f_i + \frac{\Delta f_i}{\Delta x_i} (x_{i,k} - x_i) + \right. \right. \\ &\quad \left. L(x - x_{i,k}) \text{sign}(\sin(\omega x_{i,k})) - f_i - \frac{\Delta f_i}{\Delta x_i} (x - x_i) \right) \sin(\omega x) dx + \int_{\bar{x}_{i,k}}^{\bar{x}_{i,k}} \frac{1}{2} ((1 + \text{sign}(\Delta f_i) \times \\ &\quad \text{sign}(\sin(\omega x_{i,k}))) \left(f_i + \frac{\Delta f_i}{\Delta x_i} (x_{i,k} - x_i) + L(x - x_{i,k}) \text{sign}(\sin(\omega x_{i,k})) - f_i - \frac{\Delta f_i}{\Delta x_i} (x - x_i) \right) + \\ &\quad (1 - \text{sign}(\Delta f_i) \text{sign}(\sin(\omega x_{i,k}))) \left(f_i + \frac{\Delta f_i}{\Delta x_i} (x_{i,k+1} - x_i) + L(x_{i,k+1} - x) \text{sign}(\sin(\omega x_{i,k})) - \right. \\ &\quad \left. f_i - \frac{\Delta f_i}{\Delta x_i} (x - x_i) \right) \sin(\omega x) dx + \int_{\bar{x}_{i,k}}^{x_{i,k+1}} \left(f_i + \frac{\Delta f_i}{\Delta x_i} (x_{i,k+1} - x_i) + L(x_{i,k+1} - x) \times \right. \\ &\quad \left. \text{sign}(\sin(\omega x_{i,k})) - f_i - \frac{\Delta f_i}{\Delta x_i} (x - x_i) \right) \sin(\omega x) dx + \int_{x_{N-1}, [\lfloor \omega/\pi \rfloor]}^{x_{N+1}} L(x - x_{N-1}, [\lfloor \omega/\pi \rfloor]) \times \\ &\quad \text{sign}(\sin(\omega x_{N-1}, [\lfloor \omega/\pi \rfloor])) \sin(\omega x) dx \Big) = \frac{2L}{\omega^2} \left(\left[\frac{|\omega|}{\pi} \right] - \sum_{i=0}^{N-1} \left(\left[\frac{|\omega|}{\pi} x_{i+1} \right] - \left[\frac{|\omega|}{\pi} x_i \right] \right) \times \right. \\ &\quad \left. \sin^2 \frac{|\Delta f_i| \pi}{4L \Delta x_i} + \sum_{i=1}^{N-1} \text{sign}(\Delta f_i) \times \sum_{k=0}^{k_i-1} \cos \left(\frac{\omega}{2} (x_{i,k+1} + x_{i,k}) \right) \left(\sin \frac{\omega |\Delta f_i| \pi}{2L \Delta x_i |\omega|} - \frac{|\Delta f_i|}{L \Delta x_i} \times \right. \right. \\ &\quad \left. \left. \sin \left(\frac{\omega \pi}{2|\omega|} \right) \right) + \frac{L}{|\omega|} \left| \frac{2}{\omega} \sin \left(\frac{\omega}{2} (b - \left[\frac{|\omega|}{\pi} \right] \frac{\pi}{|\omega|}) \right) \cos \left(\frac{\omega}{2} \times (2 + (\left[\frac{|\omega|}{\pi} \right] \frac{\pi}{|\omega|} - b)) \right) \right. \\ &\quad \left. - \left(1 - \left[\frac{|\omega|}{\pi} \right] \frac{\pi}{|\omega|} \right) \cos(\omega b) \right|. \end{aligned} \quad (3.30)$$

Using the explicit form of minorant function given by (3.25)–(3.27) and equality (3.28), we transform the second component under the maximum sign in (3.29)

$$\sum_{i=1}^{N+1} \sum_{k=0}^{k_i-1} \int_{x_{i,k}}^{x_{i,k+1}} (f_{i,k}^*(x) - f_{i,k}^-(x)) \sin(\omega x) dx = \sum_{i=1}^N \sum_{k=0}^{k_i-1} \left(\int_{x_{i,k}}^{\bar{x}_{i,k}} \left(f_i + \frac{\Delta f_i}{\Delta x_i} (x - x_i) - \right. \right.$$

$$\begin{aligned}
& f_i - \frac{\Delta f_i}{\Delta x_i} (x_{i,k} - x_i) + L(x - x_{i,k}) \operatorname{sign}(\sin(\omega x_{i,k})) \Big) \sin(\omega x) dx + \frac{1}{2} \int_{\bar{x}_{i,k}}^{\bar{x}_{i,k}} \left((1 - \operatorname{sign}(\Delta f_i)) \times \right. \\
& \left. \operatorname{sign}(\sin(\omega x_{i,k})) \right) \left(f_i + \frac{\Delta f_i}{\Delta x_i} (x - x_i) - f_i - \frac{\Delta f_i}{\Delta x_i} (x_{i,k} - x_i) + L(x - x_{i,k}) \operatorname{sign}(\sin(\omega x_{i,k})) \right) + \\
& (1 + \operatorname{sign}(\Delta f_i) \operatorname{sign}(\sin(\omega x_{i,k}))) \left(f_i + \frac{\Delta f_i}{\Delta x_i} (x - x_i) - f_i - \frac{\Delta f_i}{\Delta x_i} (x_{i,k+1} - x_i) + \right. \\
& L(x_{i,k+1} - x) \operatorname{sign}(\sin(\omega x_{i,k}))) \sin(\omega x) dx + \int_{\bar{x}_{i,k}}^{x_{i,k}} \left(f_i + \frac{\Delta f_i}{\Delta x_i} (x - x_i) - f_i - \right. \\
& \left. \frac{\Delta f_i}{\Delta x_i} (x_{i,k+1} - x_i) + L(x_{i,k} - x) \operatorname{sign}(\sin(\omega x_{i,k})) \right) \sin(\omega x) dx + \int_{x_{N-1}, [\omega/\pi]}^{x_{N+1}} L \times \\
& (x - x_{N-1}, [\omega/\pi]) \operatorname{sign}(\sin(\omega x_{N-1}, [\omega/\pi])) \sin(\omega x) dx = \frac{2L}{\omega^2} \left(\left[\frac{|\omega|}{\pi} \right] - \sum_{i=1}^{N-1} \left(\left[\frac{|\omega|}{\pi} x_{i+1} \right] - \right. \right. \\
& \left. \left[\frac{|\omega|}{\pi} x_i \right] \right) \sin^2 \frac{|\Delta f_i| \pi}{4L \Delta x_i} - \sum_{i=1}^{N-1} \operatorname{sign}(\Delta f_i) \times \sum_{k=0}^{k_i-1} \cos\left(\frac{\omega}{2}(x_{i,k+1} + x_{i,k})\right) \\
& \left(\sin\left(\frac{\omega |\Delta f_i| \pi}{2L \Delta x_i |\omega|}\right) - \frac{|\Delta f_i|}{L \Delta x_i} \times \sin\left(\frac{\omega}{2} \frac{\pi}{|\omega|}\right) \right) + \frac{L}{|\omega|} \left| \frac{2}{\omega} \sin\left(\frac{\omega}{2}(b - \left[\frac{|\omega|}{\pi} \right] \frac{\pi}{|\omega|})\right) \cos\left(\frac{\omega}{2} \times \right. \right. \\
& \left. \left. (2 + (\left[\frac{|\omega|}{\pi} \right] \frac{\pi}{|\omega|} - b)) \right) - (1 - \left[\frac{|\omega|}{\pi} \right] \frac{\pi}{|\omega|}) \cos(\omega b) \right|. \tag{3.31}
\end{aligned}$$

Substituting resulting equalities (3.30) and (3.31) into (3.29) yields

$$\begin{aligned}
& \bar{v}(C_{1,L,N}^1, R_2(\omega, S), f) \leq \frac{2L}{\omega^2} \left(\left[\frac{|\omega|}{\pi} \right] - \sum_{i=1}^{N-1} \left(\left[\frac{|\omega|}{\pi} x_{i+1} \right] - \left[\frac{|\omega|}{\pi} x_i \right] \right) \sin^2 \frac{\pi |\Delta f_i|}{4L \Delta x_i} + \right. \\
& \left. \left| \sum_{i=1}^{N-1} \operatorname{sign}(\Delta f_i) \sum_{k=0}^{k_i-1} \cos\left(\frac{\omega}{2}(x_{i,k+1} + x_{i,k})\right) \left(\sin\left(\frac{\omega}{2} \frac{\pi |\Delta f_i|}{L \Delta x_i |\omega|}\right) - \frac{|\Delta f_i|}{L \Delta x_i} \sin\left(\frac{\omega \pi}{2 |\omega|}\right) \right) \right| + \right. \\
& \left. \frac{L}{|\omega|} \left| \frac{2}{\omega} \sin\left(\frac{\omega}{2}(b - \left[\frac{|\omega|}{\pi} \right] \frac{\pi}{|\omega|})\right) \cos\left(\frac{\omega}{2} \left(2 + (\left[\frac{|\omega|}{\pi} \right] \frac{\pi}{|\omega|} - b) \right) \right) - \right. \right. \\
& \left. \left. \left(1 - \left[\frac{|\omega|}{\pi} \right] \frac{\pi}{|\omega|} \right) \cos(\omega b) \right|, \tag{3.32} \right.
\end{aligned}$$

that completes the proof. ■

Remark 3.2 If condition C2 is satisfied, then for quadrature formula (3.4) computing (3.2) in class $C_{1,L,N}^1$ we have the following estimate

$$\bar{v}(C_{1,L,N}^1, R_3(\omega, S), f) \leq \frac{2L}{\omega^2} \left(\left[\frac{|\omega|}{\pi} \right] - \sum_{i=1}^{N-1} \left(\left[\frac{|\omega|}{\pi} x_{i+1} \right] - \left[\frac{|\omega|}{\pi} x_i \right] \right) \cos^2 \frac{|\Delta f_i| \pi}{4L \Delta x_i} \right) +$$

$$\begin{aligned}
& \frac{L}{|\omega|} \left| \frac{2}{\omega} \cos \left(\frac{\omega}{2} \left(b - \left[\frac{|\omega|}{\pi} \right] \frac{\pi}{|\omega|} \right) \right) \times \cos \left(\frac{\omega}{2} \left(2 + \left(\left[\frac{|\omega|}{\pi} \right] \frac{\pi}{|\omega|} - b \right) \right) \right) - \right. \\
& \left. (1 - \left[\frac{|\omega|}{\pi} \right] \frac{\pi}{|\omega|}) \sin(\omega b) \right| + \frac{2L}{\omega^2} \left| \sum_{i=1}^{N-1} \text{sign}(\Delta f_i) \sum_{k=0}^{k_i-1} \sin \left(\frac{\omega}{2} (x_{i,k+1} + x_{i,k}) \right) \times \right. \\
& \left. \left(\sin \frac{\omega |\Delta f_i| \pi}{2L \Delta x_i |\omega|} - \frac{|\Delta f_i|}{L \Delta x_i} \sin \frac{\omega \pi}{2|\omega|} \right) \right|, \tag{3.33}
\end{aligned}$$

where $x_{i,k} = \frac{\pi}{|\omega|} \left(2 \left[\frac{|\omega|}{\pi} x_i + 1 \right] + 2k \right)$ are zeros of function $\cos(\omega x)$ on $[x_i, x_{i+1}]$, $x_{i,0} = x_i$, $x_{i,k_i} = x_{i+1}$, $k = 0, \dots, k_i$, $i = 1, \dots, N$.

4 Error estimates for optimal-by-order quadrature formulae in class $C_{1,L,N,\epsilon}^1$.

In this section we generalise results obtained in previous sections into class $C_{1,L,N,\epsilon}^1$. As above, we deal with the case when *a priori* information is given approximately. This precludes an efficient application of optimal-by-accuracy quadrature formulae that are based on the exact knowledge of *a priori* information.

Using the method of quasi-solutions in the construction of quadrature formulae for approximate calculation of integrals (3.1) and (3.2) in class $C_{1,L,N,\epsilon}^1$, we assume that the Lipschitz constant, L , and a certain accuracy estimate for the definition of function $f(x)$ in N nodes of an arbitrary grid are given. In this case the application of a linear spline $S(x, L)$ as an approximation of the function $f(x)$, allows smoothing input data and defining ϵ with a higher precision. In some applications the Lipschitz constant is not known, but we can estimate accuracy of the definition of $f(x)$ in nodes of the grid. Then it is reasonable to apply quadrature formulae constructed by the residual method where the value of L is not used.

We note that both quadrature formulae for computing (3.1) (those constructed with the method of quasi-solutions and the residual method), has the form (3.3). The difference consists of the fact that in the method of quasi-solutions values \tilde{f}_i , $i = 1, \dots, N$ are computed with (2.9), making use of constant L . In the residual method, for computing \tilde{f}_i , $i = 1, \dots, N$ we change constant L in (2.9) for constant M defined by (2.11).

Let $\Delta \tilde{f}_i = \tilde{f}_{i+1} - \tilde{f}_i$, $\Delta x_i = x_{i+1} - x_i$ and $Q = \Delta \tilde{f}_i / \Delta x_i$, $i = 1, \dots, N-1$. Then we introduce the following notation

$$\Delta_i = \frac{(\xi_i^+ + \xi_{i+1}^+) + (\xi_i^- + \xi_{i+1}^-) - (\Delta \xi_i^+ - \Delta \xi_i^-) \text{sign}(\sin(\omega x))}{4(Q + L) \text{sign}(\sin(\omega x))}, \tag{4.1}$$

where ξ_i^\pm , $i = 1, \dots, N$ are respectively maximal and minimal admissible solutions of the system of linear inequalities

$$\begin{cases} -\epsilon_i \leq \xi_i \leq \epsilon_i, & i = 1, \dots, N, \\ -L \Delta x_i - \Delta \tilde{f}_i \leq \xi_{i+1} - \xi_i \leq L \Delta x_i - \Delta \tilde{f}_i, & i = 1, \dots, N-1. \end{cases} \tag{4.2}$$

Let further

$$n_i = \left[\frac{|\omega|}{\pi} x_{i+1} \right] - \left[\frac{|\omega|}{\pi} x_i \right] \quad (4.3)$$

be the number of oscillations of function $\sin \omega x$ on $[x_i, x_{i+1}]$ and

$$x_{i,k} = \left(\left[\frac{|\omega|}{\pi} x_i \right] + k \right) \frac{\pi}{|\omega|}, \quad k = 1, \dots, n_i \quad (4.4)$$

be zeros of function $\sin(\omega x)$ on $[x_i, x_{i+1}]$, $x_{i,0} = \bar{x}_{i,0} = x_i$, $x_{i,n_i+1} = \bar{x}_{i,n_i+1} = x_{i+1}$, $x^1 \in [x_i, x_{i+1}]$, $x^2 \in [x_{i,n_i}, x_{i+1}]$. Finally, let

$$\begin{aligned} \tilde{x}_{i,1} &= x_{i,1} + \left(x_{i,1} - \frac{x_i + x_{i,2}}{2} \right) \frac{L \operatorname{sign}(\sin(\omega x)) - Q}{L \operatorname{sign}(\sin(\omega x)) + Q} + \Delta_i, \\ \tilde{x}_{i,n_i} &= x_{i,n_i} + \left(x_{i,n_i} - \frac{x_{i,n_i-1} + x_{i+1}}{2} \right) \frac{L \operatorname{sign}(\sin(\omega x)) - Q}{L \operatorname{sign}(\sin(\omega x)) + Q} + \Delta_i, \quad i = 1, \dots, N-1; \\ \tilde{x}_{i,k} &= x_{i,k} + \Delta_i, \quad k = 2, \dots, n_i-1, \end{aligned} \quad (4.5)$$

$$\Delta \tilde{f}_{i,k} = Q \Delta \tilde{x}_{i,k}, \quad \Delta \tilde{x}_{i,k} = \tilde{x}_{i,k+1} - \tilde{x}_{i,k}; \quad (4.6)$$

$$\hat{x}_{i,k} = \frac{\tilde{x}_{i,k} + \tilde{x}_{i,k+1}}{2} - \delta_{i,k}, \quad \hat{x}_{i,k} = \frac{\tilde{x}_{i,k} + \tilde{x}_{i,k+1}}{2} + \delta_{i,k}, \quad (4.7)$$

where

$$\begin{aligned} \delta_{i,k} &= \frac{1}{4L} (2|\Delta \tilde{f}_{i,k}| + (\xi_{i+1}^+(1 - \operatorname{sign}(x_{i+1} - x_{i,k+1})) - \\ &\quad \xi_i^+(1 - \operatorname{sign}(x_{i,k} - x_i))) (\operatorname{sign}(\Delta \tilde{f}_i) + 1) + (\xi_{i+1}^-(1 - \operatorname{sign}(x_{i+1} - \\ &\quad x_{i,k+1})) - \xi_i^-(1 - \operatorname{sign}(x_{i,k} - x_i))) (\operatorname{sign}(\Delta \tilde{f}_i) - 1)), \end{aligned} \quad (4.8)$$

$k_i = 1, \dots, n_i - 1$, $i = 1, \dots, N - 1$.

Now we are in the position to formulate the main result of this section.

Theorem 4.1 *Let $f(x) \in C_{1,L,N,\epsilon}^1$, condition C2 satisfied and $\{\tilde{f}_i\}_{i=1,\dots,N}$ are given on $[a, b]$ in x_i , $i = 1, \dots, N$ nodes of the grid with fixed ends $x_1 = a$, $x_N = b$. Then quadrature formula (3.3) for computing (3.1) constructed by the residual method or the method of quasi-solutions is optimal-by-order with constant not exceeding 2. The following estimate holds*

$$\begin{aligned} v(C_{1,L,N,\epsilon}^1, R_2(\omega, S), f) &\leq \frac{L}{\omega^2} \sum_{i=1}^{N-1} \left(2 \left| \sin \left(\omega \left(\frac{x_i + \tilde{x}_{i,1}}{2} + \frac{\xi_i^- - \xi_i^+}{2L} \right) \right) \right| \times \right. \\ &\quad \left. \cos \left(\omega \frac{2|\Delta \tilde{f}_{i,0}| - |\xi_i^+ + \xi_i^-| \operatorname{sign}(\Delta \tilde{f}_i)}{2L} \right) - |\sin(\omega x_i)| + \frac{\omega(\xi_i^+ - \xi_i^-) \operatorname{sign}(\sin(\omega x^1))}{2L} \cos(\omega x_i) + \right. \\ &\quad \left. \dots \right. \end{aligned}$$

$$\begin{aligned}
& 2 \left(\left| \sin \left(\omega \frac{\tilde{x}_{i,1} + \tilde{x}_{i,2}}{2} \right) \right| + (n_i - 3) \left| \sin \left(\omega \frac{\tilde{x}_{i,2} + \tilde{x}_{i,3}}{2} \right) \right| + \left| \sin \left(\omega \frac{\tilde{x}_{i,n_i-1} + x_{i,n_i}}{2} \right) \right| \right) \cos \frac{\pi Q}{2L} - \\
& |\sin(\omega x_{i+1})| - \frac{\omega(\xi_{i+1}^+ - \xi_{i+1}^-)\text{sign}(\sin(\omega x^2))}{2L} \cos(\omega x_{i+1}) + 2 \left| \sin \left(\omega \left(\frac{\tilde{x}_{i,n_i} + x_{i+1}}{2} + \right. \right. \right. \\
& \left. \left. \left. \frac{\xi_{i+1}^+ - \xi_{i+1}^-}{2} \right) \right) \right| \cos \left(\omega \frac{2|\Delta \tilde{f}_{i,n_i}| + |\xi_{i+1}^+ + \xi_{i+1}^-| \text{sign}(\Delta \tilde{f}_i)}{2L} \right) + \left| \frac{1}{\omega^2} \sum_{i=1}^{N-1} \left(\sum_{k=0}^{n_i} L \text{sign}(\Delta \tilde{f}_i) \times \right. \right. \\
& (\sin(\omega \hat{x}_{i,k}) - \sin(\omega \tilde{x}_{i,k})) - \frac{\Delta \tilde{f}_i}{\Delta x_i} (\sin(\omega x_{i+1}) - \sin(\omega x_i)) \left. \right) + \frac{1}{\omega} \left((\hat{f}_N - \tilde{f}_N) \cos(\omega x_{N-1}) - \right. \\
& \left. (\hat{f}_1 - \tilde{f}_1) \cos(\omega x_1) \right). \tag{4.9}
\end{aligned}$$

Proof. First, we consider quadrature formula (3.3) constructed by the method of quasi-solutions. From the definition of quantities \tilde{f}_i^+ , \tilde{f}_i^- , $i = 1, \dots, N$ follows that $S(x_i, L) \in [\tilde{f}_i - \epsilon, \tilde{f}_i + \epsilon]$, $i = 1, \dots, N$, and that $|S'(x, L)| \leq L$. Indeed, for any $x \in [x_i, x_{i+1}]$, $i = 1, \dots, N-1$ we have

$$\begin{aligned}
|S'(x, L)| &= \left| \frac{\tilde{f}_{i+1} - \tilde{f}_i}{x_{i+1} - x_i} \right| = \left| \frac{\tilde{f}_{i+1}^+ + \tilde{f}_{i+1}^- - \tilde{f}_i^+ - \tilde{f}_i^-}{2(x_{i+1} - x_i)} \right| = \frac{1}{2(x_{i+1} - x_i)} \times \\
&\left| \max_{1 \leq j \leq N} (\tilde{f}_j - L|x_j - x_{i+1}|) - \max_{1 \leq j \leq N} (-\tilde{f}_j - L|x_j - x_{i+1}|) - \max_{1 \leq j \leq N} (\tilde{f}_j - L|x_j - x_i|) + \right. \\
&\left. \max_{1 \leq j \leq N} (-\tilde{f}_j - L|x_j - x_i|) \right| = \frac{1}{2(x_{i+1} - x_i)} \times \left| \tilde{f}_{j_1} - L|x_{j_1} - x_{i+1}| + \tilde{f}_{j_2} + \right. \\
&\left. L|x_{j_2} - x_{i+1}| - \tilde{f}_{j_3} + L|x_{j_3} - x_i| - \tilde{f}_{j_4} - L|x_{j_4} - x_i| \right|, \tag{4.10}
\end{aligned}$$

where j_k , $k = 1, \dots, 4$ are the values of index j on which the following values are achieved $\max_{1 \leq j \leq N} (\tilde{f}_j - L|x_j - x_{i+1}|)$, $\max_{1 \leq j \leq N} (-\tilde{f}_j - L|x_j - x_{i+1}|)$, $\max_{1 \leq j \leq N} (\tilde{f}_j - L|x_j - x_i|)$ and $\max_{1 \leq j \leq N} (-\tilde{f}_j - L|x_j - x_i|)$. If $(-\tilde{f}_{j_2} - L|x_{j_2} - x_{i+1}|) = \max_{1 \leq j \leq N} (-\tilde{f}_j - L|x_j - x_{i+1}|)$, then $\tilde{f}_{j_2} + L|x_{j_2} - x_{i+1}| = \min_{1 \leq j \leq N} (\tilde{f}_j + L|x_j - x_{i+1}|)$. Hence changing $(\tilde{f}_{j_2} + L|x_{j_2} - x_{i+1}|)$ for $(\tilde{f}_{j_4} + L|x_{j_4} - x_{i+1}|)$, we only increase the value of the expression $(\max_{1 \leq j \leq N} (\tilde{f}_j - L|x_j - x_{i+1}|) - \max_{1 \leq j \leq N} (-\tilde{f}_j - L|x_j - x_{i+1}|))$. In a similar way, since $(\tilde{f}_{j_3} - L|x_{j_3} - x_i| = \max_{1 \leq j \leq N} (\tilde{f}_j - L|x_j - x_i|))$, we can change it for $(\tilde{f}_{j_1} - L|x_{j_1} - x_i|)$ increasing the expression $(-\max_{1 \leq j \leq N} (\tilde{f}_j - L|x_j - x_i|) + \max_{1 \leq j \leq N} (-\tilde{f}_j - L|x_j - x_i|))$. Therefore, from (4.10) we get

$$\begin{aligned}
|S'(x, L)| &\leq \frac{1}{2(x_{i+1} - x_i)} \left| \tilde{f}_{j_1} - L|x_{j_1} - x_{i+1}| + \tilde{f}_{j_4} + L|x_{j_4} - x_{i+1}| - \right. \\
&\left. \tilde{f}_{j_1} + L|x_{j_1} - x_i| - \tilde{f}_{j_4} - L|x_{j_4} - x_i| \right| \leq \frac{L}{2(x_{i+1} - x_i)} \times \\
&(|x_{j_1} - x_i - x_{j_1} + x_{i+1}| + |x_{j_4} - x_{i+1} - x_{j_4} + x_i|) = L. \tag{4.11}
\end{aligned}$$

Hence, $S(x, L) \in C_{1,L,N,\epsilon}^1$.

Error estimates of quadrature formula (3.3) can be obtained on the basis of inequalities (3.12) and (3.13). Indeed, assume that on interval $[x_i, x_{i+1}]$ there are $n_i = \left[\frac{|\omega|}{\pi} x_{i+1} \right] - \left[\frac{|\omega|}{\pi} x_i \right]$ oscillations of function $\sin(\omega x)$, $i = 1, \dots, N - 1$. Then the limit functions of class $C_{1,L,N,\epsilon}^1$ on $[x_{i,k}, x_{i,k+1}]$, $k = 0, \dots, n_i - 1$ can be written in the explicit form as follows.

(a) For $x \in [x_{i,k}, \hat{x}_{i,k}]$ we have

$$f_{i,k}^\pm(x) = \tilde{f}_i + \left(\frac{\xi_i^+ + \xi_i^-}{2} \pm \frac{\xi_i^+ - \xi_i^-}{2} \operatorname{sign}(\sin(\omega x)) \right) \times \\ (1 - \operatorname{sign}(x_{i,k} - x_i)) + Q(\tilde{x}_{i,k} - x_i) \pm L(x - x_{i,k}) \operatorname{sign}(\sin(\omega x)) \quad (4.12)$$

(b) For $x \in [\hat{x}_{i,k}, \hat{\tilde{x}}_{i,k}]$ we have

$$f_{i,k}^\pm(x) = \tilde{f}_i + \frac{1}{2} \left(1 \pm \operatorname{sign}(\Delta \tilde{f}_i) \operatorname{sign}(\sin(\omega x)) \right) \left(\left(\frac{\xi_i^+ + \xi_i^-}{2} \pm \frac{\xi_i^+ - \xi_i^-}{2} \operatorname{sign}(\sin(\omega x)) \right) \times \right. \\ (1 - \operatorname{sign}(x_{i,k} - x_i)) + Q(\tilde{x}_{i,k} - x_i) \pm L(x - x_{i,k}) \operatorname{sign}(\sin(\omega x)) \Big) + \\ \frac{1}{2} \left(1 \mp \operatorname{sign}(\Delta \tilde{f}_i) \operatorname{sign}(\sin(\omega x)) \right) \times \left(\left(\frac{\xi_{i+1}^+ + \xi_{i+1}^-}{2} \pm \frac{\xi_{i+1}^+ - \xi_{i+1}^-}{2} \operatorname{sign}(\sin(\omega x)) \right) \times \right. \\ (1 - \operatorname{sign}(x_{i+1} - x_{i,k+1})) + Q(\tilde{x}_{i,k+1} - x_i) \pm L(x_{i,k+1} - x) \operatorname{sign}(\sin(\omega x)) \Big) \quad (4.13)$$

(c) For $x \in [\hat{\tilde{x}}_{i,k}, x_{i,k+1}]$ we have

$$f_{i,k}^\pm(x) = \tilde{f}_i + \left(\frac{\xi_{i+1}^+ + \xi_{i+1}^-}{2} \pm \frac{\xi_{i+1}^+ - \xi_{i+1}^-}{2} \operatorname{sign}(\sin(\omega x)) \right) \times \\ (1 - \operatorname{sign}(x_{i+1} - x_{i,k+1})) + Q(\tilde{x}_{i,k+1} - x_i) \pm L(x_{i,k+1} - x) \operatorname{sign}(\sin(\omega x)). \quad (4.14)$$

We recall that $x_1 = a$ and $x_N = b$. Then the error of quadrature formula (3.3) in class $C_{1,L,N,\epsilon}^1$ in the case of strong oscillations of $\sin \omega x$ can be estimated as follows

$$v(C_{1,L,N,\epsilon}^1, R_2(\omega, S), f) \leq \max \left(\sum_{i=1}^{N-1} \sum_{k=0}^{n_i} \int_{x_{i,k}}^{x_{i,k+1}} (f_{i,k}^+(x) - S(x, L)) \sin(\omega x) dx, \right. \\ \left. \sum_{i=1}^{N-1} \sum_{k=0}^{n_i} \int_{x_{i,k}}^{x_{i,k+1}} (S(x, L) - f_{i,k}^-(x)) \sin(\omega x) dx \right). \quad (4.15)$$

Using (4.12)–(4.14), the integrals in (4.15) can be calculated explicitly. Taking into account explicit expressions for the majorant function (see (4.12)–(4.14)), we have

$$\sum_{i=1}^{N-1} \sum_{k=0}^{n_i} \int_{x_{i,k}}^{x_{i,k+1}} (f_{i,k}^+(x) - S(x, L)) \sin(\omega x) dx = \sum_{i=1}^{N-1} \sum_{k=0}^{n_i} \left(\int_{x_{i,k}}^{\tilde{x}_{i,k}} (\tilde{f}_i - \tilde{f}_i + Q(\tilde{x}_{i,k} - x_i) - \right.$$

$$\begin{aligned}
& \frac{\Delta \hat{f}_i}{\Delta x_i} (x - x_i) + \frac{\xi_i^+ + \xi_i^-}{2} (1 - \text{sign}(x_{i,k} - x_i)) + L(x - x_{i,k}) \text{sign}(\sin(\omega x)) \Big) \sin(\omega x) dx + \\
& \frac{1}{2} \int_{\tilde{x}_{i,k}}^{\tilde{x}_{i,k}} \left((\tilde{f}_i - \hat{f}_i + Q(\tilde{x}_{i,k} - x_i) - \frac{\Delta \hat{f}_i}{\Delta x_i} (x - x_i) + \frac{\xi_i^+ + \xi_i^-}{2} (1 - \text{sign}(x_{i,k} - x_i)) + \right. \\
& L(x - x_{i,k}) \text{sign}(\sin(\omega x)) \Big) (1 + \text{sign}(\Delta \tilde{f}_i) \text{sign}(\sin(\omega x))) + \left(\tilde{f}_i - \hat{f}_i + Q(\tilde{x}_{i,k+1} - x_i) - \right. \\
& \left. \frac{\Delta \hat{f}_i}{\Delta x_i} (x - x_i) + \frac{\xi_{i+1}^+ + \xi_{i+1}^-}{2} (1 - \text{sign}(x_{i+1} - x_{i,k+1})) + L(x_{i,k} - x) \text{sign}(\sin(\omega x)) \right) \times \\
& (1 - \text{sign}(\Delta \tilde{f}_i) \text{sign}(\sin(\omega x))) \Big) \sin(\omega x) dx + \int_{\tilde{x}_{i,k}}^{x_{i,k+1}} \left(\tilde{f}_i - \hat{f}_i + Q(\tilde{x}_{i,k+1} - x_i) - \right. \\
& \left. \frac{\Delta \hat{f}_i}{\Delta x_i} (x - x_i) + \frac{\xi_{i+1}^+ + \xi_{i+1}^-}{2} \times (1 - \text{sign}(x_{i+1} - x_{i,k+1})) \right) \sin(\omega x) dx = \\
& \frac{L}{\omega^2} \sum_{i=1}^{N-1} \left(2 \left| \sin \left(\omega \left(\frac{x_i + \tilde{x}_{i,1}}{2} + \frac{\xi_i^- - \xi_i^+}{2L} \right) \right) \right| \cos \left(\omega \frac{2|\Delta \tilde{f}_{i,0}| - |\xi_i^+ + \xi_i^-| \text{sign}(\Delta \tilde{f}_i)}{2L} \right) - \right. \\
& |\sin(\omega x_i)| + \frac{\omega(\xi_i^+ - \xi_i^-) \text{sign}(\sin(\omega x^1))}{2L} \cos(\omega x_i) + 2 \left(\left| \sin \left(\omega \frac{\tilde{x}_{i,1} + \tilde{x}_{i,2}}{2} \right) \right| + \right. \\
& (n_i - 3) \left| \sin \left(\omega \frac{\tilde{x}_{i,2} + \tilde{x}_{i,3}}{2} \right) \right| + \left| \sin \left(\omega \frac{\tilde{x}_{i,n_i-1} + \tilde{x}_{i,n_i}}{2} \right) \right| \left. \right) \cos \frac{\pi Q}{2L} - |\sin(\omega x_{i+1})| - \\
& \left. \frac{\omega(\xi_{i+1}^+ - \xi_{i+1}^-) \text{sign}(\sin(\omega x^2))}{2L} \cos(\omega x_{i+1}) + 2 \left| \sin \left(\omega \left(\frac{\tilde{x}_{i,n_i} + x_{i+1}}{2} + \frac{\xi_{i+1}^+ - \xi_{i+1}^-}{2} \right) \right) \right| \times \right. \\
& \cos \left(\omega \frac{2|\Delta \tilde{f}_{i,n_i}| + (\xi_{i+1}^+ + \xi_{i+1}^-) \text{sign}(\Delta \tilde{f}_i)}{2L} \right) + \frac{1}{\omega^2} \sum_{i=1}^{N-1} \left(L \text{sign}(\Delta \tilde{f}_i) \sum_{k=0}^{n_i} \left(\sin(\omega \tilde{x}_{i,k}) - \right. \right. \\
& \left. \sin(\omega \tilde{x}_{i,k}) \right) - \frac{\Delta \hat{f}_i}{\Delta x_i} (\sin(\omega x_{i+1} - \sin(\omega x_i)) \Big) + \frac{1}{\omega} \left((\hat{f}_N - \tilde{f}_N) \cos(\omega x_N) - \right. \\
& \left. (\hat{f}_1 - \tilde{f}_1) \cos(\omega x_1) \right). \tag{4.16}
\end{aligned}$$

Using the explicit form of the minorant (see (4.12)–(4.14)) we transform the second component under the maximum sign in (4.15) as follows

$$\begin{aligned}
& \sum_{i=1}^{N-1} \sum_{k=0}^{n_i} \int_{x_{i,k}}^{x_{i,k+1}} (S(x, L) - f_{i,k}^-(x)) \sin(\omega x) dx = \sum_{i=1}^{N-1} \sum_{k=0}^{n_i} \left(\int_{x_{i,k}}^{\tilde{x}_{i,k}} \left(\hat{f}_i - \tilde{f}_i - Q(\tilde{x}_{i,k} - x_i) + \right. \right. \\
& \left. \left. \frac{\Delta \hat{f}_i}{\Delta x_i} (x - x_i) - \frac{\xi_i^+ + \xi_i^-}{2} (1 - \text{sign}(x_{i,k} - x_i)) + L(x - x_{i,k}) \text{sign}(\sin(\omega x)) \right) \sin(\omega x) dx + \right.
\end{aligned}$$

$$\begin{aligned}
& \frac{1}{2} \int_{\tilde{x}_{i,k}}^{\hat{x}_{i,k}} \left(\left(\hat{f}_i - \tilde{f}_i - Q(\tilde{x}_{i,k} - x_i) + \frac{\Delta \hat{f}_i}{\Delta x_i} (x - x_i) - \frac{\xi_i^+ + \xi_i^-}{2} (1 - \text{sign}(x_{i,k} - x_i)) - \right. \right. \\
& L(x - x_{i,k}) \text{sign}(\sin(\omega x)) \Big) (1 - \text{sign}(\Delta \tilde{f}_i) \text{sign}(\sin(\omega x))) + \left(\tilde{f}_i - \tilde{f}_i - Q(\tilde{x}_{i,k+1} - x_i) + \right. \\
& \left. \left. \frac{\Delta \tilde{f}_i}{\Delta x_i} (x - x_i) - \frac{\xi_{i+1}^+ + \xi_{i+1}^-}{2} (1 - \text{sign}(x_{i+1} - x_{i,k+1})) - L(x_{i,k} - x) \text{sign}(\sin(\omega x)) \right) \times \right. \\
& \left. (1 + \text{sign}(\Delta \tilde{f}_i) \text{sign}(\sin(\omega x))) \right) \sin(\omega x) dx + \int_{\tilde{x}_{i,k}}^{x_{i,k+1}} \left(\hat{f}_i - \tilde{f}_i - Q(\tilde{x}_{i,k+1} - x_i) + \right. \\
& \left. \left. \frac{\Delta \hat{f}_i}{\Delta x_i} (x - x_i) - \frac{\xi_{i+1}^+ + \xi_{i+1}^-}{2} (1 - \text{sign}(x_{i+1} - x_{i,k+1})) \right) \sin(\omega x) dx \right) = \\
& \frac{L}{\omega^2} \sum_{i=1}^{N-1} \left(2 \left| \sin \left(\omega \left(\frac{x_i + \tilde{x}_{i,1}}{2} + \frac{\xi_i^- - \xi_i^+}{2L} \right) \right) \right| \cos \left(\omega \frac{2|\Delta \tilde{f}_{i,0}| - |\xi_i^+ + \xi_i^-| \text{sign}(\Delta \tilde{f}_i)}{2L} \right) - \right. \\
& \left| \sin(\omega x_i) \right| + \frac{\omega (\xi_i^+ - \xi_i^-) \text{sign}(\sin(\omega x^1))}{2L} \cos(\omega x_i) + 2 \left(\left| \sin \left(\omega \frac{\tilde{x}_{i,1} + \tilde{x}_{i,2}}{2} \right) \right| + \right. \\
& \left. (n_i - 3) \left| \sin \left(\omega \frac{\tilde{x}_{i,2} + \tilde{x}_{i,3}}{2} \right) \right| + \left| \sin \left(\omega \frac{\tilde{x}_{i,n_i-1} + \tilde{x}_{i,n_i}}{2} \right) \right| \right) \cos \frac{\pi Q}{2L} - \left| \sin(\omega x_{i+1}) \right| - \\
& \left. \frac{\omega (\xi_{i+1}^+ - \xi_{i+1}^-) \text{sign}(\sin(\omega x^2))}{2L} \cos(\omega x_{i+1}) + 2 \left| \sin \left(\omega \left(\frac{\tilde{x}_{i,n_i} + x_{i+1}}{2} + \frac{\xi_{i+1}^+ - \xi_{i+1}^-}{2} \right) \right) \right| \times \right. \\
& \left. \cos \left(\omega \frac{2|\Delta \tilde{f}_{i,n_i}| + (\xi_{i+1}^+ + \xi_{i+1}^-) \text{sign}(\Delta \tilde{f}_i)}{2L} \right) \right) + \frac{1}{\omega^3} \sum_{i=1}^{N-1} \left(L \text{sign}(\Delta \tilde{f}_i) \sum_{k=0}^{n_i} \left(\sin(\omega \tilde{x}_{i,k}) - \right. \right. \\
& \left. \left. \sin(\omega \hat{x}_{i,k}) \right) - \frac{\Delta \hat{f}_i}{\Delta x_i} (\sin(\omega x_{i+1}) - \sin(\omega x_i)) \right) + \frac{1}{\omega} \left((\hat{f}_1 - \tilde{f}_1) \cos(\omega x_1) - \right. \\
& \left. (\hat{f}_N - \tilde{f}_N) \cos(\omega x_N) \right). \tag{4.17}
\end{aligned}$$

Substituting the final results obtained in (4.16) and (4.17) into (4.15), estimate (4.9) follows immediately.

Now, let us consider quadrature formula (3.3) constructed by the residual method. Analogous to the procedure used for the quadrature formula constructed by the method of quasi-solutions, for $S(x, M)$ we have

$$S(x_i, M) \in [f_i - \epsilon_i, f_i + \epsilon_i], \quad i = 1, \dots, N \text{ and } |S'(x, M)| \leq M, \tag{4.18}$$

where

$$M = \max_{1 \leq i \leq N} \left(0, \max_{j > i} \frac{|\tilde{f}_j - \tilde{f}_i| - \epsilon_j - \epsilon_i}{x_j - x_i} \right) \leq \max_{1 \leq i \leq N} \left(0, \max_{j > i} \left(L - \frac{\epsilon_j + \epsilon_i}{x_j - x_i} \right) \right) =$$

$$\max_{1 \leq i \leq N} \left(0, L - \min_{j > i} \frac{\epsilon_j + \epsilon_i}{x_j - x_i} \right) \leq L. \quad (4.19)$$

From (4.18), (4.19) follows that $S(x, M) \in C_{1,L,N,\epsilon}^1$. Furthermore, analogously to the reasoning conducted for quadrature formula constructed by the method of quasi-solutions, it can be shown that the error of quadrature formula (3.3) constructed by the residual method, has the form (4.9). ■

Remark 4.1 For computing integral (3.2) with quadrature formula (3.4) in class $C_{1,L,N,\epsilon}^1$ we have result analogous to Theorem 4.1:

$$\begin{aligned} v(C_{1,L,N,\epsilon}^1, R_3(\omega, S), f) &\leq \frac{L}{\omega^2} \sum_{i=1}^{N-1} \left(2 \left| \cos \left(\omega \left(\frac{x_i + \tilde{x}_{i,1}}{2} + \frac{\xi_i^- - \xi_i^+}{2} \right) \right) \right| \times \right. \\ &\quad \left. \cos \omega \frac{2|\Delta \tilde{f}_{i,0}| - (\xi_i^+ + \xi_i^-)\text{sign}(\Delta \tilde{f}_i)}{2L} - |\cos(\omega x_i)| + \frac{\omega(\xi_i^+ - \xi_i^-)\text{sign}(\cos(\omega x^1))}{2L} \cos(\omega x_i) + \right. \\ &\quad 2 \left(\left| \cos \left(\omega \frac{\tilde{x}_{i,1} + \tilde{x}_{i,2}}{2} \right) \right| + (n_i - 3) \left| \cos \left(\omega \frac{\tilde{x}_{i,2} + \tilde{x}_{i,3}}{2} \right) \right| + \left| \cos \left(\omega \frac{\tilde{x}_{i,n_i-1} + \tilde{x}_{i,n_i}}{2} \right) \right| \right) \cos \frac{\pi Q}{2L} + \\ &\quad 2 \left| \cos \left(\omega \left(\frac{\tilde{x}_{i,n_i-1} + \tilde{x}_{i,n_i}}{2L} + \frac{\xi_{i+1}^+ - \xi_{i+1}^-}{2L} \right) \right) \right| \times \cos \left(\omega \left(\frac{2|\Delta \tilde{f}_{i,n_i}| + (\xi_{i+1}^+ + \xi_{i+1}^-)\text{sign}(\Delta \tilde{f}_i)}{2L} \right) \right) - \\ &\quad \left. |\cos(\omega x_{i+1})| + \frac{\omega(\xi_{i+1}^+ - \xi_{i+1}^-)\text{sign}(\cos(\omega x^2))}{2L} \sin(\omega x_{i+1}) \right) + \left| \frac{1}{\omega^2} \sum_{i=1}^{N-1} \left(L\text{sign}(\Delta \tilde{f}_i) \times \right. \right. \\ &\quad \left. \sum_{k=0}^{n_i} (\cos(\omega \hat{x}_{i,k}) - \cos(\omega \tilde{x}_{i,k})) - \frac{\Delta \hat{f}_i}{\Delta x_i} (\cos(\omega x_{i+1}) - \cos(\omega x_i)) \right) + \frac{1}{\omega} \left((\hat{f}_N - \tilde{f}_N) \sin(\omega x_N) - \right. \\ &\quad \left. \left. (\hat{f}_1 - \tilde{f}_1) \sin(\omega x_1) \right) \right|. \end{aligned} \quad (4.20)$$

We emphasise that for a specific problem when instead of L and ϵ known only their estimates, the real error of computing integrals (3.1), (3.2) using quadrature formulae (3.3) and (3.4), may be considerably less than corresponding error estimates obtained under approximate *a priori* information.

5 Computing Estimates of Fourier Transforms under Approximate A Priori Information

Let us consider the case when integrand $f(x)$ satisfies the Lipschitz condition, but the Lipschitz constant (denoted by L) is not known. The function $f(x)$ is finite on $[a, b]$, has bounded (by L) first derivative and satisfies the condition:

$$|f(x_i) - \tilde{f}_i| \leq \epsilon_i, \quad (5.1)$$

where x_i are nodes of uniform grid $\Delta = \{a = x_1 < \dots < x_N = b\}$ (i.e. we assume that the accuracy of function definition in N nodes of a uniform grid is given), \tilde{f}_i , ϵ_i , $i = 1, \dots, N-1$

are given real numbers, $N = 2^m + 1$, $m \geq 3$ is integer, $\omega_k = 2\pi k/(b - a)$, $k = 1, \dots, N - 1$. We compute N approximate values $\tilde{I}(\omega_k)$ of estimates of sin- and cos- Fourier transforms

$$I_2(\omega_k) = \int_a^b f(x) \sin \omega_k x dx, \quad I_3(\omega_k) = \int_a^b f(x) \cos \omega_k x dx \quad (5.2)$$

in a large range of frequencies $\omega_k \geq 2\pi$. We also compute an a priori estimate of the total absolute error of the problem solution

$$E = \max_{1 \leq k \leq N-1} |\tilde{I}(\omega_k) - I(\omega_k)|. \quad (5.3)$$

The algorithm based on the residual method consists of the following steps.

Algorithm 5.1.

1. Input initial data $N, a, b, \{x_i\}_{i=1,\dots,N}, \{\tilde{f}_i\}_{i=1,\dots,N}, \{\epsilon_i\}_{i=1,\dots,N}$;
2. Find the values of frequencies $\{\omega_k\}_{k=1,\dots,N-1}$ using the formula

$$\omega_k = 2\pi k/(b - a), \quad k = 1, \dots, N - 1;$$

3. Calculate constant M as follows

$$M = \max_{1 \leq i \leq N} (0, \max_{j > i} (|\tilde{f}_j - \tilde{f}_i| - \epsilon_i - \epsilon_j) / (x_j - x_i));$$

4. Calculate the grid step as $h = (b - a)/(N - 1)$;

5. Calculate $\tilde{f}_i^+, \tilde{f}_i^-$ using formulae

$$\tilde{f}_i^+ = \max_{1 \leq j \leq N} (\tilde{f}_j - M|x_j - x_i|), \quad \tilde{f}_i^- = -\max_{1 \leq j \leq N} (-\tilde{f}_j - M|x_j - x_i|);$$

6. Compute $\hat{f}_i = (\tilde{f}_i^+ + \tilde{f}_i^-)/2$, $i = 1, \dots, N$;

7. Compute $\hat{f}'_i = (\hat{f}_{i+1} - \hat{f}_i)/h$, $i = 1, \dots, N - 1$;

8. Compute values $\{\sin(\omega_k ih)\}_{i=1,\dots,N}$, $\{\cos(\omega_k ih)\}_{i=1,\dots,N}$, $k = 1, \dots, N - 1$;

9. Compute estimate $\bar{R} = \{\bar{R}_k\}_{k=1,\dots,N-1}$ for sin- Fourier transform:

$$\begin{aligned} \bar{R}_k = & \frac{1}{\omega_k^2} (\sin(\omega_k h) \sum_{i=1}^{N-1} \tilde{f}'_i \cos(\omega_k ih) - (1 - \cos(\omega_k h)) \times \\ & \sum_{i=1}^{N-1} \tilde{f}'_i \sin(\omega_k ih)) - \frac{1}{\omega_k} (\hat{f}_N \cos(\omega_k b) - \hat{f}_1 \cos(\omega_k a)), \quad k = 1, \dots, N - 1; \end{aligned}$$

10. Compute estimate $\bar{\bar{R}} = \{\bar{\bar{R}}_k\}_{k=1,\dots,N-1}$ for cos- Fourier transform:

$$\begin{aligned} \bar{\bar{R}}_k = & -\frac{1}{\omega_k^2} (\sin(\omega_k h) \sum_{i=1}^{N-1} \tilde{f}'_i \sin(\omega_k ih) + (1 - \cos(\omega_k h)) \times \\ & \sum_{i=1}^{N-1} \tilde{f}'_i \cos(\omega_k ih)) + \frac{1}{\omega_k} (\hat{f}_N \sin(\omega_k b) - \hat{f}_1 \sin(\omega_k a)), \quad k = 1, \dots, N - 1; \end{aligned}$$

11. Output data $\omega = \{\omega_k\}_{k=1,\dots,N-1}$, $\bar{R} = \{\bar{R}_k\}_{k=1,\dots,N-1}$, $\bar{\bar{R}} = \{\bar{\bar{R}}_k\}_{k=1,\dots,N-1}$.

One can also compute a priori estimates \bar{E} , $\bar{\bar{E}}$ of the total absolute errors of computing sin- and cos- Fourier transform respectively

$$\bar{E} = \Delta_1 + \bar{\Delta}_2 + \bar{\Delta}_3, \quad \bar{\bar{E}} = \Delta_1 + \bar{\bar{\Delta}}_2 + \bar{\bar{\Delta}}_3, \quad (5.4)$$

where Δ_1 is the hereditary error, $\bar{\Delta}_2$, $\bar{\bar{\Delta}}_2$ are errors of the residual method, and $\bar{\Delta}_3$, $\bar{\bar{\Delta}}_3$ are round-off errors of computing estimates of sin- and cos- Fourier transforms respectively. In the case when $\epsilon_i = 0$, $i = 1, \dots, N$ the hereditary error is zero. If ϵ_i , $i = 1, \dots, N$ are different from zero, then for the absolute hereditary error Δ_1 of computing \bar{R}_k , and $\bar{\bar{R}}_k$, $k = 1, \dots, N-1$ we have

$$\Delta_1 = \tilde{\epsilon}(b - a), \quad (5.5)$$

where $\tilde{\epsilon}$ is the maximal error of the definition $f(x)$ in the nodes x_i , $i = 1, \dots, N$:

$$|f(x_i) - \hat{f}_i| \leq \tilde{\epsilon}, \quad i = 1, \dots, N. \quad (5.6)$$

The error of the method of computing \bar{R}_k , $k = 1, \dots, N-1$ ($\bar{\Delta}_2$) has the form (3.11), and the error of the method of computing $\bar{\bar{R}}_k$, $k = 1, \dots, N-1$ ($\bar{\bar{\Delta}}_2$) has the form (3.22), provided $\epsilon_i = 0$, $i = 1, \dots, N$, $N \geq |\omega|$ and condition C1 is satisfied. In the case when $\epsilon_i = 0$, $i = 1, \dots, N$ and condition C2 is satisfied, the error $\bar{\Delta}_2$ has the form (3.23), and the error $\bar{\bar{\Delta}}_2$ has the form (3.33). Finally, if ϵ_i , $i = 1, \dots, N$ are non-zeros, then for $\bar{\Delta}_2$ and $\bar{\bar{\Delta}}_2$ we have estimates (4.9) and (4.20) respectively.

We assume that computations are conducted in the floating-point regime with the round-off of arithmetical operations on the basis of the standard rule up to τ binary digits in the normalised mantices of numbers. Then the round-off errors of computing \bar{R}_k , $\bar{\bar{R}}_k$, $k = 1, \dots, N-1$ have the form [31, 27]

$$\begin{aligned} \bar{\Delta}_3 &= \frac{M}{\omega_k^2} (8 + 1.06(N-1)) 2^{-\tau} (|\sin(\omega_k h)| \max_{1 \leq i \leq N-1} |\cos(\omega_k i h)| + \\ &|1 - \cos(\omega_k h)| \max_{1 \leq i \leq N-1} |\sin(\omega_k i h)|) + 5 \left| \frac{1}{\omega_k} (\hat{f}_N \cos(\omega_k b) - \hat{f}_1 \cos(\omega_k a)) \right| 2^{-\tau}; \end{aligned} \quad (5.7)$$

$$\begin{aligned} \bar{\bar{\Delta}}_3 &= \frac{M}{\omega_k^2} (8 + 1.06(N-1)) 2^{-\tau} (|\sin(\omega_k h)| \max_{1 \leq i \leq N-1} |\sin(\omega_k i h)| + \\ &|1 - \cos(\omega_k h)| \max_{1 \leq i \leq N-1} |\cos(\omega_k i h)|) + 5 \left| \frac{1}{\omega_k} (\hat{f}_N \sin(\omega_k b) - \hat{f}_1 \sin(\omega_k a)) \right| 2^{-\tau}. \end{aligned} \quad (5.8)$$

Example 1. Let $f(x) = x$, values \bar{f}_i are given with errors ϵ_i at nodes x_i of a uniform grid on $[0, 1]$, $i = 1, \dots, 2^7 + 1$. We take $\epsilon_i = 10^{-2}$ when i is even, and $\epsilon_i = -2 \times 10^{-2}$ when i is odd (this rule is also used in further examples). By RS and RC we denote errors of sin- and cos- Fourier transforms respectively obtained by the application of the proposed method. By ST and CT we denote the values of integrals $I_1(\omega_k)$ and $I_2(\omega_k)$ (see (5.2)) computed analytically.

Frequency	RS	RC	ST	CT
7.0685830	-.86228190	.093400980	-.86228190	.093400980
159.1740000	.003096217	.005389304	.003096217	.005389304
516.0066000	-.001371164	.001373098	-.001371164	.001373098
864.9852000	.000576047	-.001002487	.000576047	-.001002487
4741.7110000	.000105379	-.000182761	.000105379	-.000182761

Table 1: $f(x) = x$ (for the residual method and the method of quasi-solutions).

The results given in Table 1 confirms the fact that quadrature formulae (3.3), (3.4), constructed by the residual method, are exact on linear functions (taking into account round-off error). Below we provide results on three other illustrative examples of the application of quadrature formulae constructed with the residual method.

Example 2. Let $f(x) = \frac{1}{2}x^2$, values of \tilde{f}_i are given with error ϵ_i in nodes x_i of a uniform grid on $[1, 2]$, $i = 1, \dots, 2^8 + 1$. The results of computations are given in Table 2.

Frequency	RS	RC	ST	CT
7.0685830	.076986650	.238778800	.073891510	.217943200
159.1740000	.004804598	-.013605410	.004608938	-.013621320
516.0066000	.000682280	.003166615	.000690012	.003188090
864.9852000	.000924193	.002433502	.000870774	.002502229
4741.7110000	.000158427	.000452690	.000158636	.000456420

Table 2: $f(x) = x^2/2$ (for the residual method).

Example 3. Let $f(x) = \frac{1}{2}x^3$, values of \tilde{f}_i are given with error ϵ_i in nodes x_i of a uniform grid on $[1, 2]$, $i = 1, \dots, 2^7 + 1$. The results are given in Table 3.

Frequency	RS	RC	ST	CT
7.0685830	.139202300	.554518200	.142515500	.484506100
159.1740000	.010960590	-.024687310	.010736880	-.024570550
516.0066000	.000664492	.007070152	.000703684	.007062656
864.9852000	.002015193	.004482241	.002032279	.004502165
4741.7110000	.000369616	.000817584	.000370015	.000821442

Table 3: $f(x) = x^3/2$ (for the residual method).

Example 4. Let $f(x) = \exp(x)$, values of \tilde{f}_i are given with error ϵ_i in nodes x_i of a uniform grid on $[0, 1]$, $i = 1, \dots, 2^7 + 1$. The results of computations are given in Table 4.

The above examples demonstrate that with increasing frequency ω , accuracy of computations using quadrature formulae (3.3), (3.4) increases. For the same frequency, variation of h does not substantially influence accuracy that confirm the theoretical conclusion that accuracy of quadrature formulae (3.3), (3.4) constructed by the residual method is only weakly dependent on mutual location of nodes of the grid and zeros of oscillating factor ($\sin \omega x$ or $\cos \omega x$ respectively).

Frequency	RS	RC	ST	CT
7.0685830	-.089598670	.282172300	-.090178490	.284681300
159.1740000	.014738880	.014555340	.014913670	.014695670
516.0066000	-.001788848	.003717981	-.001779803	.003728426
864.9852000	.002706216	-.002713822	.002724084	-.002724782
4741.7110000	.000494088	-.000494915	.000497213	-.000496694

Table 4: $f(x) = \exp(x)$ (for the residual method).

Now we consider problem (5.2) with condition (5.6) assuming that \tilde{f}_i are given numbers and the Lipschitz constant L is known. We do not assume as *a priori*, the accuracy of the definition of $f(x)$ in N nodes of a uniform grid. Then Algorithm 5.1 has to be modified as follows

Algorithm 5.2.

1. Input $n, a, b, \{x_i\}_{i=1,\dots,N}, \{f_i\}_{i=1,\dots,N}, L$;
2. Compute h and frequencies $\{\omega_k\}$, $k = 1, \dots, N - 1$ as in Algorithm 5.1;
3. Calculate $\tilde{f}_i^+, \tilde{f}_i^-$ using formulae

$$\tilde{f}_i^+ = \max_{1 \leq j \leq N} (\tilde{f}_j - L|x_j - x_i|), \quad \tilde{f}_i^- = -\max_{1 \leq j \leq N} (-\tilde{f}_j - L|x_j - x_i|);$$

4. The rest of this algorithm coincides with the corresponding steps of Algorithm 5.1.

Quadrature formulae (3.3), (3.4) constructed with the method of quasi-solutions for $f(x) = x$ with exactly given f_i in nodes x_i of a uniform grid on $[0, 1]$ ($i = 1, \dots, 2^7 + 1$), $L = 1$ give results identical to those in Table 1. This support the conclusion that quadrature formulae (3.3), (3.4) constructed with the method of quasi-solutions are exact on linear functions (taking into account round-off errors). For $f(x) = \frac{1}{2}x^2$ with given f_i in nodes x_i of the uniform grid on $[1, 2]$ ($i = 1, \dots, 2^8 + 1$), $L = 2$ the results are given in Table 5.

Frequency	RS	RC
7.0685830	.077655620	.238451700
159.1740000	.004720655	-.013705960
516.0066000	.000684278	.003185287
864.9852000	.000870180	.002504955
4741.7110000	.000158477	.000456333

Table 5: $f(x) = x^2/2$ (the method of quasi-solutions).

With the same data as in Examples 3 and 4 (setting $L = 6$ and $L = e$ respectively), quadrature formulae (3.3), (3.4) constructed with the method of quasi-solutions give the results presented in Table 6. the results are presented in Table 6.

6 Conclusions and Future Directions

We derived and tested optimal-by-order, with constant not exceeding 2, quadrature formulae constructed by the residual method and the method of quasi-solutions. Numerical results

Frequency	RS ($x^3/2$)	RC ($x^3/2$)	RS ($\exp(x)$)	RC ($\exp(x)$)
7.0685830	.138503100	.554798100	-.089598670	.282172300
159.1740000	.011005560	-.024583290	.014738880	.014555340
516.0066000	.000670536	.007042121	-.001788848	.003717981
864.9852000	.002016058	.004461390	.002706216	-.002713822
4741.7110000	.000369576	.000813939	.000494088	-.000494915

Table 6: $f(x) = x^3/2$ and $f(x) = \exp(x)$ (the method of quasi-solutions).

support theoretical conclusions of the paper that

- these quadrature formulae are exact for linear functions (taking into account round-off errors);
- when $\epsilon_i = 0, i = 1, \dots, N$ results obtained by the residual method and the method of quasi-solutions coincide (taking into account round-off errors);
- with increasing frequency ω the accuracy of computations by these formulae increases;
- accuracy of the quadrature formulae practically independent on the mutual arrangement of grid nodes and zeros of oscillating factors.

In order to achieve efficiency in computing \bar{R}_k and $\tilde{\bar{R}}_k, k = 1, \dots, N - 1$ we used the Fast-Fourier-Transform (FFT) algorithm for the following expressions

$$\begin{aligned} \bar{S}_k &= \frac{1}{\omega_k^2} (\sin(\omega_k h) \sum_{i=1}^{N-1} \hat{f}'_i \cos(\omega_k i h) - (1 - \cos(\omega_k h)) \sum_{i=1}^{N-1} \hat{f}'_i \sin(\omega_k i h)), \\ \tilde{\bar{S}}_k &= -\frac{1}{\omega_k^2} (\sin(\omega_k h) \sum_{i=1}^{N-1} \hat{f}'_i \sin(\omega_k i h) + (1 - \cos(\omega_k h)) \sum_{i=1}^{N-1} \hat{f}'_i \cos(\omega_k i h)), \end{aligned}$$

$k = 1, \dots, N - 1$. The computation of $\bar{S}_k, \tilde{\bar{S}}, k = 1, \dots, N - 1$ by the standard procedure requires N^2 operations of additions and multiplications. The application of FFT allows us to speed up calculations by the factor 100 for $N = 2^{10}$ due to its requirement for only $N \log_2 N$ arithmetic operations.

The main emphasis in the paper was given to algorithms that guarantee optimal-by-order, rather than optimal-by-accuracy, solution of the problem of computing integrals with fast oscillatory functions in classes $C_{1,L,N}^1$ and $C_{1,L,N,\epsilon}^1$. Such algorithms are especially effective in the case when *a priori* information about the problem is given approximately. This case is typical in majority of applications. However, our algorithms can also be applied in the case when *a priori* information is assumed to be given precisely. In such a case the error of our results will not exceed optimal-by-accuracy results by more than two times.

The results obtained in this paper has been recently used by authors in the development of algorithms for optimal integration of fast oscillatory functions of two variables (recent survey on the topic may be found in [6]) and the solution of the problem of optimal-by-accuracy recovery of such functions (see, for example, [27]). These issues will be discussed elsewhere.

Acknowledgements

The second author was partially supported by Australian Research Council Grant 179406. The authors thank Professors V. Zadiraka and T. Sag for their helpful comments and Peter Dunn for his helpful assistance at the final stage of preparation of this paper.

References

- [1] Ahlberg, J.H., Nilson, E.N. and Walsh, J.L., *The Theory of Splines and Their Applications*, NY, Academic Press, 1967.
- [2] Alaylioglu, A., Numerical Evaluation of Finite Fourier Integrals, *J. Comput. Appl. Math.*, **9**, 1983, 305–313.
- [3] Berezovskii, A.I., Nechiporenko, N.E., Optimal Accuracy Approximation of Functions and Their Derivatives, *Journal of S. Mathematics*, **54**, 1991, 799–812.
- [4] Busarova, T.H., On Optimal Approximate Integration of Fast Oscillatory Functions, *Ukr. Mathematical Journal*, **38**, 1986, 89–93.
- [5] Chen, T.H.C., Numerical Evaluation of Certain Oscillatory Integrals, *ZAMM*, **65**, No. 10, 1985, 518–520.
- [6] Cools, R., Constructing Cubature Formulae: The Science Behind the Art, *Acta Numerica*, Cambridge University Press, 1997, 1–54.
- [7] Davis, P. and Rabinowitz, P., *Methods of Numerical Integration*, Academic Press, 1984.
- [8] Drachman, B. and Ross, J., Approximation of Certain Functions Given by Integrals with Highly Oscillatory Integrands, *IEEE Transactions on Antennas and Propagation*, **42**, No. 9, 1994, 1355–1356.
- [9] Einarson, B., Numerical Calculation of Fourier Integrals with Cubic Splines, *BIT*, **8**, No. 3, 1968, 279–286.
- [10] Filon, L.N., On a Quadrature Formula for Trigonometric Integrals, *Proc. Roy. Soc. Edinburg*, **49**, 1928, 38–47.
- [11] Flinn, E.A., A Modification of Filon's Method of Numerical Integration, *J. Assoc. Comput. Mach.*, **7**, 1960, 181–184.
- [12] Gold, B. and Rader, C. M., *Digital Processing of Signals*, Krieger, 1983.
- [13] Haider, Q. and Liu, L.C., Fourier and Bessel Transformations of Highly Oscillatory Functions, *J. Phys. A: Math. Gen.*, **25**, 1992, 6755–6760.
- [14] Ivanov, V.K., Vasin, V.V. and Tanana, V.P., *Theory of Linear Ill-Posed Problems and its Applications*, Moscow, Nauka, 1978.
- [15] Ivanov, V.K. and Tanana, V.P., On the Well-Posedness of Conditionally Well-Posed Problems in Locally Convex Spaces, *Doklady. Mathematics*, **46**, No.1, 1993, 165.
- [16] Korneichuk, N.P., Ligun, A.A. and Babenko, V.F., *Extremal Properties of Polynomials and Splines*, NY, Nova Science Publishers, 1996.
- [17] Leondes, C.T., editor, *Digital Control and Signal Processing Systems and Techniques*, San-Diego, Calif.: Academic Press, 1996.

- [18] Levin, D., Fast Integration of Rapidly Oscillatory Functions, *J. Comput. Appl. Math.*, **67**, 1996, 95–101.
- [19] Luke, Y.L., On the Computation of Oscillatory Integrals. Part 2, *Proc. Cambridge Philos. Soc.*, **50**, 1964, 269–277.
- [20] Melnik, K.N. and Melnik, R.V.N., On Computational Aspects of Certain Optimal Digital Signal Processing Algorithms, *Proc. of Computational Technique and Applications Conference: CTAC97*, Eds. J. Noye, M. Teubner and A. Gill, World Scientific, 1998, 433–440.
- [21] Morozov, V.A., *Methods of Solving Incorrectly Posed Problems*, Springer-Verlag, 1984.
- [22] Nikolskii, S.M., *Quadrature Formulae*, Delhi, Hindustan Pub. Corp, 1964, International Monographs on Advanced Mathematics and Physics, 29.
- [23] Nurnberger, G., *Approximation by Spline Functions*, Springer-Verlag, 1989.
- [24] Patterson, T.N.L., On High Precision Methods for the Evaluation of Fourier Integrals with Finite and Infinite Limits, *Numer. Math.*, **27**, 1976, 41–52.
- [25] Piessens, R. and Branders, M., An Algorithm for Computation of Bessel Functions Integrals, *J. Comput. Appl. Math.*, **11**, 1984, 119–137.
- [26] Piessens, R. and Branders, M., A Survey on Numerical Methods for Computation of Bessel Functions Integrals, *Rend. Sem. Mat. Univ. Politec. Torino*, 1985, 249–265.
- [27] Sucharev, A., *Minimax Models in the Theory of Numerical Methods*, Kluwer Academic Publishers, 1992.
- [28] Tikhonov, A.N. et al. *Numerical Methods for the Solution of Ill-Posed Problems*, Dordrecht, Kluwer Academic, 1995.
- [29] Tikhonov, A.N., Leonov, A.S. and Yagola, A.G., *Nonlinear Ill-Posed Problems*, London, Chapman & Hall, 1996.
- [30] Traub, J.F. and Wozniakowski, H., *A General Theory of Optimal Algorithms*, New York: Academic Press, 1980.
- [31] Zadiraka, V.K., Abatov, N.T., Optimally Exact Algorithms for Solutions of a Certain Numerical Integration Problem, *Ukr. Mathematical Journal*, **43**, 1991, 43–54.
- [32] Zhensykbaev, A.A., Monosplines of Minimal Norms and Optimal Quadrature Formulae, *UMN*, **36**, Issue 4(220), 1981, 107–159.
- [33] Zhileikin, Ya.M. and Kukarkin, A.B., A Fast Fourier-Bessel Transformation Algorithm, *Computational Mathematics and Mathematical Physics*, **35**, No. 7, 1995, 901.
- [34] Zhileikin, Ya.M. and Kukarkin, A.B., *Approximate Computation of Integrals from Fast Oscillatory Functions*, Moscow University Press, 1987.

USQ



TOOWOOMBA

**MODELLING DYNAMICS OF
PIEZOELECTRIC SOLIDS IN THE
TWO-DIMENSIONAL CASE**

Roderick V Nicholas Melnik
Department of Mathematics & Computing, USQ

K N Melnik
Electronic Data Systems
Faculty of Sciences Working Paper Series
SC-MC-9708

THE UNIVERSITY OF
S O U T H E R N
Q U E E N S L A N D

Working Papers

FACULTY
OF
SCIENCES

**MODELLING DYNAMICS OF
PIEZOELECTRIC SOLIDS IN THE
TWO-DIMENSIONAL CASE**

Roderick V Nicholas Melnik
Department of Mathematics & Computing, USQ

K N Melnik
Electronic Data Systems
Faculty of Sciences Working Paper Series
SC-MC-9708
December 1997

MODELLING DYNAMICS OF PIEZOELECTRIC SOLIDS IN THE TWO-DIMENSIONAL CASE

R. V. N. Melnik *

Department of Mathematics and Computing,
University of Southern Queensland, QLD 4350, Australia

K. N. Melnik
Electronic Data Systems,
60 Waymouth Street, Adelaide, 5000, Australia

Abstract

This article deals with the accuracy issues for a numerical scheme applied to coupled dynamic problems of electroelasticity. The computational results are discussed with examples for thin finite-dimension piezoceramic solids poled radially and circularly.

Key words: electro-mechanical interactions, hollow piezoceramic cylinders, corner points, accuracy estimates.

1 Introduction.

In many applications, advanced structure design with integrated self-monitoring and control capabilities has become increasingly important. Due to the coupling phenomenon between electric and elastic fields, piezoelectric materials are widely used in such design as sensors, actuators, and transducers [17, 5]. As a rapidly developing area of such utilisation we mention the integral incorporation of mechanical actuation and sensing microstructures into electronic chips [1, 4]. Microelectromechanical structures and piezoelectric semiconductors have features that may not be attained by purely electronic means. These new horizons of piezoelectric applications together with their traditional areas of application have stimulated a greater interest in rigorous mathematical approaches for the investigation of coupling phenomenon in piezoelectric materials.

This paper contributes to the subject of mathematical modelling of piezoelectric structures and deals with dynamic rather than steady-state problems for which coupling between electric and elastic fields may be substantial. We investigate a numerical procedure for the solution of such problems

*Corresponding author, E-mail: melnik@usq.edu.au

applied to thin finite-dimensional structures. In particular, we present computational results for thin hollow piezoceramic structures. These structures have become an important element of design in many technical devices and have potential for future applications [7, 10]. They may also be used as a basis for further investigation of piezoelectric effect in bones [8].

The trend towards miniaturisation of piezoelectric sensors and actuators leads to the situation when standard approaches based on thickness averaging (for mechanical components of electroelastic-fields) and the use of Kirchoff-type hypotheses may not be appropriate. Under such circumstances the development of effective numerical methods becomes important in the investigation of statics and dynamics of coupled electroelastic fields [3, 12, 13, 14, 15, 16, 18, 19, 20, 26]. For those applications that require smaller size and improved resolution of devices (for example, in biomedical imaging, nondestructive evaluation etc), thin structures produced from hollow spheres or cylinders may be good candidates to satisfy size and performance requirements [7].

The paper is organized as follows.

- Mechanical and mathematical notation is introduced in Section 2.
- Section 3 and 4 provide the reader with the mathematical model and the numerical scheme for coupled problems of dynamic electroelasticity in the two-dimensional case;
- In Section 5 and 6 we give the rigorous mathematical justification of the numerical method. Accuracy estimates are also derived in these sections.
- Section 7 describes the results of computational experiments conducted for finite-length piezoceramic cylinders poled radially and circularly.
- Conclusions are given in Section 8.

2 Notation

The following notations are used throughout this paper

Mechanical notation:

- u_r and u_z are components of the displacement vector;
- f_i , $i = 1, 2$ are components of the vector of mass forces;
- f_3 is the volume charge function;
- ρ is the density of the piezoceramic material;
- E_r and E_z are vector components of the stress of electric field;
- D_r and D_z are vector components of electric induction;
- c_{kl} tensor components of elastic quantities;
- e_{ij} tensor components of electro-elastic quantities;
- c_{kl} tensor components of electric quantities;

Mathematical notation:

- $\bar{Q}_T = \bar{G} \times \bar{I}$ is the space-time region of interest with $\bar{G} = \{(r, z) : R_0 \leq r \leq R_1, Z_0 \leq z \leq Z_1\}$ (see Figure 1) and $\bar{I} = \{t : 0 \leq t \leq T\}$;
- $\gamma_1 = \{(r, z) : R_0 < r < R_1, z = Z_0\}$, $\gamma_2 = \{(r, z) : R_0 < r < R_1, z = Z_1\}$, $\gamma_3 = \{(r, z) : r = R_0, Z_0 < z < Z_1\}$, $\gamma_4 = \{(r, z) : r = R_1, Z_0 < z < Z_1\}$, are boundaries of the spatial region \bar{G} ;

- $\gamma_{13} = \{r = R_0, z = Z_0\}$, $\gamma_{23} = \{r = R_0, z = Z_1\}$, $\gamma_{24} = \{r = R_1, z = Z_1\}$, $\gamma_{14} = \{r = R_1, z = Z_0\}$ are corner points of \bar{G} ;
- $\bar{\omega}_{h\tau} = \bar{\omega}_h \times \bar{\omega}_\tau$ is the discrete grid that covers the region \bar{Q}_T with $\bar{\omega}_h = \bar{\omega}_{h_1} \times \bar{\omega}_{h_2}$, $\bar{\omega}_{h_1} = \{r_i : r_i = R_0 + ih_1, i = 0, 1, \dots, N, h_1 = (R_1 - R_0)/N\}$, $\bar{\omega}_{h_2} = \{z_j : z_j = Z_0 + jh_2, j = 0, 1, \dots, M, h_2 = (Z_1 - Z_0)/M\}$, and $\bar{\omega}_\tau = \{t_k : t_k = k\tau, \tau = T/L, k = 0, 1, \dots, L\}$;
- $\bar{r} = r - h_1/2$, $\bar{z} = z - h_2/2$ are “flux” nodes where deformations and stresses are defined;
- $\omega_{h_1} = \{r_i = R_0 + ih_1, i = 1, \dots, N-1\}$, $\omega_{h_1}^+ = \{r_i = R_0 + ih_1, i = 1, \dots, N\}$, and $\omega_{h_1}^- = \{r_i = R_0 + ih_1, i = 0, \dots, N-1\}$ are auxiliary grids in the r -direction (analogously we define grids $\omega_{h_2}, \omega_{h_2}^+, \omega_{h_2}^-$, ω_h etc);
- if $y(r, z, t) \in \bar{\omega}_{h\tau}$ is a grid function, then y_r , y_r , y_τ , and y_{rr} denote the first backward, forward, central and the second central difference derivatives in the r -direction respectively (for difference derivatives in the z -direction and temporal difference derivatives we use analogous notation);
- other notation is explained in the text when appropriate.

3 Mathematical Model

In order to describe the propagation process of electroelastic waves in hollow finite piezoceramic cylinders we use a coupled nonstationary system of partial differential equations which includes

- the equation of motion of piezoelastic continuum media in cylindrical coordinates

$$\rho \frac{\partial^2 u_r}{\partial t^2} = \frac{\partial \sigma_r}{\partial r} + \frac{\partial \sigma_{rz}}{\partial z} + \frac{\sigma_r - \sigma_\theta}{r} + f_1, \quad (3.1)$$

$$\rho \frac{\partial^2 u_z}{\partial t^2} = \frac{\partial \sigma_{rz}}{\partial r} + \frac{\partial \sigma_z}{\partial z} + \frac{\sigma_{rz}}{r} + f_2, \quad (3.2)$$

- the Maxwell equation for piezoelectrics (in the acoustic range of frequencies, it is the forced electrostatic equation of dielectrics)

$$\frac{1}{r} \frac{\partial}{\partial r} (r D_r) + \frac{\partial D_z}{\partial z} = f_3, \quad (3.3)$$

- and state equations for piezoceramics with radial preliminary polarisation

$$\sigma_r = c_{33}\epsilon_r + c_{13}(\epsilon_\theta + \epsilon_z) - e_{33}E_r, \quad \sigma_\theta = c_{13}\epsilon_r + c_{11}\epsilon_\theta + c_{12}\epsilon_z - e_{13}E_r, \quad (3.4)$$

$$\sigma_z = c_{13}\epsilon_r + c_{12}\epsilon_\theta + c_{11}\epsilon_z - e_{13}E_r, \quad \sigma_{rz} = c_{44}\epsilon_{rz} - e_{15}E_z, \quad (3.5)$$

$$D_r = e_{33}\epsilon_r + e_{13}(\epsilon_\theta + \epsilon_z) + e_{33}E_r, \quad D_z = 2e_{15}\epsilon_{rz} + \epsilon_{11}E_z \quad (3.6)$$

(the reader may consult, for example, [2, 11] for similar expressions in the case of circular preliminary polarisation).

The relationship between deformations and displacements is of the Cauchy type

$$\epsilon_r = \frac{\partial u_r}{\partial r}, \quad \epsilon_\theta = \frac{u_r}{r}, \quad \epsilon_z = \frac{\partial u_z}{\partial z}, \quad \epsilon_{rz} = \frac{1}{2} \left(\frac{\partial u_r}{\partial z} + \frac{\partial u_z}{\partial r} \right) \quad (3.7)$$

and the function of electrostatic potential is introduced by formulae

$$E_r = -\frac{\partial \varphi}{\partial r}, \quad E_z = -\frac{\partial \varphi}{\partial z} \quad (3.8)$$

The system (3.1)–(3.8) is considered in the space-time region \bar{Q}_T , and supplemented by the following initial conditions

$$u_r(r, z, 0) = u_r^{(0)}(r, z), \quad \frac{\partial u_r(r, z, 0)}{\partial t} = u_r^{(1)}(r, z), \quad (3.9)$$

$$u_z(r, z, 0) = u_z^{(0)}(r, z), \quad \frac{\partial u_z(r, z, 0)}{\partial t} = u_z^{(1)}(r, z). \quad (3.10)$$

The boundary conditions for this problem are defined as follows

- mechanical boundary conditions

$$\sigma_r(R_i, z, t) = p_r^{(i)}(z, t), \quad \sigma_z(r, Z_i, t) = p_z^{(i)}(r, t), \quad (3.11)$$

$$\sigma_{rz}(R_i, z, t) = p_{rz}^{(i)}(z, t), \quad \sigma_{rz}(r, Z_i, t) = p_{rz}^{(i)}(r, t), \quad i = 0, 1; \quad (3.12)$$

- electric boundary conditions

$$\varphi(R_i, z, t) = 0, \quad D_z(r, Z_i, t) = 0, \quad i = 0, 1. \quad (3.13)$$

Finally, we assume non-negativity of the potential energy of deformation

$$\delta_1 \sum_{i=1}^4 \xi_i^2 \leq c_{33} \xi_1^2 + c_{11} (\xi_2^2 + \xi_3^2) + 2c_{13} (\xi_2 \xi_1 + \xi_3 \xi_1) + 2c_{12} \xi_3 \xi_2 + 2c_{44} \xi_4^2, \quad \delta_1 > 0. \quad (3.14)$$

The results on the existence and uniqueness of generalised solutions for the model (3.1)–(3.13) can be found in [19]. We also recall that the total inner energy described by the model (3.1)–(3.13) is the sum $\mathcal{E} = K + W + P$, where

$$K = \frac{\rho}{2} \int \int_{\Omega} r \left\{ \left(\frac{\partial u_r}{\partial t} \right)^2 \left(\frac{\partial u_z}{\partial t} \right)^2 \right\} d\Omega$$

is the kinetic energy of the system,

$$W = \frac{1}{2} \int \int_{\Omega} r \left\{ c_{33} \epsilon_r^2 + c_{11} (\epsilon_\theta^2 + \epsilon_z^2) + 2c_{13} (\epsilon_\theta \epsilon_r + \epsilon_z \epsilon_r) + 2c_{12} \epsilon_x \epsilon_\theta + 2c_{44} \epsilon_{rz}^2 \right\} d\Omega$$

is the energy of elastic deformation, and

$$P = \frac{\epsilon_{33}}{2} \int \int_{\Omega} r E_r^2 d\Omega + \frac{\epsilon_{11}}{2} \int \int_{\Omega} r E_z^2 d\Omega,$$

is the energy of the electric field.

It was shown in [21] that \mathcal{E} satisfies the following energy balance equation

$$\begin{aligned} \frac{d\mathcal{E}}{dt} = & \int \int_{\Omega} r \left[\frac{\partial D_r}{\partial t} E_r + \frac{\partial D_z}{\partial t} E_z \right] d\Omega + \int_{R_0}^{R_1} r \left[\sigma_{rz} \frac{\partial u_r}{\partial t} + \sigma_z \frac{\partial u_z}{\partial t} \right] dr \Big|_{Z_0}^{Z_1} + \\ & \int_{Z_0}^{Z_1} r \left[\sigma_r \frac{\partial u_r}{\partial t} + \sigma_{rz} \frac{\partial u_z}{\partial t} \right] dz \Big|_{R_0}^{R_1} + \int \int_{\Omega} r \left[f_1 \frac{\partial u_r}{\partial t} + f_2 \frac{\partial u_z}{\partial t} \right] d\Omega, \end{aligned} \quad (3.15)$$

Moreover, the functional \mathcal{E} is bounded and the solution of (3.1)–(3.13) possesses the property given by the following theorem [21].

Theorem 3.1 If the condition (3.14) is satisfied, the solution of the problem (3.1)–(3.13) is characterised by the following energy bound

$$\begin{aligned} \mathcal{E}(t_1) \leq M & \left\{ \rho \int \int_{\Omega} r \left[(u_r^{(1)})^2 + (u_z^{(1)})^2 \right] d\Omega + \int \int_{\Omega} r \left[c_{33}\epsilon_r^2 + c_{11}(\epsilon_\theta^2 + \epsilon_z^2) + 2c_{13}(\epsilon_\theta + \right. \right. \\ & + \epsilon_z)\epsilon_r + 2c_{12}\epsilon_z\epsilon_\theta + 2c_{44}\epsilon_{rz}^2 \Big|_{t=0} d\Omega + \int_{R_0}^{R_1} \left[\sum_{i,j=0}^1 \left(|p_{rt}^{(i)}(r, t_j)|^2 + |p_z^{(i)}(r, t_j)|^2 \right) \right] dr + \\ & \int_{Z_0}^{Z_1} \left[\sum_{i,j=0}^1 \left(|p_r^{(i)}(z, t_j)|^2 + |p_{zt}^{(i)}(z, t_j)|^2 \right) \right] dz + \int_0^{t_1} \int_{R_0}^{R_1} \sum_{i=0}^1 \left[\left(\frac{\partial p_{rt}^{(i)}}{\partial t} \right)^2 + \left(\frac{\partial p_z^{(i)}}{\partial t} \right)^2 \right] dr dt + \\ & \left. \int_0^{t_1} \int_{Z_0}^{Z_1} \sum_{i=0}^1 \left[\left(\frac{\partial p_r^{(i)}}{\partial t} \right)^2 + \left(\frac{\partial p_{zt}^{(i)}}{\partial t} \right)^2 \right] dz dt + \int \int_{\Omega} r \lambda^2 \Big|_{t=0} d\Omega + \int_0^{t_1} \int \int_{\Omega} r(f_1^2 + f_2^2) d\Omega dt \right\}, \end{aligned}$$

where $\mathcal{E}(t)$ is the total inner energy of the electro-mechanical system at time t , and λ is defined by the relationships

$$\frac{\partial \lambda}{\partial r} + \frac{\partial \lambda}{\partial z} = f_3, \quad \lambda(R_0, z, t) = \lambda(r, Z_0, t) = 0.$$

4 Numerical Schemes

The numerical scheme for the solution of problem (3.1)–(3.13) was derived from (3.15) using the concept of generalised solutions [21]. The scheme has the following form

$$\begin{cases} \rho y_{tt} = \Lambda_1(y, g, \mu) + F_1, \\ \rho g_{tt} = \Lambda_2(y, g, \mu) + F_2, \\ \Lambda_3(y, g, \mu) = F_3, \end{cases} \quad (4.1)$$

where functions y , g and μ are fully-discrete functions that give approximations to $u_r(r, z, t)$, $u_z(r, z, t)$ and $\varphi(r, z, t)$ respectively. The explicit form of the operators Λ_i and the right hand sides from (4.1) are given in the Appendix.

The approximation of state equations (3.4)–(3.6) has the following form

$$\bar{\sigma}_r = c_{33}\bar{\epsilon}_r + c_{13}(\bar{\epsilon}_\theta + \bar{\epsilon}_z) - e_{33}\bar{E}_r, \quad \bar{\sigma}_\theta = c_{13}\bar{\epsilon}_r + c_{11}\bar{\epsilon}_\theta + c_{12}\bar{\epsilon}_z - e_{13}\bar{E}_r, \quad (4.2)$$

$$\bar{\sigma}_z = c_{13}\bar{\epsilon}_r + c_{12}\bar{\epsilon}_\theta + c_{11}\bar{\epsilon}_z - e_{13}\bar{E}_r, \quad \bar{\sigma}_{rz} = c_{44}\bar{\epsilon}_{rz} - e_{15}\bar{E}_z, \quad (4.3)$$

$$\bar{D}_r = \bar{E}_r + e_{33}\bar{\epsilon}_r + e_{13}(\bar{\epsilon}_\theta + \bar{\epsilon}_z), \quad \bar{D}_z = \epsilon_{11}\bar{E}_z + 2e_{15}\bar{\epsilon}_{rz}, \quad (4.4)$$

where the expressions

$$\bar{E}_r = \frac{1}{2} (\mu_r + \mu_r^{(-1,z)}), \quad \bar{E}_z = \frac{1}{2} (\mu_z + \mu_z^{(-1,r)}), \quad (4.5)$$

$$\bar{\epsilon}_r = \frac{1}{2} (y_r + y_r^{(-1,z)}), \quad \bar{\epsilon}_\theta = \frac{1}{4r} (y + y^{(-1,r)} + y^{(-1,z)} + y^{(-1,-1)}), \quad (4.6)$$

$$\bar{\epsilon}_z = \frac{1}{2} (g_z + g_z^{(-1_r)}) , \quad 2\bar{\epsilon}_{rz} = \frac{1}{2} (y_z + y_z^{(-1_r)} + g_r + g_r^{(-1_z)}) \quad (4.7)$$

give approximations to the Cauchy relations (3.7) and electrostatic potential formulae (3.8).

The approximation of time derivatives in the initial conditions (3.9), (3.10) is performed with the fictitious-time-layer technique [20, 23, 24]. For $t = 0$ we have:

$$y(r, z, 0) = u_r^{(0)}(r, z), \quad g(r, z, 0) = u_z^{(0)}(r, z). \quad (4.8)$$

$$\rho y_t = \rho u_r^{(1)} + \frac{\tau}{2} (F_1 + \Lambda_1(y, g, \mu)), \quad (4.9)$$

$$\rho g_t = \rho u_z^{(1)} + \frac{\tau}{2} (F_2 + \Lambda_2(y, g, \mu)). \quad (4.10)$$

The discrete analogue of the total inner energy of the electromechanical system is introduced as follows [19, 21]

$$\begin{aligned} \tilde{\mathcal{E}}(t) = & \rho \sum_{\omega_h^-} r \hbar_1 \hbar_2 (y_t^2 + g_t^2) + \sum_{\omega_h^+} \bar{r} \hbar_1 \hbar_2 \{ c_{33} \Phi(\bar{\epsilon}_r) + c_{11} (\Phi(\bar{\epsilon}_\theta) + \Phi(\bar{\epsilon}_z)) + \\ & c_{13} [\bar{\epsilon}_r (\bar{\epsilon}_\theta + \bar{\epsilon}_z) + \bar{\epsilon}_r (\bar{\epsilon}_\theta + \bar{\epsilon}_z) - \tau^2 (\bar{\epsilon}_r)_t ((\bar{\epsilon}_\theta)_t + (\bar{\epsilon}_z)_t)] + \\ & c_{12} [\bar{\epsilon}_z \bar{\epsilon}_\theta + \bar{\epsilon}_z \bar{\epsilon}_\theta - \tau^2 (\bar{\epsilon}_z)_t (\bar{\epsilon}_\theta)_t] + 2c_{44} \Phi(\bar{\epsilon}_{rz}) + c_{33} \Phi(\bar{E}_r) + c_{11} \Phi(\bar{E}_z) \}, \end{aligned} \quad (4.11)$$

where

$$\Phi(y) = \frac{(y + \dot{y})^2}{4} - \frac{\tau^2}{4} (y_t)^2.$$

Then the stability conditions for the numerical scheme (4.1)–(4.10) follow from the non-negativity of (4.11) (see details in [22]):

$$\left\{ \begin{array}{l} \frac{\tau^2}{h_1^2} c_1^2 \left[\left(1 + \frac{h_1}{2R_0}\right) \frac{1 + K_1/\epsilon^M}{1 + K_1} + \frac{c_{13}}{8c_{33}(1 + K_1)} \frac{h_1}{h_2} + \frac{c_{13}h_1}{4R_0c_{33}(1 + K_1)} + \right. \\ \left. \frac{1}{4R_0^2c_{33}(1 + K_1)} \left(c_{11} + \frac{c_{13}^2}{\epsilon_{33}\epsilon^M}\right) h_1^2 \right] + \frac{\tau^2}{h_2^2} c_2^2 \left[\left(1 + \frac{h_1}{2R_0}\right) \times \right. \\ \left. \frac{1 + 2K_2/\epsilon^M}{1 + K_2} + \frac{c_{13}}{8c_{44}(1 + K_2)} \frac{h_2}{h_1} + \frac{c_{12}}{8R_0c_{44}(1 + K_2)} h_2 \right] \leq 1 - \epsilon_1, \\ \frac{\tau^2}{h_2^2} c_3^2 \left[\left(1 + \frac{h_1}{2R_0}\right) \frac{1 + K_3/\epsilon^M}{1 + K_3} + \frac{c_{13}}{8c_{11}(1 + K_3)} \frac{h_2}{h_1} + \frac{c_{12}h_2}{8R_0c_{11}(1 + K_1)} \right] + \\ \frac{\tau^2}{h_1^2} c_2^2 \left[\left(1 + \frac{h_1}{2R_0}\right) \frac{1 + 2K_2/\epsilon^M}{1 + K_2} + \frac{c_{13}}{8c_{44}(1 + K_2)} \frac{h_1}{h_2} \right] \leq 1 - \epsilon_2. \end{array} \right. \quad (4.12)$$

where

$$c_1 = \sqrt{\frac{c_{33}(1 + K_1)}{\rho}}, \quad c_2 = \sqrt{\frac{c_{44}(1 + K_2)}{\rho}}, \quad c_3 = \sqrt{\frac{c_{11}(1 + K_3)}{\rho}},$$

are velocities of the three plane waves (quasi-longitudinal and two quasi-transverse) that in the general case propagate in an anisotropic electro-elastic medium (see [2, 6]), and

$$K_1 = \frac{e_{33}^2}{\epsilon_{33}c_{33}}, \quad K_2 = \frac{e_{15}^2}{\epsilon_{11}c_{44}}, \quad K_3 = \frac{e_{13}^2}{\epsilon_{11}c_{11}}$$

are constants of electromechanical coupling, $\epsilon^M = \min\{\frac{1}{2}, \frac{\epsilon_{11}}{\epsilon_{33}}\}$, and $\epsilon_i, i = 1, 2$ are positive constants that do not depend on steps τ, h_1 and h_2 .

Finally we recall (see [22] for details) that the solution of numerical model (4.1)–(4.10) is the subject of the discrete analogue of Theorem 3.1.

Theorem 4.1 *If conditions (4.12) are satisfied then the solution of the discrete model (4.1)–(4.10) satisfies the following estimate*

$$\begin{aligned} \bar{\mathcal{E}}(t_1 + \tau) \leq M & \left\{ \rho \sum_{\tilde{\omega}_h} r \hbar_1 \hbar_2 \left((y_t(0))^2 + (g_t(0))^2 \right) + \sum_{\omega_h^+} \bar{r} \hbar_1 \hbar_2 \left\{ c_{33} \left[(\bar{\epsilon}_r(0))^2 + \frac{\tau^2}{4} ((\bar{\epsilon}_r(0))_t)^2 \right] + \right. \right. \\ & c_{11} \left[(\bar{\epsilon}_\theta(0))^2 + (\bar{\epsilon}_z(0))^2 + \frac{\tau^2}{4} ((\bar{\epsilon}_\theta(0))_t)^2 + ((\bar{\epsilon}_z(0))_t)^2 \right] + c_{13} [\bar{\epsilon}_r(0) (\bar{\epsilon}_\theta(0) + \bar{\epsilon}_z(0)) + \right. \\ & \left. \left. \frac{\tau}{2} (\bar{\epsilon}_r(0) ((\bar{\epsilon}_\theta(0))_t + (\bar{\epsilon}_z(0))_t) + (\bar{\epsilon}_r(0))_t (\bar{\epsilon}_\theta(0) + \bar{\epsilon}_z(0))) \right] + c_{12} \left[\bar{\epsilon}_z(0) \bar{\epsilon}_\theta(0) + \frac{\tau}{2} (\bar{\epsilon}_z(0) \times \right. \right. \\ & (\bar{\epsilon}_\theta(0))_t + (\bar{\epsilon}_z(0))_t \bar{\epsilon}_\theta(0))] + 2c_{44} \left[(\bar{\epsilon}_{rz}(0))^2 + \frac{\tau^2}{4} ((\bar{\epsilon}_{rz}(0))_t)^2 \right] \right\} + \sum_{\tilde{\omega}_{h_1}} r \hbar_1 \times \\ & 0, \max_{\tau, t_1, t_1 + \tau} \left[\sum_{k=0}^1 ((p_z^{(k)})^2 + (p_{rt}^{(k)})^2) \right] + \sum_{\tilde{\omega}_{h_2}} r \hbar_2 0, \max_{\tau, t_1, t_1 + \tau} \left[\sum_{k=0}^1 ((p_r^{(k)})^2 + (p_{rz}^{(k)})^2) \right] + \\ & \sum_{t'=r}^{t_1} \left\{ \sum_{\tilde{\omega}_{h_1}} r \hbar_1 \left[\sum_{k=0}^1 (|(p_z^{(k)})_t|^2 + |(p_{rt}^{(k)})_t|^2) \right] + \sum_{\tilde{\omega}_{h_2}} r \hbar_2 \left[\sum_{k=0}^1 (|(p_r^{(k)})_t|^2 + \right. \right. \\ & \left. \left. |(p_{rz}^{(k)})_t|^2) \right] \right\} + \sum_{\omega_h^+} \bar{r} \hbar_1 \hbar_2 (\bar{\lambda}(0))^2 + \sum_{t'=\tau}^{t_1} \tau \sum_{\omega_h} r \hbar_1 \hbar_2 (f_1^2 + f_2^2) \right\}. \end{aligned} \quad (4.13)$$

with the discrete energy function defined by (4.11) and the function $\bar{\lambda}$ defined by the following relationships

$$\left(\bar{r} \frac{\bar{\lambda} + \bar{\lambda}^{(+1_r)}}{2} \right)_r + \left(\frac{\bar{r} \bar{\lambda} + \bar{r}^{(+1)} \bar{\lambda}^{(+1_r)}}{2} \right)_z = r f_3, \quad (4.14)$$

$$\bar{\lambda}^{(+1_r)} = \bar{D}_r^{(+1_r)} \text{ for } r = R_0, \text{ and } \bar{\lambda}^{(+1_z)} = \bar{D}_z^{(+1_z)} \text{ for } z = Z_0 \quad \forall t \in \tilde{\omega}_\tau. \quad (4.15)$$

5 Convergence of Discrete Approximations and Accuracy of Numerical Schemes

The error of the scheme (4.1)–(4.10),

$$z_1 = y - u_r, \quad z_2 = g - u_z, \quad \zeta = \mu - \varphi, \quad (5.1)$$

can be defined as the solution of the following operator-difference scheme

$$\begin{cases} \rho(z_1)_{tt} = \Lambda_1(z_1, z_2, \zeta) + \psi_1, & t \in \omega_\tau, \\ \rho(z_2)_{tt} = \Lambda_2(z_1, z_2, \zeta) + \psi_2, & t \in \omega_\tau, \\ \Lambda_3(z_1, z_2, \zeta) = \psi_3, & t \in \tilde{\omega}_\tau \end{cases} \quad (5.2)$$

with the initial conditions

$$z_i = 0, \quad \rho(z_i)_t = \psi_i, \quad i = 1, 2 \quad \text{when } t = 0. \quad (5.3)$$

The right hand sides in (5.2), (5.3) are defined as follows

$$\psi_1 = \begin{cases} -\rho(u_r)_R + (F_1 - \Lambda_1(u_r, u_z, \varphi)) & \text{for } t \in \omega_\tau, \\ \rho u_r^{(1)} - \rho(u_r)_t + \frac{\tau}{2} (F_1 - \Lambda_1(u_r, u_z, \varphi)) & \text{for } t = 0, \end{cases} \quad (5.4)$$

$$\psi_2 = \begin{cases} -\rho(u_z)_R + (F_2 - \Lambda_2(u_r, u_z, \varphi)) & \text{for } t \in \omega_\tau, \\ \rho u_z^{(1)} - \rho(u_z)_t + \frac{\tau}{2} (F_2 - \Lambda_2(u_r, u_z, \varphi)) & \text{for } t = 0, \end{cases} \quad (5.5)$$

$$\psi_3 = F_3 - \Lambda_3(u_r, u_z, \varphi) \quad \text{for } t \in \bar{\omega}_\tau. \quad (5.6)$$

As in the one-dimensional case [20], it can be shown that if the solution of (3.1)–(3.13) belongs to the Sobolev class $(W_2^4(Q_T))^2 \times L_2(I, W_2^4(G))$, then the approximation errors ψ_i , $i = 1, 2, 3$ defined by (5.4)–(5.6) can be represented in the form

$$\psi_i = \check{\psi}_i + \delta_1(h_1) \check{\psi}_i + \delta_2(h_2) \check{\psi}_i, \quad i = 1, 2, \quad \psi_3 = \check{\psi}_3 + \delta_3(h_2) \check{\psi}_3, \quad (5.7)$$

for any $t \in \bar{\omega}_\tau$, where

$$\delta_1(h_1) = \begin{cases} 0, & \omega_h \cup \gamma_1 \cup \gamma_2, \\ -\frac{2}{h_1}, & \bar{\gamma}_3, \\ \frac{2}{h_1}, & \bar{\gamma}_4, \end{cases} \quad \delta_2(h_2) = \begin{cases} 0, & \omega_h \cup \gamma_3 \cup \gamma_4, \\ -\frac{2}{h_2}, & \bar{\gamma}_1, \\ \frac{2}{h_2}, & \bar{\gamma}_2, \end{cases}$$

$$\delta_3(h_2) = \begin{cases} 0, & \omega_h \cup \bar{\gamma}_3 \cup \bar{\gamma}_4, \\ -\frac{2}{h_2}, & \gamma_1, \\ \frac{2}{h_2}, & \gamma_2, \end{cases}$$

$$\check{\psi}_i = \mathcal{O}(|h|^2) \quad \forall (r, z) \in \bar{\omega}_h, \quad |h|^2 = h_1^2 + h_2^2, \quad i = 1, 2; \quad \check{\psi}_3 = \mathcal{O}(|h|^2), \quad (5.8)$$

and the functionals $\check{\psi}_i$, $i = 1, 2$, $\check{\psi}_j$, $j = 1, 2, 3$ have the second order of smallness with respect to $|h|$ except at corner points. In the corner points we have

$$\check{\psi}_i = \mathcal{O}(h_2 + h_1^2), \quad i = 1, 2; \quad \check{\psi}_j = \mathcal{O}(h_1 + h_2^2), \quad j = 1, 2, 3. \quad (5.9)$$

Assuming that the stability conditions (4.12) are satisfied from (4.13) (Theorem 4.1) we obtain the following accuracy estimate for our numerical method:

$$\begin{aligned} \check{\mathcal{E}}(t_1 + \tau) \leq M & \left\{ \sum_{\omega_h} r h_1 h_2 ((\check{\psi}_1(r, z, 0))^2 + (\check{\psi}_2(r, z, 0))^2) + \sum_{\omega^+} \bar{r} h_1 h_2 \{ c_{33}(\check{\epsilon}_r(0))^2 + \right. \\ & c_{11} ((\check{\epsilon}_\theta(0))^2 + (\check{\epsilon}_z(0))^2) + c_{13} [\check{\epsilon}_r(0)(\check{\epsilon}_\theta(0) + \check{\epsilon}_z(0))] + c_{12} \check{\epsilon}_z(0) \check{\epsilon}_\theta(0) + 2c_{44} (\check{\epsilon}_{rz}(0))^2 \} + \\ & \sum_{\omega_1} r \tilde{h}_1 \max_{0, \tau, t_1, t_1 + \tau} ((\check{\psi}_1)^2 + (\check{\psi}_2)^2) + \sum_{\omega_2} r \tilde{h}_2 \max_{0, \tau, t_1, t_1 + \tau} ((\check{\psi}_1)^2 + (\check{\psi}_2)^2) + \\ & \sum_{t'=\tau}^{t_1} \tau \left\{ \sum_{\tilde{\omega}_{h_1}} ((\check{\psi}_1)_{t'})^2 + ((\check{\psi}_2)_{t'})^2 \right\} + \sum_{\tilde{\omega}_{h_2}} ((\check{\psi}_1)_{t'})^2 + ((\check{\psi}_2)_{t'})^2 \} + \\ & \left. \sum_{t'=\tau}^{t_1} \tau \sum_{\omega_h} r h_1 h_2 ((\check{\psi}_1)^2 + (\check{\psi}_2)^2) + \sum_{\omega_h^+} \bar{r} h_1 h_2 (\kappa(0))^2 \right\}, \end{aligned} \quad (5.10)$$

where functions $\check{\epsilon}_r$, $\check{\epsilon}_\theta$, $\check{\epsilon}_z$, $\check{\epsilon}_{rz}$ in the right hand side of (5.10) and the function $\check{\mathcal{E}}$ are obtained from the corresponding functions $\bar{\epsilon}_r$, $\bar{\epsilon}_\theta$, $\bar{\epsilon}_z$, $\bar{\epsilon}_{rz}$ and $\bar{\mathcal{E}}$ by the change of y for z_1 , g for z_2 , and φ for ζ . Finally, the function κ in (5.10) is defined by the following relationships

$$r \psi_3 = \left(\bar{r} \frac{\kappa + \kappa^{(+1_r)}}{2} \right)_r + \left(\frac{\bar{r} \kappa + \bar{r}^{(+1_r)} \kappa^{(+1_r)}}{2} \right)_z, \quad (5.11)$$

$$\kappa^{(+1_r)} = \hat{D}_r^{(+1_r)} \quad \text{for } r = R_0; \quad \kappa^{(+1_z)} = \hat{D}_z^{(+1_z)} \quad \text{for } r = Z_0, \quad (5.12)$$

where \hat{D}_r and \hat{D}_z are obtained from \bar{D}_r and \bar{D}_z by the change of y for u_r , g for u_z , and μ for φ .

Hence, taking into account representations (5.7)–(5.9) the following result follows immediately from the estimate (5.10).

Theorem 5.1 *If stability conditions (4.12) are satisfied, then the solution of (4.1)–(4.10) converges in the energy norm to the solution of the differential problem (3.1)–(3.13) from the class $(W_2^4(Q_T))^2 \times L_2(I, W_2^4(G))$ with the speed $\mathcal{O}(h_1^{3/2} + h_2^{3/2} + \tau^2)$. The accuracy estimate (5.10) holds for any $t_1 > 0$.*

6 The Improvement of Approximations at Corner Points

Compared to the one-dimensional case [20], Theorem 5.1 gives us a weaker result due to the loss of half-order in the convergence speed. We note that the decrease in the order of spatial approximations is caused by the approximations of the mechanical boundary conditions in corner points. Similar difficulties arise even in the classical elasticity theory when, for example, on the one side of a rectangular plate we are given stresses, whereas on the adjoint side we are given one component of stress and one component of displacement.

An effective technique for the improvement of approximations at the corner points for mechanical boundary conditions in coupled problems of electroelasticity was proposed in [19]. It can be shown (for example, by using the Taylor's formula with the remainder in the integral form) that for the solution (u_r, u_z, φ) from $(W_2^4(Q_T))^2 \times L_2(I, W_2^4(G))$ we have

$$\psi_1 = \begin{cases} \frac{h_2}{h_1} \frac{\partial \sigma_r}{\partial z} + \frac{h_1}{h_2} \frac{1}{r} \frac{\partial}{\partial r} (r \sigma_{rz}) + \mathcal{O}(|h|^2 + \tau^2) & \text{when } (r, z) \in \gamma_{13} \cup \gamma_{24}, \\ -\frac{h_1}{h_2} \frac{1}{r} \frac{\partial}{\partial r} (r \sigma_{rz}) - \frac{h_2}{h_1} \frac{\partial \sigma_r}{\partial z} + \mathcal{O}(|h|^2 + \tau^2), & \text{when } (r, z) \in \gamma_{23} \cup \gamma_{14}, \end{cases} \quad (6.1)$$

$$\psi_2 = \begin{cases} \frac{h_1}{h_2} \frac{1}{r} \frac{\partial}{\partial r} (r \sigma_z) + \frac{h_2}{h_1} \frac{\partial \sigma_{rz}}{\partial z} + \mathcal{O}(|h|^2 + \tau^2), & \text{when } (r, z) \in \gamma_{13} \cup \gamma_{24}, \\ -\frac{h_2}{h_1} \frac{\partial \sigma_{rz}}{\partial z} - \frac{h_1}{h_2} \frac{1}{r} \frac{\partial}{\partial r} (r \sigma_z) + \mathcal{O}(|h|^2 + \tau^2) & \text{when } (r, z) \in \gamma_{23} \cup \gamma_{14}. \end{cases} \quad (6.2)$$

Then taking into account the conditions (3.11, (3.12) we set

$$F_1^c = \begin{cases} F_1, & \omega_h \cup \gamma_1 \cup \gamma_2 \cup \gamma_3 \cup \gamma_4, \\ f_1 - \frac{2}{h_1} p_r^{(0)} - \frac{2}{h_2} p_{rt}^{(0)} - \frac{h_2}{h_1} \frac{\partial p_r^{(0)}}{\partial z} - \frac{h_1}{h_2} \frac{1}{r} \frac{\partial}{\partial r} (rp_{rt}^{(0)}), & \gamma_{13}, \\ f_1 + \frac{2}{h_2} p_{rt}^{(1)} - \frac{2}{h_1} p_r^{(0)} + \frac{h_1}{h_2} \frac{1}{r} \frac{\partial}{\partial r} (rp_{rt}^{(1)}) + \frac{h_2}{h_1} \frac{\partial p_r^{(0)}}{\partial z}, & \gamma_{23}, \\ f_1 - \frac{2}{h_2} p_{rt}^{(0)} + \frac{2}{h_1} p_r^{(1)} + \frac{h_1}{h_2} \frac{1}{r} \frac{\partial}{\partial r} (rp_{rt}^{(0)}) + \frac{h_2}{h_1} \frac{\partial p_r^{(1)}}{\partial z}, & \gamma_{14}, \\ f_1 + \frac{2}{h_2} p_{rt}^{(1)} + \frac{2}{h_1} p_r^{(1)} - \frac{h_1}{h_2} \frac{1}{r} \frac{\partial}{\partial r} (rp_{rt}^{(1)}) - \frac{h_2}{h_1} \frac{\partial p_r^{(1)}}{\partial z}, & \gamma_{24}, \end{cases} \quad (6.3)$$

$$F_2^c = \begin{cases} F_2, & \omega_h \cup \gamma_1 \cup \gamma_2 \cup \gamma_3 \cup \gamma_4, \\ f_2 - \frac{2}{h_2} p_z^{(0)} - \frac{2}{h_1} p_{zt}^{(0)} - \frac{h_1}{h_2} \frac{1}{r} \frac{\partial}{\partial r} (rp_z^{(0)}) - \frac{h_2}{h_1} \frac{\partial p_z^{(0)}}{\partial z}, & \gamma_{13}, \\ f_2 + \frac{2}{h_2} p_z^{(1)} - \frac{2}{h_1} p_{zt}^{(0)} + \frac{h_1}{h_2} \frac{1}{r} \frac{\partial}{\partial r} (rp_z^{(1)}) + \frac{h_2}{h_1} \frac{\partial p_z^{(0)}}{\partial z}, & \gamma_{23}, \\ f_2 - \frac{2}{h_2} p_z^{(0)} + \frac{2}{h_1} p_{zt}^{(1)} + \frac{h_1}{h_2} \frac{1}{r} \frac{\partial}{\partial r} (rp_z^{(0)}) + \frac{h_2}{h_1} \frac{\partial p_z^{(1)}}{\partial z}, & \gamma_{14}, \\ f_2 + \frac{2}{h_2} p_z^{(1)} + \frac{2}{h_1} p_{zt}^{(1)} - \frac{h_1}{h_2} \frac{1}{r} \frac{\partial}{\partial r} (rp_z^{(1)}) - \frac{h_2}{h_1} \frac{\partial p_z^{(1)}}{\partial z}, & \gamma_{24}, \end{cases} \quad (6.4)$$

We consider a new numerical scheme

$$\begin{cases} \rho y_{tt} = \Lambda_1(y, g, \mu) + F_1^c, \\ \rho g_{tt} = \Lambda_2(y, g, \mu) + F_2^c, \\ \Lambda_3(y, g, \mu) = F_3 \end{cases} \quad (6.5)$$

with F_1^c, F_2^c defined by (6.3), (6.4), and the following initial conditions

$$\begin{cases} \rho y_t = \rho u_r^{(1)} + \frac{\tau}{2} (F_1^c + \Lambda_1(y, g, \mu)), \\ \rho g_t = \rho u_z^{(1)} + \frac{\tau}{2} (F_2^c + \Lambda_2(y, g, \mu)). \end{cases} \quad (6.6)$$

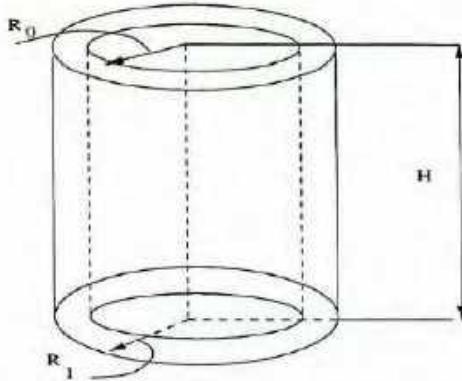


Figure 1: Hollow piezoceramic cylinder ($l = R_1 - R_0$, $l/H < R_1/H = a$)

Approximations of state equations, Cauchy relations and those initial conditions not involving derivatives have the forms (4.2)–(4.8).

The error approximation of the numerical scheme (6.5), (6.6), (4.2)–(4.8) can be represented in the form analogous to (5.7), namely

$$\psi_i^c = \check{\psi}_i^c + \delta_1(h_1) \dot{\psi}_i^c + \delta_2(h_2) \ddot{\psi}_i^c, \quad i = 1, 2. \quad (6.7)$$

However, in the latter case if the solution of (3.1)–(3.13) belongs to the class $(W_2^4(Q_T))^2 \times L_2(I, W_2^4(G))$ all functionals $\check{\psi}_i^c, \dot{\psi}_i^c, \ddot{\psi}_i^c$ have the second order of smallness with respect to $|h|$.

As a result we have proved the following theorem

Theorem 6.1 *The solution of the numerical scheme (6.5), (6.6), (4.2)–(4.8) converges to the solution of the differential problem (3.1)–(3.13) from the class $(W_2^4(Q_T))^2 \times L_2(I, W_2^4(G))$ in the energy norm with the second order accuracy with respect to space-time discretisation subject to the stability conditions (4.12). The accuracy estimate analogous to (5.10) holds for any $t_1 > 0$ with the change of functionals $\check{\psi}_i, \dot{\psi}_i, \ddot{\psi}_i$ ($i = 1, 2$) for $\check{\psi}_i^c, \dot{\psi}_i^c, \ddot{\psi}_i^c$ respectively.*

7 Computational Experiments

In this section we consider results for modelling piezoceramic solids under nonstationary conditions. As in [20], the main emphasis is the dynamics of radial displacements on the external surface of cylinders.

The equations (3.1), (3.2) are scaled to the following form

$$\begin{aligned} \frac{\partial^2 u_r}{\partial t^2} &= \frac{\partial \sigma_r}{\partial r} + a \frac{\partial \sigma_{rz}}{\partial z} + \frac{\sigma_r - \sigma_\theta}{r} + f_1, \\ \frac{\partial^2 u_z}{\partial t^2} &= \frac{\partial \sigma_{rz}}{\partial r} + a \frac{\partial \sigma_z}{\partial z} + \frac{\sigma_{rz}}{r} + f_2, \end{aligned}$$

where a is the parameter that characterises the ratio of the cylinder thickness to its length (see Figure 1).

The Maxwell equation (3.3) is transformed to the form

$$\frac{\epsilon_{33}}{\epsilon_{11}} \frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial \varphi}{\partial r} \right) + \frac{\partial^2 \varphi}{\partial z^2} = F$$

with

$$F = -\frac{f_3}{\epsilon_{11}} + \frac{1}{\epsilon_{11}} \frac{1}{r} \frac{\partial}{\partial r} [r (\epsilon_{33}\epsilon_r + e_{13}(\epsilon_\theta + \epsilon_z))].$$

In all computations we assume that solids are made from PZT-piezoceramic materials for which scaled piezoelectric characteristics are as follows

$$\begin{aligned} c_{11} &= 1, c_{12} = 0.559712, c_{13} = 0.534532, c_{33} = 0.827338, c_{44} = 0.220144, \\ e_{13} &= -0.18605, e_{33} = 0.54027, e_{15} = 0.454383, \epsilon_{11} = 0.87, \epsilon_{33} = 1. \end{aligned}$$

Mechanical boundary conditions follow from the assumption of zero stress on the interior and exterior surfaces. Initially, piezoceramic is assumed to be unexcited. The given potential difference ($2V = 1$) is maintained between interior and exterior surfaces of the cylinders

$$\varphi = \pm 0.5, \quad r = R_0, R_1 \quad \text{and} \quad \frac{\partial \varphi}{\partial z} = \frac{2e_{15}}{\epsilon_{11}} \epsilon_{rz}, \quad z = Z_0, Z_1.$$

Computation has been conducted using the improved scheme (6.5), (6.6), (4.2)–(4.8). For the solution of the Maxwell equation we use the package ALTPACK which implements the alternating-triangular method [23, 24].

Results obtained with the two-dimensional model for cylinders with small values of a (typically for $a = 0.001$) practically coincide with the results obtained earlier for infinite-length cylinders with the one-dimensional model [20]. Therefore, in such situations the use of two-dimensional models is unnecessary and it is reasonable to confine themselves to the one-dimensional model.

For finite-length cylinders the situation is different. The analysis of radial displacements in time on the external surface show that with the decrease in cylinder thickness the amplitude of oscillations increases for both types of cylinders, poled radially and circularly (see Figure 2). However, the increase in amplitude for cylinders poled radially takes place much quicker. For small-thickness cylinders the amplitude of oscillations attains considerable values, many times greater than for cylinders poled circularly. This fact is a consequence of the strong coupling of elastic and electric fields in the case of radial polarisation compared to a weak coupling for cylinders poled circularly. Therefore in general, for modelling finite-length piezoceramic cylinders it is essential to use two-dimensional models for cylinders poled both radially and circularly. We also note that compared to circular-poled cylinders, the error obtained with the one-dimensional model for cylinders poled radially could be considerably greater. Figures 2 and 3 show the evolution of the radial displacements for finite-length cylinders with radial preliminary polarisation. On Figure 4, 5 we present the same characteristics for cylinders poled circularly.

In the design of various technical devices based on hollow piezoceramic cylinders, radiating properties of the external surface may essentially influence the overall device performance. If the radiation from the external surface of thin hollow piezoceramic cylinders has to be increased then such cylinders have to be poled radially. In contrast, for thick hollow piezoceramic cylinders circular preliminary polarisation will lead to an increase in the degree of radiation from the external surface.

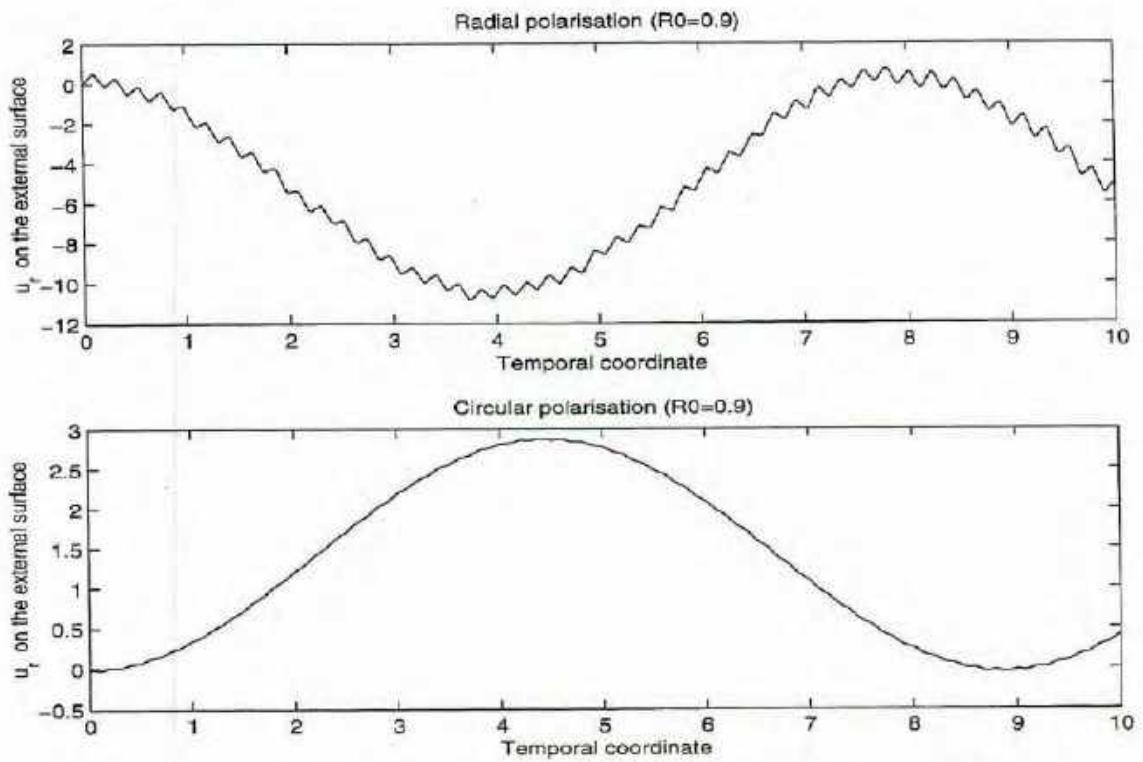


Figure 2: Time dependency of radial displacements on the the external surface of the piezoceramic cylinder ($l = 0.1$, $a = 1$).

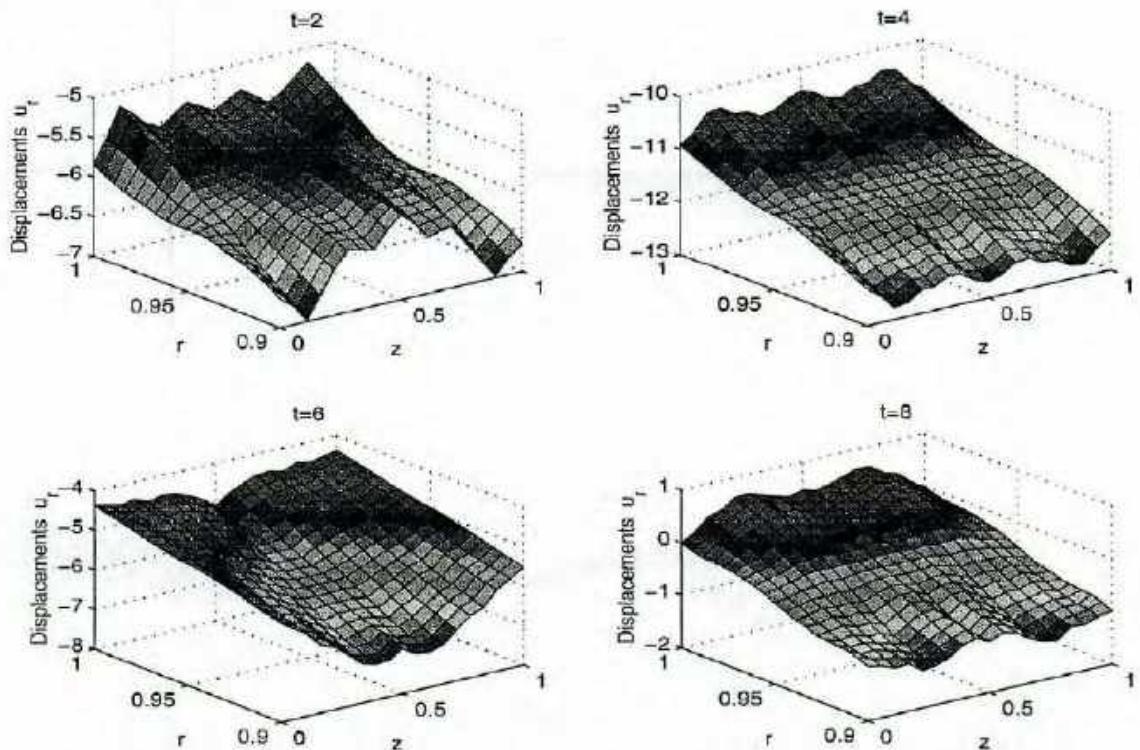


Figure 3: Dynamics of radial displacements in the thin PZT-cylinder poled radially ($l = 0.1$, $a = 1$).

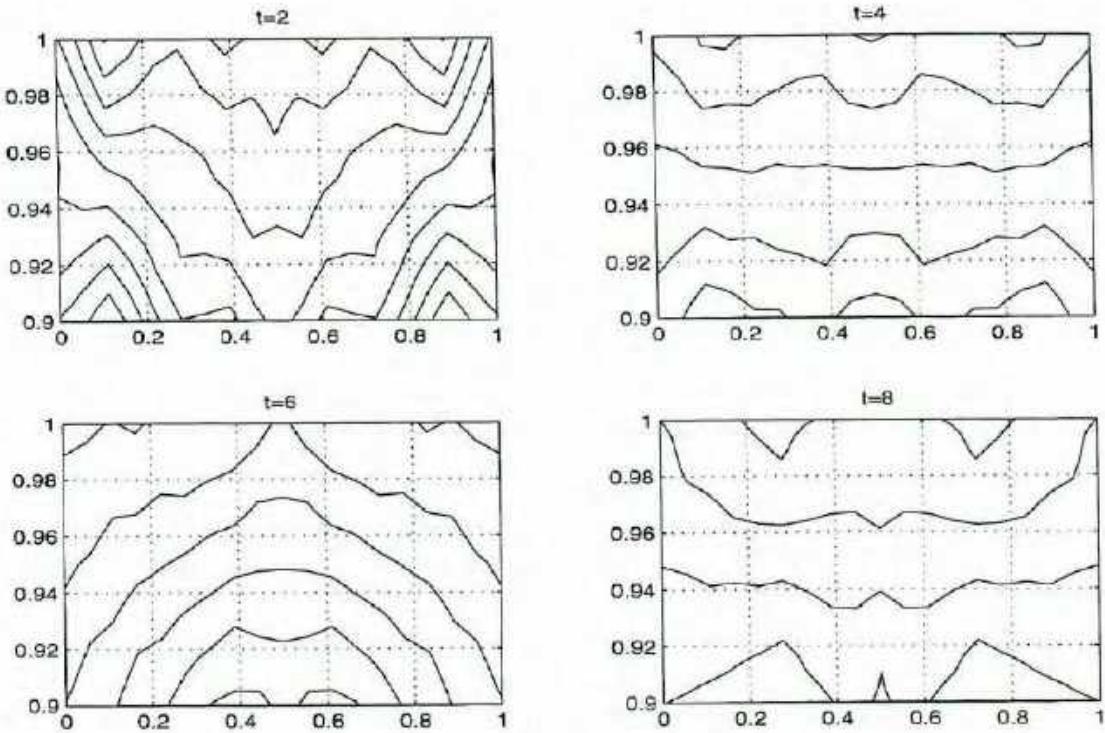


Figure 4: Level curves of radial displacements in the thin piezoceramic cylinder poled radially ($l = 0.1$, $a = 1$).

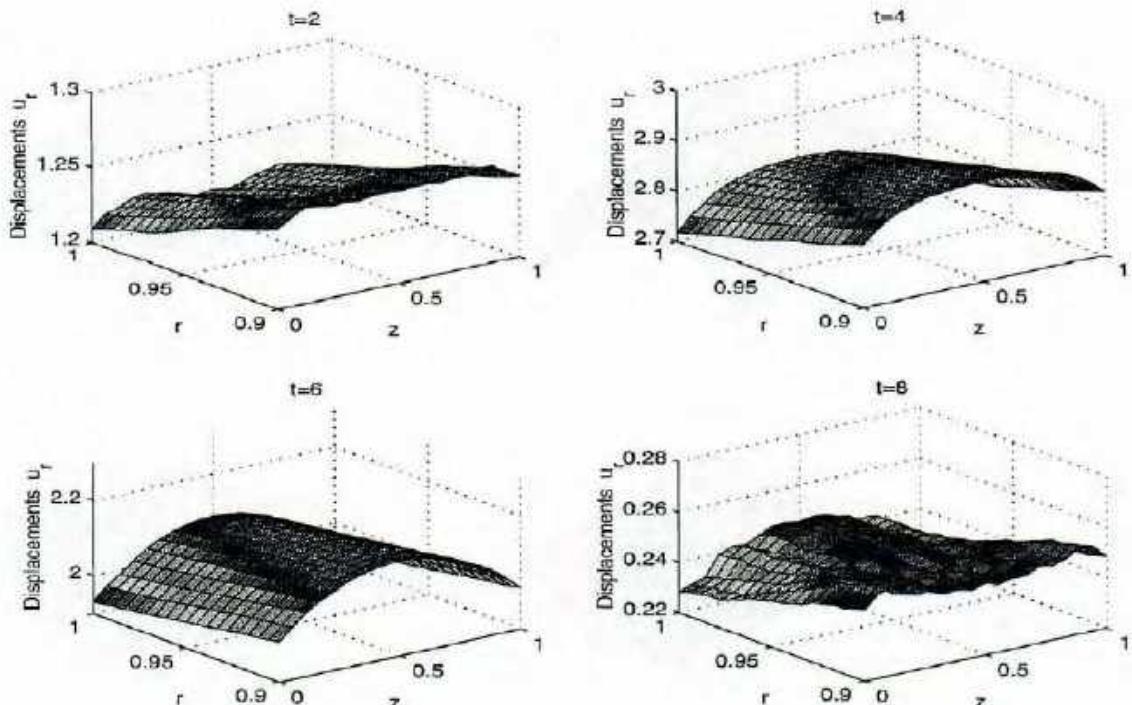


Figure 5: Dynamics of radial displacements in the thin PZT-cylinders poled circularly ($l = 0.1$, $a = 1$).

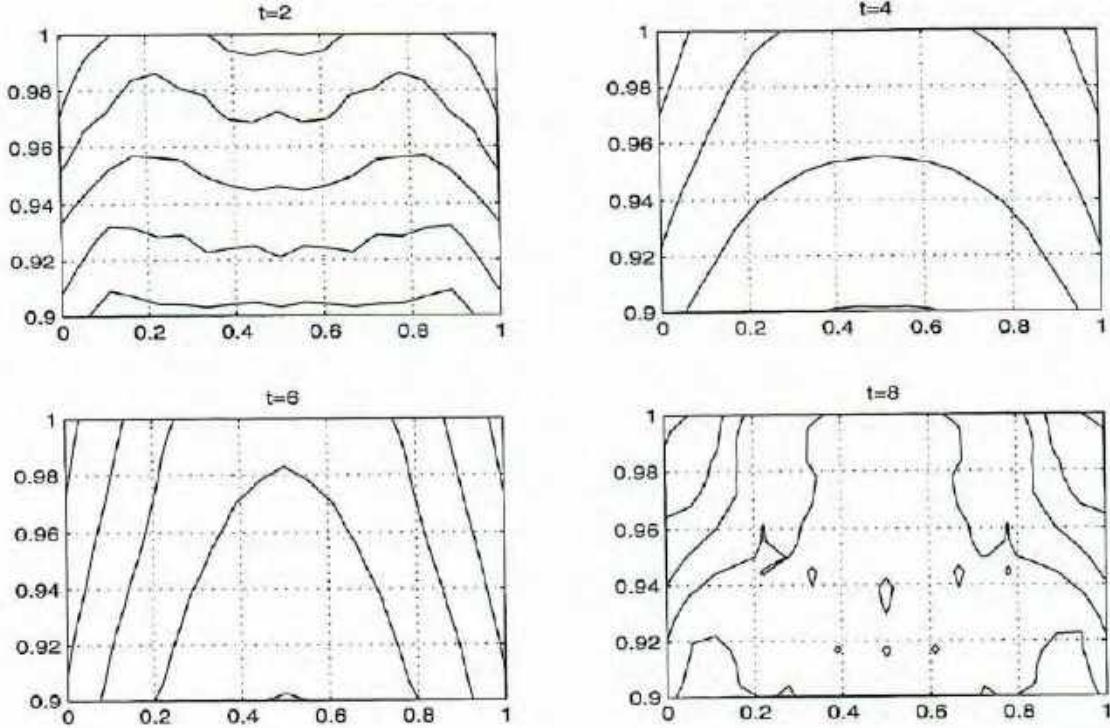


Figure 6: Level curves of radial displacements in the thin piezoceramic cylinder poled circularly ($l = 0.1$, $a = 1$).

8 Conclusions and Future directions.

The question of accuracy is one of the most important issues in the theory and practice of approximate methods. In this paper we proved the accuracy theorems for an effective numerical method designed for the solution of a quite general class of coupled problems in dynamic theory of electroelasticity. Using numerical examples we demonstrated that taking into account the coupling of elastic and electric fields, as well as anisotropy of physical properties of piezoelectric materials may essentially influence the quality of description of wave phenomena in piezoelectric solids.

An important direction in the future development of effective numerical procedures in coupled electroelasticity is connected with a wider application of piezocomposite materials. With piezoelectric material alone it is often difficult to simultaneously satisfy such requirements as broad operation frequency bandwidth, media adjustable acoustic impedance, and a high electromechanical coupling [9]. In such situations the use of piezoceramic-polymer composites may be advantageous. Models in this field should often incorporate such important effects as non-local memory effects, aging, interaction of coupled electroelastic fields with magnetic and thermal fields. This requires taking into account dissipations that may have different origins. The technique based on the Bloch expansion (see [25] and references therein) may prove to be useful in such situations.

Acknowledgements

This work was partially supported by grant USQ-PTRP 17989. The authors are grateful to Dr David Smith for his suggestions of improvement and helpful assistance at the final stage of preparation of this paper.

References

- [1] Ballato, A., Piezoelectricity: Old Effect, New Thrusts, *IEEE Trans. on Ultrasonics, Ferroelectrics, and Frequency Control*, 42, No. 5, 916, 1995.
- [2] Berlincourt, D.A., Curran, D.R., and Jaffe, H. *Piezoelectric and Piezomagnetic Materials and Their Function in Transducers*, in "Physical Acoustics", Vol. 1A, Ed. W.P. Mason, New York and London: Academic Press, 1964, 204-236.
- [3] Buchanan, G.R., Peddeson, Jr., J., Vibration of Infinite Piezoelectric Cylinders with Anisotropic Properties Using Cylindrical Finite Elements, *IEEE Trans. Ultrason., Ferroelect., Freq. Contr.*, Vol. 38, No.3, 1991, 291-301.
- [4] *Ceramic Materials for Electronics: Processing, Properties, and Applications*, Ed. R.C. Buchanan, Marcel Dekker, 1991.
- [5] Crawley, E.F., Intelligent structures for aerospace: a technology overview and assessment, *AIAA Journal*, 32, 1689-1699.
- [6] Dieulesaint, E., Royer, D., *Elastic Waves in Solids: Applications to Signal Processing*, Chichester; N.Y.: J.Wiley, 1980.
- [7] Fielding, J.T. et al, Characterization of PZT Hollow-Sphere Transducers, *Proceedings of the IX IEEE International Symposium on Applications of Ferroelectrics*, 1994, 202-205.
- [8] Fukada, E., Poiseuille Medal Award Lecture: Piezoelectricity of biopolymers, *Biorheology*, 32, 593, 1995.
- [9] Geng, X., Zhang, Q.M., Evaluation of piezocomposites for ultrasonic transducer applications - influence of the unit cell dimensions and the properties of constituents on the performance of 2-2 piezocomposites, *IEEE Trans. Ultrason., Ferroelect., Freq. Contr.*, Vol.44, No. 4, 1997, 857-872.
- [10] Gururaja, T.R., Piezoelectric Transducers for Medical Ultrasonic Imaging, *American Ceramic Society Bulletin*, Vol. 73, No.5, 1994, 50-55.
- [11] Ikeda, T., *Fundamentals of Piezoelectricity*, Oxford: Oxford University Press, 1990.
- [12] Kagawa, Y., Tsuchiya, T., Kawashima, T., Finite Element Simulation of Piezoelectric Vibrator Gyroscopes, *IEEE Trans. Ultrason., Ferroelect., Freq. Contr.*, Vol. 43, No. 4, 1996, 509-120.
- [13] Kagawa, Y., Tsuchiya, T., Furukawa, G., Finite Element Simulation of Dynamic Responses of Piezoelectric Actuators, *J. of Sound and Vibration*, Vol. 191, No. 4, 1996, 519-528.
- [14] Lee, J.S., Boundary Element Method for Electroelastic Interaction in Piezoceramics, *Engineering Analysis with Boundary Elements*, Vol. 15, No. 4, 1995, 321-328.
- [15] Le Letty, R., Claeysen, F., Bossut, R., Combined Finite Element-Normal Mode Expansion Methods in Electroelasticity and Their Application to Piezoactive Motors, *Int. J. Numer. Methods Eng.*, Vol. 40, No. 18, 1997, 3385-3395.

- [16] Lerch, R., Simulation of Piezoelectric Devices by Two- and Three-Dimensional Finite Elements, *IEEE Trans. Ultrason., Ferroelect., Freq. Contr.*, Vol. 37, No. 3, 1990, 233-243.
- [17] Liang, Y.C., Hwu, C., Electromechanical analysis of defects in piezoelectric materials, *Smart Mater. Struct.*, 5, 1996, 314-320.
- [18] Lu, P., Mahrenholtz, O., A Variational Boundary Element Formulation for Piezoelectricity, *Mechanics Research Communications*, Vol. 21, No.6, 1994, 605-615.
- [19] Melnik, R.V.N., Moskalkov, M.N., Difference Schemes for and Analysis of Approximate Solutions of Two-Dimensional Nonstationary Problems in Coupled Electroelasticity, *Differential Equations*, Vol. 27, No. 7, 1991, 1220-1230 (by Plenum Publishing Corporation/Consultants Bureau, N.Y., 1992, 860-867).
- [20] Melnik, R.V.N., The stability condition and energy estimate for non-stationary problems of coupled electroelasticity, *Mathematics and Mechanics of Solids*, Vol. 2, No. 2, 1997 153-180.
- [21] Melnik, R.V.N., Generalised solutions, discrete models, and energy estimates for a 2D problem of coupled field theory, *Centre for Industrial and Applied Mathematics TR 1997-12, School of Mathematics, University of South Australia, 1997, submitted*.
- [22] Melnik, R.V.N., Melnik, K.N., Courant-Friederichs-Lewy type stability conditions in two-dimensional dynamic electroelasticity, *TR Series, Faculty of Sciences, Department of Mathematics & Computing, University of Southern Queensland, SC-MC-9707, 1997, submitted*.
- [23] Samarskii, A.A., *Theorie der Differenzenverfahren*, (Series: Mathematik und ihre Anwendungen in Physik und Technik), Aufl. Leipzig: Academische Verlagsgesellschaft Geest & Portig, 1984.
- [24] Samarskii, A.A., Nikolaev, E.S., *Numerical Methods for Grid Equations*, Basel, Boston: Birkhauser Verlag, 1989.
- [25] Turbe, N., Maugin, G.A., On the linear piezoelectricity of composite materials, *Mathematical Methods in the Applied Sciences*, Vol. 14, 1991, 403-412.
- [26] Vatulyan, A.O., Kublikov, V.L., Boundary Element Method in Electroelasticity, *Boundary Elements Communications*, Vol. 6, No. 2, 1995, 59-61.

Appendix

The difference operators Λ_i and right hand sides F_i , $i = 1, 2, 3$ in (5.2)–(5.5), (6.3)–(6.6) are defined as follows

$$\Lambda_1(y, g, \mu) = \begin{cases} \frac{1}{r} \left(\bar{r} \frac{\bar{\sigma}_r + \bar{\sigma}_r^{(+1_z)}}{2} \right)_r + \frac{1}{r} \left(\bar{r} \bar{\sigma}_{rz} + \bar{r}^{(+1)} \bar{\sigma}_{rz}^{(+1_r)} \right)_z - \\ \frac{\bar{\sigma}_\theta + \bar{\sigma}_\theta^{(+1_r)} + \bar{\sigma}_\theta^{(+1_z)} + \bar{\sigma}_\theta^{(+1,+1)}}{4r}, & (r, z) \in \omega_h, \\ \frac{1}{r} \left(\bar{r} \bar{\sigma}_r^{(+1_z)} \right)_r + \frac{1}{r} \frac{2}{h_2} \left(\frac{\bar{r} \bar{\sigma}_{rz}^{(+1_z)} + \bar{r}^{(+1)} \bar{\sigma}_{rz}^{(+1,+1)}}{2} \right) - \\ \frac{\bar{\sigma}_\theta^{(+1_z)} + \bar{\sigma}_\theta^{(+1,+1)}}{2r}, & (r, z) \in \gamma_1, \\ \frac{1}{r} (\bar{r} \bar{\sigma}_r)_r - \frac{1}{r} \frac{2}{h_2} \left(\frac{\bar{r} \bar{\sigma}_{rz} + \bar{r}^{(+1)} \bar{\sigma}_{rz}^{(+1_r)}}{2} \right) - \frac{\bar{\sigma}_\theta + \bar{\sigma}_\theta^{(+1_r)}}{2r}, & (r, z) \in \gamma_2, \\ \frac{1}{r} \bar{r}^{(+1)} \left(\bar{\sigma}_{rz}^{(+1_r)} \right)_z + \frac{1}{r} \frac{2}{h_1} \bar{r}^{(+1)} \left(\frac{\bar{\sigma}_r^{(+1_r)} + \bar{\sigma}_r^{(+1,+1)}}{2} \right) - \\ \frac{\bar{\sigma}_\theta^{(+1_r)} + \bar{\sigma}_\theta^{(+1,+1)}}{2r}, & (r, z) \in \gamma_3, \\ \frac{1}{r} \bar{r} (\bar{\sigma}_{rz})_z - \frac{1}{r} \frac{2}{h_1} \bar{r} \left(\frac{\bar{\sigma}_r + \bar{\sigma}_r^{(+1_z)}}{2} \right) - \frac{\bar{\sigma}_\theta + \bar{\sigma}_\theta^{(+1_z)}}{2r}, & (r, z) \in \gamma_4, \\ \frac{1}{r} \frac{2}{h_1} \bar{r}^{(+1)} \bar{\sigma}_r^{(+1_r+1)} + \frac{1}{r} \frac{2}{h_2} \bar{r}^{(+1)} \bar{\sigma}_{rz}^{(+1,+1)} - \frac{\bar{\sigma}_\theta^{(+1,+1)}}{r}, & (r, z) \in \gamma_{13}, \\ \frac{1}{r} \frac{2}{h_1} \bar{r}^{(+1)} \bar{\sigma}_r^{(+1_r)} - \frac{1}{r} \frac{2}{h_2} \bar{r}^{(+1)} \bar{\sigma}_{rz}^{(+1_r)} - \frac{\bar{\sigma}_\theta^{(+1_r)}}{r}, & (r, z) \in \gamma_{23}, \\ -\frac{1}{r} \frac{2}{h_1} \bar{r} \bar{\sigma}_r^{(+1_z)} + \frac{1}{r} \frac{2}{h_2} \bar{r} \bar{\sigma}_{rz}^{(+1_z)} - \frac{\bar{\sigma}_\theta^{(+1_z)}}{r}, & (r, z) \in \gamma_{14}, \\ -\frac{1}{r} \frac{2}{h_1} \bar{r} \bar{\sigma}_r - \frac{1}{r} \frac{2}{h_2} \bar{r} \bar{\sigma}_{rz} - \frac{\bar{\sigma}_\theta}{r}, & (r, z) \in \gamma_{24}, \end{cases}$$

$$\Lambda_2(y, g, \mu) = \begin{cases} \frac{1}{r} \left(\bar{r} \frac{\bar{\sigma}_{rz} + \bar{\sigma}_{rz}^{(+1_z)}}{2} \right)_r + \frac{1}{r} \left(\bar{r} \bar{\sigma}_z + \bar{r}^{(+1)} \bar{\sigma}_z^{(+1_r)} \right)_z, & (r, z) \in \omega_h, \\ \frac{1}{r} \left(\bar{r} \bar{\sigma}_{rz}^{(+1_z)} \right)_r + \frac{1}{r} \frac{2}{h_2} \left(\frac{\bar{r} \bar{\sigma}_z^{(+1_z)} + \bar{r}^{(+1)} \bar{\sigma}_z^{(+1, +1)}}{2} \right), & (r, z) \in \gamma_1, \\ \frac{1}{r} (\bar{r} \bar{\sigma}_{rz})_r - \frac{1}{r} \frac{2}{h_2} \left(\frac{\bar{r} \bar{\sigma}_z + \bar{r}^{(+1)} \bar{\sigma}_z^{(+1_r)}}{2} \right) - \frac{\bar{\sigma}_\theta + \bar{\sigma}_\theta^{(+1_r)}}{2r}, & (r, z) \in \gamma_2, \\ \frac{1}{r} \bar{r}^{(+1)} \left(\bar{\sigma}_z^{(+1_r)} \right)_z + \frac{1}{r} \frac{2}{h_1} \bar{r}^{(+1)} \left(\frac{\bar{\sigma}_{rz}^{(+1_r)} + \bar{\sigma}_{rz}^{(+1, +1)}}{2} \right), & (r, z) \in \gamma_3, \\ \frac{1}{r} \bar{r} (\bar{\sigma}_z)_z - \frac{1}{r} \frac{2}{h_1} \bar{r} \left(\frac{\bar{\sigma}_{rz} + \bar{\sigma}_{rz}^{(+1_z)}}{2} \right), & (r, z) \in \gamma_4, \\ \frac{1}{r} \frac{2}{h_2} \bar{r}^{(+1)} \bar{\sigma}_z^{(+1, +1)} + \frac{1}{r} \frac{2}{h_1} \bar{r}^{(+1)} \bar{\sigma}_{rz}^{(+1, +1)}, & (r, z) \in \gamma_{13}, \\ \frac{1}{r} \frac{2}{h_1} \bar{r}^{(+1)} \bar{\sigma}_{rz}^{(+1_r)} - \frac{1}{r} \frac{2}{h_2} \bar{r}^{(+1)} \bar{\sigma}_z^{(+1_r)}, & (r, z) \in \gamma_{23}, \\ -\frac{1}{r} \frac{2}{h_1} \bar{r} \bar{\sigma}_{rz}^{(+1_z)} + \frac{1}{r} \frac{2}{h_2} \bar{r} \bar{\sigma}_z^{(+1_z)}, & (r, z) \in \gamma_{14}, \\ -\frac{1}{r} \frac{2}{h_1} \bar{r} \bar{\sigma}_{rz} - \frac{1}{r} \frac{2}{h_2} \bar{r} \bar{\sigma}_z, & (r, z) \in \gamma_{24}, \end{cases}$$

$$\Lambda_3(y, g, \mu) = \begin{cases} \frac{1}{r} \left(\bar{r} \frac{\bar{D}_r + \bar{D}_r^{(+1_z)}}{2} \right)_r + \frac{1}{r} \left(\bar{r} \bar{D}_z + \bar{r}^{(+1)} \bar{D}_z^{(+1_r)} \right)_z, & (r, z) \in \omega_h, \\ \frac{1}{r} \left(\bar{r} \bar{D}_r^{(+1_z)} \right)_r + \frac{1}{r} \frac{2}{h_2} \left(\frac{\bar{r} \bar{D}_z^{(+1_z)} + \bar{r}^{(+1)} \bar{D}_z^{(+1, +1)}}{2} \right), & (r, z) \in \gamma_1, \\ \frac{1}{r} (\bar{r} \bar{D}_r)_r - \frac{1}{r} \frac{2}{h_2} \left(\frac{\bar{r} \bar{D}_z + \bar{r}^{(+1)} \bar{D}_z^{(+1_r)}}{2} \right), & (r, z) \in \gamma_2, \\ \mu, & (r, z) \in \bar{\omega}_h / (\omega_h \cup \gamma_1 \cup \gamma_2), \end{cases}$$

$$F_3 = \begin{cases} f_3, & (r, z) \in \omega_h \cup \gamma_1 \cup \gamma_2, \\ 0 & (r, z) \in \bar{\omega}_h / (\omega_h \cup \gamma_1 \cup \gamma_2), \end{cases}$$

$$F_1 = f_1 + \begin{cases} 0, & (r, z) \in \omega_h, \\ -\frac{2}{h_2} p_{rt}^{(0)}, & (r, z) \in \gamma_1, \\ \frac{2}{h_2} p_{rt}^{(1)}, & (r, z) \in \gamma_2, \\ -\frac{2}{h_1} p_r^{(0)}, & (r, z) \in \gamma_3, \\ \frac{2}{h_1} p_{rt}^{(1)}, & (r, z) \in \gamma_4, \\ -\frac{2}{h_1} p_r^{(0)} - \frac{2}{h_2} p_{rt}^{(0)}, & (r, z) \in \gamma_{13}, \\ -\frac{2}{h_1} p_r^{(0)} + \frac{2}{h_2} p_{rt}^{(1)}, & (r, z) \in \gamma_{23}, \\ \frac{2}{h_1} p_r^{(1)} - \frac{2}{h_2} p_{rt}^{(0)}, & (r, z) \in \gamma_{14}, \\ \frac{2}{h_1} p_r^{(1)} + \frac{2}{h_2} p_{rt}^{(1)}, & (r, z) \in \gamma_{24}, \end{cases} \quad F_2 = f_2 + \begin{cases} 0, & (r, z) \in \omega_h, \\ -\frac{2}{h_2} p_z^{(0)}, & (r, z) \in \gamma_1, \\ \frac{2}{h_2} p_z^{(1)}, & (r, z) \in \gamma_2, \\ -\frac{2}{h_1} p_{zt}^{(0)}, & (r, z) \in \gamma_3, \\ \frac{2}{h_1} p_{zt}^{(1)}, & (r, z) \in \gamma_4, \\ -\frac{2}{h_1} p_{zt}^{(0)} - \frac{2}{h_2} p_z^{(0)}, & (r, z) \in \gamma_{13}, \\ -\frac{2}{h_1} p_{zt}^{(0)} + \frac{2}{h_2} p_z^{(1)}, & (r, z) \in \gamma_{23}, \\ \frac{2}{h_1} p_{zt}^{(1)} - \frac{2}{h_2} p_z^{(0)}, & (r, z) \in \gamma_{14}, \\ \frac{2}{h_1} p_{zt}^{(1)} + \frac{2}{h_2} p_z^{(1)}, & (r, z) \in \gamma_{24}. \end{cases}$$

USQ



TOOWOOMBA

**COURANT-FRIEDERICH-S-LEWY TYPE
STABILITY CONDITION IN TWO
DIMENSIONAL DYNAMIC
ELECTROELASTICITY**

Roderick V Nicholas Melnik
Department of Mathematics & Computing, USQ

K N Melnik
Electronic Data Systems

THE UNIVERSITY OF
SOUTHERN
QUEENSLAND

Working Papers

FACULTY
OF
SCIENCES

**COURANT-FRIEDERICHSS-LEWY TYPE
STABILITY CONDITION IN TWO
DIMENSIONAL DYNAMIC
ELECTROELASTICITY**

Roderick V Nicholas Melnik
Department of Mathematics & Computing, USQ

K N Melnik
Electronic Data Systems
Faculty of Sciences Working Paper Series
SC-MC-9707
December 1997

COURANT-FRIEDERICH-S-LEWY TYPE STABILITY CONDITION IN TWO DIMENSIONAL DYNAMIC ELECTROELASTICITY

R. V. N. Melnik *

Department of Mathematics and Computing,
University of Southern Queensland, QLD 4350, Australia

K. N. Melnik
Electronic Data Systems,
60 Waymouth Street, Adelaide, 5000, Australia

Abstract

In this article, the authors derive the discrete analogue of the conservation law for coupled electromechanical systems in the two-dimensional case. As a consequence of non-negativity of the discrete energy operator, the stability conditions for the numerical model are derived in the explicit form. Results of computational experiments for hollow piezoceramic cylinders are presented.

Key words: piezoelectricity, discrete conservation laws, Courant-Friederichs-Lewy condition, mixed electroelastic waves.

1 Introduction.

The continuous growth of applications of piezoelectric materials in a variety of fields require more close attention to the development of effective methods for the analysis of vibrations in piezoelectric-based devices and structures. In many cases such analysis has to be conducted for dynamic rather than stationary problems for which the coupling coefficient (a dimensionless measurement of the efficiency in energy conversion) is fairly large. Mathematically, we have to solve a coupled nonstationary problem that is described by a mixed-type system of partial differential equations with appropriate boundary and initial conditions. Except for quite special situations the analytical technique has a limited success in the solution of these types of

*Corresponding author, E-mail: melnik@usq.edu.au

problems. As a result numerical methods become the most natural and effective approach for the solution of these problems [12, 13, 14, 15, 16, 18, 19].

Along with the traditional applications of piezoelectrics, new and emerging technologies based on smart structures and piezoelectric biopolymers have reemphasised the importance of piezoelectrics in the aerospace industry, microprocessor design, aeroelastic control, oceanography, biophysics, medical imaging, consumer electronics and in many other areas of human endeavour [1, 5, 23, 25, 26, 30, 3, 9, 8].

In this paper we contribute to the topic of the development of effective numerical methods for the analysis of coupled electroelastic waves in piezoelectrics. We consider the process of propagation of electroelastic waves in hollow finite-length piezoceramic cylinders with radial/circular preliminary polarisation (see [16, 18, 2, 10] and references therein). Mathematical models for the description of such a process include

- the coupled system of equations of motion and the Maxwell equation for piezoelectrics (in the acoustic range of frequencies, the latter is the forced electrostatic equation of dielectrics) in cylindrical coordinates

$$\begin{cases} \rho \frac{\partial^2 u_r}{\partial t^2} = \frac{\partial \sigma_r}{\partial r} + \frac{\partial \sigma_{rz}}{\partial z} + \frac{\sigma_r - \sigma_\theta}{r} + f_1, \\ \rho \frac{\partial^2 u_z}{\partial t^2} = \frac{\partial \sigma_{rz}}{\partial r} + \frac{\partial \sigma_z}{\partial z} + \frac{\sigma_{rz}}{r} + f_2, \\ \frac{1}{r} \frac{\partial}{\partial r} (r D_r) + \frac{\partial D_z}{\partial z} = f_3, \end{cases} \quad (1.1)$$

- state equations for piezoceramic with radial preliminary polarisation

$$\begin{cases} \sigma_r = c_{33}\epsilon_r + c_{13}(\epsilon_\theta + \epsilon_z) - e_{33}E_r, \quad \sigma_\theta = c_{13}\epsilon_r + c_{11}\epsilon_\theta + c_{12}\epsilon_z - e_{13}E_r, \\ \sigma_z = c_{13}\epsilon_r + c_{12}\epsilon_\theta + c_{11}\epsilon_z - e_{13}E_r, \quad \sigma_{rz} = c_{44}\epsilon_{rz} - e_{15}E_z, \\ D_r = e_{33}\epsilon_r + e_{13}(\epsilon_\theta + \epsilon_z) + \epsilon_{33}E_r, \quad D_z = 2e_{15}\epsilon_{rz} + \epsilon_{11}E_z \end{cases} \quad (1.2)$$

(or state equations for piezoceramic poled circularly),

- initial

$$\begin{cases} u_r(r, z, 0) = u_r^{(0)}(r, z), \quad \frac{\partial u_r(r, z, 0)}{\partial t} = u_r^{(1)}(r, z), \\ u_z(r, z, 0) = u_z^{(0)}(r, z), \quad \frac{\partial u_z(r, z, 0)}{\partial t} = u_z^{(1)}(r, z), \end{cases} \quad (1.3)$$

- and boundary conditions

$$\begin{cases} \sigma_r(R_i, z, t) = p_r^{(i)}(z, t), \quad \sigma_z(r, Z_i, t) = p_z^{(i)}(r, t), \\ \sigma_{rz}(R_i, z, t) = p_{rz}^{(i)}(z, t), \quad \sigma_{rz}(r, Z_i, t) = p_{rz}^{(i)}(r, t), \\ \varphi(R_i, z, t) = 0, \quad D_z(r, Z_i, t) = 0, \quad i = 0, 1. \end{cases} \quad (1.4)$$

In (1.1)–(1.4) we have used the following notation

- u_r and u_z are components of the displacement vector;

- D_r and D_z are vector components of electric induction;
- E_r and E_z are vector components of the electric field stress;
- $\sigma_r, \sigma_\theta, \sigma_z, \sigma_{rz}$ are components of the mechanical field stress;
- $f_i, i = 1, 2$ are components of the vector of mass forces;
- f_3 is the volume charge function;
- ρ is the density of the piezoceramic material;
- c_{kl} denotes tensor components of elastic quantities;
- e_{ij} denotes tensor components of electro-elastic quantities;
- $c_{kl}^{(i)}$ denotes tensor components of electric quantities;
- $u_r^{(i)}, u_z^{(i)}, p_r^{(i)}, p_z^{(i)}, p_{zt}^{(i)}, p_{rt}^{(i)} (i = 0, 1)$ are given functions.

The homogeneity of the electrical conditions in (1.4) does not restrict generality of the model since problems with nonhomogeneous conditions can be reduced to the homogeneous case by the known procedure [16]. The formulation of electrical boundary conditions depends on the character of electric loading and on the location of electrodes on the body surface. We assume that the lateral surfaces of the cylinder are covered by infinitely thin short circuiting electrodes, and that the dielectric permittivity of the surrounding media is much less than the dielectric permittivity of ceramics. This is true for vacuum and air.

As usual the function of electrostatic potential, φ , is introduced by the formulae

$$E_r = -\frac{\partial \varphi}{\partial r}, \quad E_z = -\frac{\partial \varphi}{\partial z} \quad (1.5)$$

and we assume the Cauchy-type relationship between displacements and deformations

$$\epsilon_r = \frac{\partial u_r}{\partial r}, \quad \epsilon_\theta = \frac{u_r}{r}, \quad \epsilon_z = \frac{\partial u_z}{\partial z}, \quad \epsilon_{rz} = \frac{1}{2} \left(\frac{\partial u_r}{\partial z} + \frac{\partial u_z}{\partial r} \right). \quad (1.6)$$

The model (1.1)–(1.6) is considered in the space-time region $\bar{Q}_T = \bar{G} \times \bar{I}$, where

$$\bar{G} = \{(r, z) : R_0 \leq r \leq R_1, Z_0 \leq z \leq Z_1\}, \quad \bar{I} = \{t : 0 \leq t \leq T\}$$

and the condition of non-negativity of the potential energy of deformation,

$$\delta_1 \sum_{i=1}^4 \xi_i^2 \leq c_{33} \xi_1^2 + c_{11} (\xi_2^2 + \xi_3^2) + 2c_{13} (\xi_2 \xi_1 + \xi_3 \xi_1) + 2c_{12} \xi_3 \xi_2 + 2c_{44} \xi_4^2, \quad \delta_1 > 0, \quad (1.7)$$

is assumed.

The strong coupling between electric and elastic fields manifests itself not only through the system (1.1) but also through boundary conditions for stresses. This fact complicates a rigorous mathematical investigation of the model.

The rest of the paper is organised as follows.

- Section 2 provides the reader with the basic notation used throughout the paper;
- In Section 3 we formulate the discrete model derived from the the energy balance equation with the use of the generalised solution technique.
- Section 4 is devoted to the derivation of the discrete conservation law for coupled electromechanical systems in the two-dimensional case. We derive the conditions of system stability from non-negativity requirements imposed on the discrete analogue of the energy functional.
- The results of computational experiments are discussed in Section 5.
- Concluding remarks are given in Section 6.

2 Notation

The following notations are used throughout this paper.

- $\bar{\omega}_{h\tau} = \bar{\omega}_h \times \bar{\omega}_\tau$ is the difference grid that covers the space-time region \bar{Q}_T ;
- $\bar{\omega}_h = \bar{\omega}_{h_1} \times \bar{\omega}_{h_2}$ is the spatial grid with $\bar{\omega}_{h_1} = \{r_i : r_i = R_0 + ih_1, i = 0, 1, \dots, N, h_1 = (R_1 - R_0)/N\}$, $\bar{\omega}_{h_2} = \{z_j : z_j = Z_0 + jh_2, j = 0, 1, \dots, M, h_2 = (Z_1 - Z_0)/M\}$;
- $\bar{\omega}_\tau = \{t_k : t_k = k\tau, \tau = T/L, k = 0, 1, \dots, L\}$ is the temporal grid;
- $\bar{r} = r - h_1/2$, $\bar{z} = z - h_2/2$ are “flux” nodes where values of deformations and stresses will be determined;
- $\omega_{h_1} = \{r_i = R_0 + ih_1, i = 1, \dots, N-1\}$, $\omega_{h_1}^+ = \{r_i = R_0 + ih_1, i = 1, \dots, N\}$, $\omega_{h_1}^- = \{r_i = R_0 + ih_1, i = 0, \dots, N-1\}$ are auxiliary spatial grids in the r -direction (in a similar way we define auxiliary spatial grids in the z -direction $\omega_{h_2}, \omega_{h_2}^+, \omega_{h_2}^-$);
- $\gamma_1 = \{(r, z) : R_0 < r < R_1, z = Z_0\}$, $\gamma_2 = \{(r, z) : R_0 < r < R_1, z = Z_1\}$, $\gamma_3 = \{(r, z) : r = R_0, Z_0 < z < Z_1\}$, $\gamma_4 = \{(r, z) : r = R_1, Z_0 < z < Z_1\}$, are the boundaries of the spatial region G ;
- $\gamma_{13} = \{r = R_0, z = Z_0\}$, $\gamma_{23} = \{r = R_0, z = Z_1\}$, $\gamma_{24} = \{r = R_1, z = Z_1\}$, $\gamma_{14} = \{r = R_1, z = Z_0\}$ are corner points of the region G ;
- $\|y\|_C = \max_{\bar{\omega}_h} |y(x)| \equiv \max_{0 \leq i \leq N} |y_i|$ is a grid analogue of the Chebyshev norm in the functional space C ;
- $\|y\|_0^2 = \sum_{\omega_h^+} y^2 h$ is a grid analogue of the norm in the functional space L_2 (note that the norm $\|y\|_0 = \sum_{\omega_h} y^2 h$ can also be considered as a grid analogue of L_2);
- for any grid function $y(r, z, t) \in \bar{\omega}_{h\tau}$ we use the following notation for difference derivatives
 - $y_r = (y(r, z, t) - y(r - h_1, z, t))/h_1$ is the first backward difference derivative in the direction of r ;
 - $y_r = (y(r + h_1, z, t) - y(r, z, t))/h_1$ is the first forward difference derivative in the direction of r ;
 - $y_{\bar{r}} \equiv y_r = (y(r + h_1, z, t) - y(r - h_1, z, t))/(2h_1)$ is the first central difference derivative in the direction of r ;
 - $y_{\bar{r}r} = (y(r + h_1, z, t) - 2y(r, z, t) + y(r - h_1, z, t))/h_1^2$ is the second (“central”) difference derivative in the direction of r (in an analogous way we denote difference derivatives in the direction of z and temporal difference derivatives);
- $\hat{y} \equiv y(r, z, t + \tau)$ is the discrete function y from the “upper” time layer;
- $\check{y} \equiv y(r, z, t - \tau)$ is the discrete function y from the “lower” time layer;
- other notation used in this paper are standard in the theory of difference scheme [27, 28, 16, 29].

3 Generalised Solutions and Variational Difference Schemes

It is well-known that in practice, solutions of dynamic problems of electroelasticity do not have to be smooth. They might exhibit steep gradients or even discontinuities. In order to

treat these situations numerical methods for the solution of coupled electroelasticity problems were constructed directly from the definition of generalised solutions using the energy balance equation [16, 19, 18].

We recall that the inner energy of the system described by the model (1.1)–(1.6) consists of the three coupled parts, kinetic energy, the energy of elastic deformation, and the energy of electric field, namely

$$\mathcal{E} = \frac{\rho}{2} \int \int_{\Omega} r \left\{ \left(\frac{\partial u_r}{\partial t} \right)^2 \left(\frac{\partial u_z}{\partial t} \right)^2 \right\} d\Omega + \frac{1}{2} \int \int_{\Omega} r \left\{ c_{33} \epsilon_r^2 + c_{11} (\epsilon_\theta^2 + \epsilon_z^2) + 2c_{13} (\epsilon_\theta \epsilon_r + \epsilon_z \epsilon_r) + 2c_{12} \epsilon_z \epsilon_\theta + 2c_{44} \epsilon_{rz}^2 \right\} d\Omega + \frac{\epsilon_{33}}{2} \int \int_{\Omega} r E_r^2 d\Omega + \frac{\epsilon_{11}}{2} \int \int_{\Omega} r E_z^2 d\Omega. \quad (3.1)$$

The functional (3.1) is bounded and an estimate for such a bound was derived in [18].

Theorem 3.1 *If the condition (1.7) is fulfilled, then the solution of the problem (1.1)–(1.6) satisfies the following energy bound*

$$\begin{aligned} \mathcal{E}(t_1) \leq M & \left\{ \rho \int \int_{\Omega} r [(u_r^{(1)})^2 + (u_z^{(1)})^2] d\Omega + \int \int_{\Omega} r [c_{33} \epsilon_r^2 + c_{11} (\epsilon_\theta^2 + \epsilon_z^2) + 2c_{13} (\epsilon_\theta + \right. \\ & \left. + \epsilon_z) \epsilon_r + 2c_{12} \epsilon_z \epsilon_\theta + 2c_{44} \epsilon_{rz}^2] \Big|_{t=0} d\Omega + \int_{R_0}^{R_1} \left[\sum_{i,j=0}^1 (|p_{rt}^{(i)}(r, t_j)|^2 + |p_z^{(i)}(r, t_j)|^2) \right] dr + \right. \\ & \left. \int_{Z_0}^{Z_1} \left[\sum_{i,j=0}^1 (|p_r^{(i)}(z, t_j)|^2 + |p_{zt}^{(i)}(z, t_j)|^2) \right] dz + \int_0^{t_1} \int_{R_0}^{R_1} \sum_{i=0}^1 \left[\left(\frac{\partial p_{ri}^{(i)}}{\partial t} \right)^2 + \left(\frac{\partial p_z^{(i)}}{\partial t} \right)^2 \right] dr dt + \right. \\ & \left. \int_0^{t_1} \int_{Z_0}^{Z_1} \sum_{i=0}^1 \left[\left(\frac{\partial p_r^{(i)}}{\partial t} \right)^2 + \left(\frac{\partial p_{zt}^{(i)}}{\partial t} \right)^2 \right] dz dt + \int \int_{\Omega} r \lambda^2 \Big|_{t=0} d\Omega + \right. \\ & \left. \int_0^{t_1} \int \int_{\Omega} r (f_1^2 + f_2^2) d\Omega dt \right\}, \end{aligned} \quad (3.2)$$

where $\mathcal{E}(t)$ is the total energy of the electro-mechanical system at time t , and λ is defined by following relationships

$$\frac{\partial \lambda}{\partial r} + \frac{\partial \lambda}{\partial z} = f_3, \quad \lambda(R_0, z, t) = \lambda(r, Z_0, t) = 0. \quad (3.3)$$

The difference scheme considered in this article can be formally obtained from the discretised version of the following energy balance equation (see details in [18])

$$\begin{aligned} \frac{d\mathcal{E}}{dt} = \int \int_{\Omega} r & \left[\frac{\partial D_r}{\partial t} E_r + \frac{\partial D_z}{\partial t} E_z \right] d\Omega + \int_{R_0}^{R_1} r \left[\sigma_{rz} \frac{\partial u_r}{\partial t} + \sigma_z \frac{\partial u_z}{\partial t} \right] dr \Big|_{Z_0}^{Z_1} + \\ & \int_{Z_0}^{Z_1} r \left[\sigma_r \frac{\partial u_r}{\partial t} + \sigma_{rz} \frac{\partial u_z}{\partial t} \right] dz \Big|_{R_0}^{R_1} + \int \int_{\Omega} r \left[f_1 \frac{\partial u_r}{\partial t} + f_2 \frac{\partial u_z}{\partial t} \right] d\Omega. \end{aligned} \quad (3.4)$$

The scheme has the form

$$\begin{cases} \rho y_{tt} = \Lambda_1(y, g, \mu) + F_1, \\ \rho g_{tt} = \Lambda_2(y, g, \mu) + F_2, \\ \Lambda_3(y, g, \mu) = F_3, \end{cases} \quad (3.5)$$

where functions y , g and μ are discrete-argument functions that give approximations to the functions $u_r(r, z, t)$, $u_z(r, z, t)$ and $\varphi(r, z, t)$ respectively. The difference operators Λ_i and right hand sides F_i , $i = 1, 2, 3$ in (3.5) are defined as follows

$$\Lambda_1(y, g, \mu) = \begin{cases} \frac{1}{r} \left(\bar{r} \frac{\bar{\sigma}_r + \bar{\sigma}_r^{(+1_z)}}{2} \right)_r + \frac{1}{r} \left(\bar{r} \bar{\sigma}_{rz} + \bar{r}^{(+1)} \bar{\sigma}_{rz}^{(+1_r)} \right)_z - \\ \frac{\bar{\sigma}_\theta + \bar{\sigma}_\theta^{(+1_r)} + \bar{\sigma}_\theta^{(+1_z)} + \bar{\sigma}_\theta^{(+1,+1)}}{4r}, & (r, z) \in \omega_h, \\ \frac{1}{r} \left(\bar{r} \bar{\sigma}_r^{(+1_z)} \right)_r + \frac{1}{r} \frac{2}{h_2} \left(\frac{\bar{r} \bar{\sigma}_{rz}^{(+1_z)} + \bar{r}^{(+1)} \bar{\sigma}_{rz}^{(+1,+1)}}{2} \right) - \\ \frac{\bar{\sigma}_\theta^{(+1_z)} + \bar{\sigma}_\theta^{(+1,+1)}}{2r}, & (r, z) \in \gamma_1, \\ \frac{1}{r} (\bar{r} \bar{\sigma}_r)_r - \frac{1}{r} \frac{2}{h_2} \left(\frac{\bar{r} \bar{\sigma}_{rz} + \bar{r}^{(+1)} \bar{\sigma}_{rz}^{(+1_r)}}{2} \right) - \frac{\bar{\sigma}_\theta + \bar{\sigma}_\theta^{(+1_r)}}{2r}, & (r, z) \in \gamma_2, \\ \frac{1}{r} \bar{r}^{(+1)} \left(\bar{\sigma}_{rz}^{(+1_r)} \right)_z + \frac{1}{r} \frac{2}{h_1} \bar{r}^{(+1)} \left(\frac{\bar{\sigma}_r^{(+1_r)} + \bar{\sigma}_r^{(+1,+1)}}{2} \right) - \\ \frac{\bar{\sigma}_\theta^{(+1_r)} + \bar{\sigma}_\theta^{(+1,+1)}}{2r}, & (r, z) \in \gamma_3, \\ \frac{1}{r} \bar{r} (\bar{\sigma}_{rz})_z - \frac{1}{r} \frac{2}{h_1} \bar{r} \left(\frac{\bar{\sigma}_r + \bar{\sigma}_r^{(+1_z)}}{2} \right) - \frac{\bar{\sigma}_\theta + \bar{\sigma}_\theta^{(+1_z)}}{2r}, & (r, z) \in \gamma_4, \\ \frac{1}{r} \frac{2}{h_1} \bar{r}^{(+1)} \bar{\sigma}_r^{(+1,+1)} + \frac{1}{r} \frac{2}{h_2} \bar{r}^{(+1)} \bar{\sigma}_{rz}^{(+1,+1)} - \frac{\bar{\sigma}_\theta^{(+1,+1)}}{r}, & (r, z) \in \gamma_{13}, \\ \frac{1}{r} \frac{2}{h_1} \bar{r}^{(+1)} \bar{\sigma}_r^{(+1_r)} - \frac{1}{r} \frac{2}{h_2} \bar{r}^{(+1)} \bar{\sigma}_{rz}^{(+1_r)} - \frac{\bar{\sigma}_\theta^{(+1_r)}}{r}, & (r, z) \in \gamma_{23}, \\ -\frac{1}{r} \frac{2}{h_1} \bar{r} \bar{\sigma}_r^{(+1_z)} + \frac{1}{r} \frac{2}{h_2} \bar{r} \bar{\sigma}_{rz}^{(+1_z)} - \frac{\bar{\sigma}_\theta^{(+1_z)}}{r}, & (r, z) \in \gamma_{14}, \\ -\frac{1}{r} \frac{2}{h_1} \bar{r} \bar{\sigma}_r - \frac{1}{r} \frac{2}{h_2} \bar{r} \bar{\sigma}_{rz} - \frac{\bar{\sigma}_\theta}{r}, & (r, z) \in \gamma_{24}, \end{cases}$$

$$\Lambda_2(y, g, \mu) = \begin{cases} \frac{1}{r} \left(\bar{r} \frac{\bar{\sigma}_{rz} + \bar{\sigma}_{rz}^{(+1_z)}}{2} \right)_r + \frac{1}{r} \left(\bar{r} \bar{\sigma}_z + \bar{r}^{(+1)} \bar{\sigma}_z^{(+1_r)} \right)_z, & (r, z) \in \omega_h, \\ \frac{1}{r} \left(\bar{r} \bar{\sigma}_{rz}^{(+1_z)} \right)_r + \frac{1}{r} \frac{2}{h_2} \left(\frac{\bar{r} \bar{\sigma}_z^{(+1_z)} + \bar{r}^{(+1)} \bar{\sigma}_z^{(+1,+1)}}{2} \right), & (r, z) \in \gamma_1, \\ \frac{1}{r} (\bar{r} \bar{\sigma}_{rz})_r - \frac{1}{r} \frac{2}{h_2} \left(\frac{\bar{r} \bar{\sigma}_z + \bar{r}^{(+1)} \bar{\sigma}_z^{(+1_r)}}{2} \right) - \frac{\bar{\sigma}_\theta + \bar{\sigma}_\theta^{(+1_r)}}{2r}, & (r, z) \in \gamma_2, \\ \frac{1}{r} \bar{r}^{(+1)} \left(\bar{\sigma}_z^{(+1_r)} \right)_z + \frac{1}{r} \frac{2}{h_1} \bar{r}^{(+1)} \left(\frac{\bar{\sigma}_{rz}^{(+1_r)} + \bar{\sigma}_{rz}^{(+1,+1)}}{2} \right), & (r, z) \in \gamma_3, \\ \frac{1}{r} \bar{r} (\bar{\sigma}_z)_z - \frac{1}{r} \frac{2}{h_1} \bar{r} \left(\frac{\bar{\sigma}_{rz} + \bar{\sigma}_{rz}^{(+1_z)}}{2} \right), & (r, z) \in \gamma_4, \\ \frac{1}{r} \frac{2}{h_2} \bar{r}^{(+1)} \bar{\sigma}_z^{(+1,+1)} + \frac{1}{r} \frac{2}{h_1} \bar{r}^{(+1)} \bar{\sigma}_{rz}^{(+1,+1)}, & (r, z) \in \gamma_{13}, \\ \frac{1}{r} \frac{2}{h_1} \bar{r}^{(+1)} \bar{\sigma}_{rz}^{(+1_r)} - \frac{1}{r} \frac{2}{h_2} \bar{r}^{(+1)} \bar{\sigma}_z^{(+1_r)}, & (r, z) \in \gamma_{23}, \\ -\frac{1}{r} \frac{2}{h_1} \bar{r} \bar{\sigma}_{rz}^{(+1_z)} + \frac{1}{r} \frac{2}{h_2} \bar{r} \bar{\sigma}_z^{(+1_z)}, & (r, z) \in \gamma_{14}, \\ -\frac{1}{r} \frac{2}{h_1} \bar{r} \bar{\sigma}_{rz} - \frac{1}{r} \frac{2}{h_2} \bar{r} \bar{\sigma}_z, & (r, z) \in \gamma_{24}, \end{cases}$$

$$\Lambda_3(y, g, \mu) = \begin{cases} \frac{1}{r} \left(\bar{r} \frac{\bar{D}_r + \bar{D}_r^{(+1_z)}}{2} \right)_r + \frac{1}{r} \left(\bar{r} \bar{D}_z + \bar{r}^{(+1)} \bar{D}_z^{(+1_r)} \right)_z, & (r, z) \in \omega_h, \\ \frac{1}{r} \left(\bar{r} \bar{D}_r^{(+1_z)} \right)_r + \frac{1}{r} \frac{2}{h_2} \left(\frac{\bar{r} \bar{D}_z^{(+1_z)} + \bar{r}^{(+1)} \bar{D}_z^{(+1,+1)}}{2} \right), & (r, z) \in \gamma_1, \\ \frac{1}{r} (\bar{r} \bar{D}_r)_r - \frac{1}{r} \frac{2}{h_2} \left(\frac{\bar{r} \bar{D}_z + \bar{r}^{(+1)} \bar{D}_z^{(+1_r)}}{2} \right), & (r, z) \in \gamma_2, \\ \mu, & (r, z) \in \bar{\omega}_h / (\omega_h \cup \gamma_1 \cup \gamma_2), \end{cases}$$

$$F_3 = \begin{cases} f_3, & (r, z) \in \omega_h \cup \gamma_1 \cup \gamma_2, \\ 0 & (r, z) \in \bar{\omega}_h / (\omega_h \cup \gamma_1 \cup \gamma_2), \end{cases}$$

$$F_1 = f_1 + \begin{cases} 0, & (r, z) \in \omega_h, \\ -\frac{2}{h_2} p_{rt}^{(0)}, & (r, z) \in \gamma_1, \\ \frac{2}{h_2} p_{rt}^{(1)}, & (r, z) \in \gamma_2, \\ -\frac{2}{h_1} p_r^{(0)}, & (r, z) \in \gamma_3, \\ \frac{2}{h_1} p_{rt}^{(1)}, & (r, z) \in \gamma_4, \\ -\frac{2}{h_1} p_r^{(0)} - \frac{2}{h_2} p_{rt}^{(0)}, & (r, z) \in \gamma_{13}, \\ -\frac{2}{h_1} p_r^{(0)} + \frac{2}{h_2} p_{rt}^{(1)}, & (r, z) \in \gamma_{23}, \\ \frac{2}{h_1} p_r^{(1)} - \frac{2}{h_2} p_{rt}^{(0)}, & (r, z) \in \gamma_{14}, \\ \frac{2}{h_1} p_r^{(1)} + \frac{2}{h_2} p_{rt}^{(1)}, & (r, z) \in \gamma_{24}, \end{cases} \quad F_2 = f_2 + \begin{cases} 0, & (r, z) \in \omega_h, \\ -\frac{2}{h_2} p_z^{(0)}, & (r, z) \in \gamma_1, \\ \frac{2}{h_2} p_z^{(1)}, & (r, z) \in \gamma_2, \\ -\frac{2}{h_1} p_{zt}^{(0)}, & (r, z) \in \gamma_3, \\ \frac{2}{h_1} p_{zt}^{(1)}, & (r, z) \in \gamma_4, \\ -\frac{2}{h_1} p_{zt}^{(0)} - \frac{2}{h_2} p_z^{(0)}, & (r, z) \in \gamma_{13}, \\ -\frac{2}{h_1} p_{zt}^{(0)} + \frac{2}{h_2} p_z^{(1)}, & (r, z) \in \gamma_{23}, \\ \frac{2}{h_1} p_{zt}^{(1)} - \frac{2}{h_2} p_z^{(0)}, & (r, z) \in \gamma_{14}, \\ \frac{2}{h_1} p_{zt}^{(1)} + \frac{2}{h_2} p_z^{(1)}, & (r, z) \in \gamma_{24}. \end{cases}$$

The approximation of the state equations has the following form

$$\begin{cases} \bar{\sigma}_r = c_{33}\bar{\epsilon}_r + c_{13}(\bar{\epsilon}_\theta + \bar{\epsilon}_z) - e_{33}\bar{E}_r, & \bar{\sigma}_\theta = c_{13}\bar{\epsilon}_r + c_{11}\bar{\epsilon}_\theta + c_{12}\bar{\epsilon}_z - e_{13}\bar{E}_r, \\ \bar{\sigma}_z = c_{13}\bar{\epsilon}_r + c_{12}\bar{\epsilon}_\theta + c_{11}\bar{\epsilon}_z - e_{13}\bar{E}_r, & \bar{\sigma}_{rz} = c_{44}\bar{\epsilon}_{rz} - e_{15}\bar{E}_z, \\ \bar{D}_r = \bar{E}_r + e_{33}\bar{\epsilon}_r + e_{13}(\bar{\epsilon}_\theta + \bar{\epsilon}_z), & \bar{D}_z = \epsilon_{11}\bar{E}_z + 2e_{15}\bar{\epsilon}_{rz}, \end{cases} \quad (3.6)$$

where

$$\begin{aligned} \bar{E}_r &= \frac{1}{2} (\mu_r + \mu_r^{(-1_z)}), \quad \bar{E}_z = \frac{1}{2} (\mu_z + \mu_z^{(-1_r)}), \\ \bar{\epsilon}_r &= \frac{1}{2} (y_r + y_r^{(-1_z)}), \quad \bar{\epsilon}_\theta = \frac{1}{4\bar{r}} (y + y^{(-1_r)} + y^{(-1_z)} + y^{(-1,-1)}), \\ \bar{\epsilon}_z &= \frac{1}{2} (g_z + g_z^{(-1_r)}), \quad 2\bar{\epsilon}_{rz} = \frac{1}{2} (y_z + y_z^{(-1_r)} + g_r + g_r^{(-1_z)}). \end{aligned}$$

The approximation of initial conditions is performed as follows

$$y(r, z, 0) = u_r^{(0)}(r, z), \quad g(r, z, 0) = u_z^{(0)}(r, z), \quad (3.7)$$

$$\rho y_t = \rho u_r^{(1)} + \frac{\tau}{2} (F_1 + \Lambda_1(y, g, \mu)), \quad \rho g_t = \rho u_z^{(1)} + \frac{\tau}{2} (F_2 + \Lambda_2(y, g, \mu)). \quad (3.8)$$

The discrete analogue of the energy conservation law (3.4) is fundamental in the investigation of the system stability. One of the key factors in this investigation is establishing some bounds for the energy functional at any given moment of time. Ultimately, it is these bounds that allow us to guarantee the stability of the corresponding difference problem (3.5)–(3.8) under certain conditions on the time and space discretisations.

4 Discrete Conservation Laws in 2D Electroelasticity

The aim of this section is to establish the analogue of (3.4) for the discrete model (3.5)–(3.8).

We start from the following identity

$$\begin{aligned} & \sum_{\tilde{\omega}_h} \rho r \hbar_1 \hbar_2 [(y_t - y_{\tilde{t}})(y_t + y_{\tilde{t}}) + (g_t - g_{\tilde{t}})(g_t + g_{\tilde{t}})] + \\ & 2\tau \left[\sum_{\omega_h^+} \bar{r} h_1 h_2 (\bar{\sigma}_r(\bar{\epsilon}_r)_{\tilde{t}} + \bar{\sigma}_\theta(\bar{\epsilon}_\theta)_{\tilde{t}} + \bar{\sigma}_z(\bar{\epsilon}_z)_{\tilde{t}} + 2\bar{\sigma}_{rz}(\bar{\epsilon}_{rz})_{\tilde{t}}) \right] = \\ & 2\tau \left\{ \sum_{\tilde{\omega}_h} r \hbar_1 \hbar_2 (f_1 v + f_2 w) + \sum_{\tilde{\omega}_{h_1}} r \hbar_1 [\bar{\sigma}_{rz} v + \bar{\sigma}_z w]|_{Z_0}^{Z_1} + \sum_{\tilde{\omega}_{h_2}} r \hbar_2 [\bar{\sigma}_r v + \bar{\sigma}_{rz} w]|_{R_0}^{R_1} \right\}. \end{aligned} \quad (4.1)$$

The identity (4.1) is called the difference energy identity and can be easily obtained using the procedure described in [16] for the one-dimensional case. In (4.1) we have denoted $y_{\tilde{t}}$ by v , $g_{\tilde{t}}$ by w , and used the equality $(\bar{D}_r)_{\tilde{t}} = (\bar{D}_z)_{\tilde{t}} = 0$ (see [16] for details).

Taking into account the state equations (3.6) and the easily verified identities,

$$y = \frac{\hat{y} + 2y + \check{y}}{4} - \frac{\tau^2}{4} y_{\tilde{t}}, \quad 2\tau y_{\tilde{t}} = \tau(y_t + y_{\tilde{t}}) = \hat{y} - \check{y},$$

we can derive from (4.1) the discrete analogue of energy conservation law for the electromechanical system described by the model (1.1)–(1.6). We have

$$\begin{aligned} \bar{\mathcal{E}}(t + \tau) = & \bar{\mathcal{E}}(t) + 2\tau \sum_{\tilde{\omega}_h} r \hbar_1 \hbar_2 (f_1 v + f_2 w) + \\ & 2\tau \left\{ \sum_{\tilde{\omega}_{h_1}} [\bar{\sigma}_{rz} v + \bar{\sigma}_z w]|_{Z_0}^{Z_1} + \sum_{\tilde{\omega}_{h_2}} r \hbar_2 [\bar{\sigma}_r v + \bar{\sigma}_{rz} w]|_{R_0}^{R_1} \right\}, \end{aligned} \quad (4.2)$$

where $\bar{\mathcal{E}}(t)$ is the discrete analogue of the total energy of the electromechanical system defined as follows (first introduced in [15])

$$\begin{aligned} \bar{\mathcal{E}}(t) = & \rho \sum_{\tilde{\omega}_h} r \hbar_1 \hbar_2 (y_{\tilde{t}}^2 + g_{\tilde{t}}^2) + \sum_{\omega_h^+} \bar{r} h_1 h_2 \{ c_{33} \Phi(\bar{\epsilon}_r) + c_{11} (\Phi(\bar{\epsilon}_\theta) + \Phi(\bar{\epsilon}_z)) + \\ & c_{13} [\bar{\epsilon}_r(\bar{\epsilon}_\theta + \bar{\epsilon}_z) + \bar{\epsilon}_r(\bar{\epsilon}_\theta + \bar{\epsilon}_z) - \tau^2(\bar{\epsilon}_r)_{\tilde{t}} ((\bar{\epsilon}_\theta)_{\tilde{t}} + (\bar{\epsilon}_z)_{\tilde{t}})] + \\ & c_{12} [\bar{\epsilon}_z \bar{\epsilon}_\theta + \bar{\epsilon}_z \bar{\epsilon}_\theta - \tau^2(\bar{\epsilon}_z)_{\tilde{t}} (\bar{\epsilon}_\theta)_{\tilde{t}}] + 2c_{44} \Phi(\bar{\epsilon}_{rz}) + \epsilon_{33} \Phi(\bar{E}_r) + \epsilon_{11} \Phi(\bar{E}_z) \}, \end{aligned} \quad (4.3)$$

and

$$\Phi(y) = \frac{(y + \check{y})^2}{4} - \frac{\tau^2}{4} (y_{\tilde{t}})^2.$$

We impose the non-negativeness requirement on the difference analogue of the energy functional (4.3) (later in this section we derive conditions when it is true). Summing (4.2) over t from τ to a certain t_1 ($\tau \leq t_1 \leq T$), we obtain

$$\begin{aligned} \bar{\mathcal{E}}(t_1 + \tau) = & \bar{\mathcal{E}}(\tau) + 2\tau \sum_{t'=\tau}^{t_1} \sum_{\tilde{\omega}_h} r \hbar_1 \hbar_2 (f_1 v + f_2 w) + \\ & 2\tau \sum_{t'=\tau}^{t_1} \left\{ \sum_{\tilde{\omega}_{h_1}} r \hbar_1 [\bar{\sigma}_{rz} v + \bar{\sigma}_z w]|_{Z_0}^{Z_1} + \sum_{\tilde{\omega}_{h_2}} r \hbar_2 [\bar{\sigma}_r v + \bar{\sigma}_{rz} w]|_{R_0}^{R_1} \right\}. \end{aligned} \quad (4.4)$$

Additives in the curl brackets of the right hand side of (4.4) are estimated analogously to each other. As an example we will estimate the second of them. The technique of estimation is similar to that described in [16]. We use an easily proved identity

$$u_t v = (uv)_t - \frac{\hat{u}v_t - \hat{u}v_t}{2},$$

the ϵ -inequality [27, 28, 16]

$$|(u, v)| \leq \|u\| \|v\| \leq \epsilon \|u\|^2 + \frac{1}{4\epsilon} \|v\|^2$$

and grid analogues of the Sobolev embedding theorems which we formulate below (the reader may consult [27, 28] for the proof and further references).

Lemma 4.1 *For any grid function $y(x)$ given on an arbitrary (possibly non-uniform) grid $\bar{\omega}_h \in [0, L]$ the following inequality*

$$\|y\|_C \leq \epsilon \|y_{\bar{x}}\|_0^2 + \left(\frac{1}{\epsilon} + \frac{1}{L} \right) \|y\|_0^2$$

holds for any positive constant ϵ .

Lemma 4.2 *For any grid function $y(x)$ given on an arbitrary (possibly non-uniform) grid $\bar{\omega}_h \in [0, L]$ the following inequality*

$$\|y_x\|_0^2 = \kappa_0 y^2(0) + \kappa_1 y^2(L) \geq M_2 \|y\|_0^2$$

holds for any positive constants κ_0 and κ_1 such that $\kappa_0 + \kappa_1 > 0$ and

$$M_2 = \frac{8(\kappa_0 + \kappa_1 + L\kappa_0\kappa_1)^2}{L(2 + L\kappa_0)(2 + L\kappa_1)(2\kappa_0 + 2\kappa_1 + L\kappa_0\kappa_1)}.$$

If $y(0) = y(L) = 0$, then the last inequality may be simplified to the Friedrichs inequality

$$\|y_{\bar{x}}\|_0 \geq \frac{8}{L^2} \|y\|_0^2.$$

This allows us to obtain the following result

$$\begin{aligned} \sum_{\tau=\tau}^{t_1} 2\tau \bar{\sigma}_z \omega |_{Z_0}^{Z_1} &\leq M_1 \bar{\mathcal{E}}(t_1 + \tau) + M_2 \left\{ \max_{t=0, \tau, t_1, t_1+\tau} \sum_{k=0}^1 |p_z^{(k)}|^2 + \right. \\ &\quad \left. \sum_{\tau=\tau}^{t_1} \tau \sum_{k=0}^1 |(p_z^{(k)})_t|^2 \right\} + M_3 \left\{ \sum_{\tau=\tau}^{t_1} \tau \max_{\tau-\tau, \tau'+\tau} \sum_{k=0}^1 |g(R_k, Z_0, t')|^2 + \right. \\ &\quad \left. \sum_{\tau=\tau}^{t_1} \tau \max_{\tau-\tau, \tau'+\tau} \sum_{k=0}^1 |g(R_k, Z_1, t')|^2 \right\}. \end{aligned}$$

Taking into account the latter inequality and using a similar technique for estimating other additives in (4.4) it is not difficult to obtain

$$\begin{aligned}
\bar{\mathcal{E}}(t_1 + \tau) &\leq M_4 \bar{\mathcal{E}}(\tau) + M_5 \bar{\mathcal{E}}(t_1 + \tau) + M_6 \left\{ \sum_{\tilde{\omega}_{h_1}} \max_{0, \tau, t_1, t_1 + \tau} \left[\sum_{k=0}^1 ((p_z^{(k)})^2 + \right. \right. \\
&\quad \left. \left. (p_{rt}^{(k)})^2) \right] + \sum_{\tilde{\omega}_{h_2}} r \hbar_2 \max_{0, \tau, t_1, t_1 + \tau} \left[\sum_{k=0}^1 ((p_r^{(k)})^2 + p_{rz}^{(k)})^2 \right] \right\} + \\
&\quad \sum_{t'=\tau}^{t_1} \tau \left\{ \sum_{\tilde{\omega}_{h_1}} r \hbar_1 \left[\sum_{k=0}^1 (|(p_z^{(k)})_{\tilde{t}}|^2 + |(p_{rt}^{(k)})_{\tilde{t}}|^2) \right] + \sum_{\tilde{\omega}_{h_2}} r \hbar_2 \left[\sum_{k=0}^1 (|(p_r^{(k)})_{\tilde{t}}|^2 + |(p_{rz}^{(k)})_{\tilde{t}}|^2) \right] \right\} + \\
&\quad M_7 \sum_{t'=\tau}^{t_1} \tau \left\{ \max_{t'-\tau, t'+\tau} \sum_{i,j=0}^1 [|g(R_i, Z_j, t')|^2 + |y(R_i, Z_j, t')|^2] + \right. \\
&\quad \left. \sum_{\tilde{\omega}_h} r \hbar_1 \hbar_2 (y_{\tilde{t}}^2 + g_{\tilde{t}}^2) + \sum_{\tilde{\omega}_h} r \hbar_1 \hbar_2 (f_1^2 + f_2^2) \right\}. \tag{4.5}
\end{aligned}$$

Then we use the technique of the proof of Theorem 6.1 from [18] and apply the discrete analogue of the Gronwall lemma formulated below (see, for example, [27, 28])

Lemma 4.3 *Let $g_j \geq 0$, $j = 1, 2, \dots$ and $f_j \geq 0$, $j = 0, 1, \dots$ be nonnegative grid functions (for example, $f_j \equiv f(t_j)$, $t_j = j\tau$, $j = 0, 1, \dots, n_0$). If f_j is a nondecreasing function (i.e. $f_{j+1} \geq f_j$), then from the inequality*

$$g_{j+1} \leq c_0 \sum_{k=1}^j \tau g_k + f_j, \quad j = 1, 2, \dots, g_1 \leq f_0, \quad c_0 = \text{const} > 0$$

follows the estimate

$$g_{j+1} \leq \exp(c_0 t_j) f_j.$$

Applying Lemma 4.3 from (4.5) we derive the following inequality

$$\begin{aligned}
\bar{\mathcal{E}}(t_1 + \tau) &\leq M_8 \bar{\mathcal{E}}(\tau) + M_9 \left\{ \sum_{\tilde{\omega}_{h_1}} r \hbar_1 \max_{0, \tau, t_1, t_1 + \tau} \left[\sum_{k=0}^1 ((p_z^{(k)})^2 + (p_{rt}^{(k)})^2) \right] + \right. \\
&\quad \left. \sum_{\tilde{\omega}_{h_2}} r \hbar_2 \max_{0, \tau, t_1, t_1 + \tau} \left[\sum_{k=0}^1 ((p_r^{(k)})^2 + (p_{rz}^{(k)})^2) \right] + \sum_{t'=\tau}^{t_1} \tau \left\{ \sum_{\tilde{\omega}_{h_1}} r \hbar_1 \left[\sum_{k=0}^1 (|(p_z^{(k)})_{\tilde{t}}|^2 + \right. \right. \right. \\
&\quad \left. \left. \left. |(p_{rt}^{(k)})_{\tilde{t}}|^2) \right] + \sum_{\tilde{\omega}_{h_2}} r \hbar_2 \left[\sum_{k=0}^1 (|(p_r^{(k)})_{\tilde{t}}|^2 + |(p_{rz}^{(k)})_{\tilde{t}}|^2) \right] \right\} + \sum_{\tilde{\omega}_h} r \hbar_1 \hbar_2 (f_1^2 + f_2^2) \right\}. \tag{4.6}
\end{aligned}$$

In order to obtain a bound on the discrete energy function we have to estimate the quantity $\bar{\mathcal{E}}(\tau)$. The most difficult part of this is the estimation of additives responsible for the electric field. Applying the Cauchy-Schwartz inequality to the difference analogue of the Maxwell equation (forced electrostatic of dielectrics) we arrive at the following result

$$\epsilon_{33} \sum_{\omega_h^+} \bar{r} h_1 h_2 \bar{E}_r^2 + \epsilon_{11} \sum_{\omega_h^+} \bar{r} h_1 h_2 \bar{E}_z^2 \leq \epsilon_{33} \left(\sum_{\omega_h^+} \bar{r} h_1 h_2 \bar{\epsilon}_r^2 \right)^{1/2} \left(\sum_{\omega_h^+} \bar{r} h_1 h_2 \bar{E}_r^2 \right)^{1/2} +$$

$$\begin{aligned}
& e_{13} \left(\sum_{\omega_h^+} \bar{r} h_1 h_2 (\bar{\epsilon}_\theta + \bar{\epsilon}_z)^2 \right)^{1/2} \left(\sum_{\omega_h^+} \bar{r} h_1 h_2 \bar{E}_r^2 \right)^{1/2} + 2e_{15} \left(\sum_{\omega_h^+} \bar{r} h_1 h_2 \bar{\epsilon}_{rz}^2 \right)^{1/2} \times \\
& \left(\sum_{\omega_h^+} \bar{r} h_1 h_2 \bar{E}_z^2 \right)^{1/2} + \left(\sum_{\omega_h^+} \bar{r} h_1 h_2 \bar{\lambda}^2 \right)^2 \left(\sum_{\omega_h^+} \bar{r} h_1 h_2 \bar{E}_r^2 \right)^{1/2} + \\
& \left(\sum_{\omega_h^+} \bar{r} h_1 h_2 \bar{\lambda}^2 \right)^{1/2} \left(\sum_{\omega_h^+} \bar{r} h_1 h_2 \bar{E}_z^+ \right)^{1/2}, \tag{4.7}
\end{aligned}$$

where the function $\bar{\lambda}$ is defined by the relationships

$$\left(\bar{r} \frac{\bar{\lambda} + \bar{\lambda}^{(+1_z)}}{2} \right)_r + \left(\frac{\bar{r} \bar{\lambda} + \bar{r}^{(+1)} \bar{\lambda}^{(+1_r)}}{2} \right)_z = r f_3, \tag{4.8}$$

$$\bar{\lambda}^{(+1_r)} = \bar{D}_r^{(+1_r)} \text{ for } r = R_0, \text{ and } \bar{\lambda}^{(+1_z)} = \bar{D}_z^{(+1_z)} \text{ for } z = Z_0 \quad \forall t \in \bar{\omega}_\tau. \tag{4.9}$$

Using the Cauchy inequality

$$\left(\sum_{i=1}^n a_i b_i \right) \leq \left(\sum_{i=1}^n a_i^2 \right) \left(\sum_{i=1}^n b_i^2 \right),$$

from (4.7) we get

$$\begin{aligned}
& \left(\epsilon_{33} \sum_{\omega_h^+} \bar{r} h_1 h_2 \bar{E}_r^2 + \epsilon_{11} \sum_{\omega_h^+} \bar{r} h_1 h_2 \bar{E}_z^2 \right)^2 \leq \left[\frac{\epsilon_{33}^2}{\epsilon_{33}} \sum_{\omega_h^+} \bar{r} h_1 h_2 \bar{\epsilon}_r^2 + \right. \\
& \left. \frac{e_{13}^2}{\epsilon_{33}} \sum_{\omega_h^+} \bar{r} h_1 h_2 (\bar{\epsilon}_\theta + \bar{\epsilon}_z)^2 + \frac{4e_{15}^2}{\epsilon_{33}} \sum_{\omega_h^+} \bar{r} h_1 h_2 \bar{\epsilon}_{rz}^2 + 2 \sum_{\omega_h^+} \bar{r} h_1 h_2 \bar{\lambda}^2 \right] \times \\
& \left(3 \sum_{\omega_h^+} \bar{r} h_1 h_2 \bar{E}_r^2 + 2 \sum_{\omega_h^+} \bar{r} h_1 h_2 \bar{E}_z^2 \right) \epsilon_{33}. \tag{4.10}
\end{aligned}$$

It is not difficult to find a constant M_{10} so that the inequality

$$M_{10} \left(3 \epsilon_{33} \sum_{\omega_h^+} \bar{r} h_1 h_2 \bar{E}_r^2 + 2 \epsilon_{33} \sum_{\omega_h^+} \bar{r} h_1 h_2 \bar{E}_z^2 \right) \leq \epsilon_{33} \sum_{\omega_h^+} \bar{r} h_1 h_2 \bar{E}_r^2 + \epsilon_{11} \sum_{\omega_h^+} \bar{r} h_1 h_2 \bar{E}_z^2 \tag{4.11}$$

holds, for example, by setting

$$M_{10} = \min \left\{ \frac{1}{3}, \frac{\epsilon_{11}}{2\epsilon_{33}} \right\}.$$

Therefore, using (4.11) from (4.10) we obtain

$$\begin{aligned}
& \epsilon_{33} \sum_{\omega_h^+} \bar{r} h_1 h_2 \bar{E}_r^2 + \epsilon_{11} \sum_{\omega_h^+} \bar{r} h_1 h_2 \bar{E}_z^2 \leq \frac{1}{M_{10}} \left[\frac{\epsilon_{33}^2}{\epsilon_{33}} \sum_{\omega_h^+} \bar{r} h_1 h_2 \bar{\epsilon}_r^2 + \right. \\
& \left. \frac{e_{13}^2}{\epsilon_{33}} \sum_{\omega_h^+} \bar{r} h_1 h_2 (\bar{\epsilon}_\theta + \bar{\epsilon}_z)^2 + \frac{4e_{15}^2}{\epsilon_{33}} \sum_{\omega_h^+} \bar{r} h_1 h_2 \bar{\epsilon}_{rz}^2 + 2 \sum_{\omega_h^+} \bar{r} h_1 h_2 \bar{\lambda}^2 \right]. \tag{4.12}
\end{aligned}$$

Taking into account (4.12) and the non-negativity of the potential energy of deformation (see (1.7)) from (4.6) we obtain the following estimate for the solution of the discrete problem (3.5)–(3.8)

$$\begin{aligned} \bar{\mathcal{E}}(t_1 + \tau) \leq M & \left\{ \rho \sum_{\omega_h^+} r \bar{h}_1 \bar{h}_2 \left((y_t(0))^2 + (g_t(0))^2 \right) + \sum_{\omega_h^+} \bar{r} h_1 h_2 \left\{ c_{33} \left[(\bar{\epsilon}_r(0))^2 + \frac{\tau^2}{4} ((\bar{\epsilon}_r(0))_t)^2 \right] + \right. \right. \\ & c_{11} \left[(\bar{\epsilon}_\theta(0))^2 + (\bar{\epsilon}_z(0))^2 + \frac{\tau^2}{4} ((\bar{\epsilon}_\theta(0))_t)^2 + ((\bar{\epsilon}_z(0))_t)^2 \right] + c_{13} [\bar{\epsilon}_r(0) (\bar{\epsilon}_\theta(0) + \bar{\epsilon}_z(0)) + \right. \\ & \left. \frac{\tau}{2} (\bar{\epsilon}_r(0) ((\bar{\epsilon}_\theta(0))_t + (\bar{\epsilon}_z(0))_t) + (\bar{\epsilon}_r(0))_t (\bar{\epsilon}_\theta(0) + \bar{\epsilon}_z(0))) \right] + c_{12} \left[\bar{\epsilon}_z(0) \bar{\epsilon}_\theta(0) + \frac{\tau}{2} (\bar{\epsilon}_z(0) \times \right. \\ & (\bar{\epsilon}_\theta(0))_t + (\bar{\epsilon}_z(0))_t \bar{\epsilon}_\theta(0))] + 2c_{44} \left[(\bar{\epsilon}_{rz}(0))^2 + \frac{\tau^2}{4} ((\bar{\epsilon}_{rz}(0))_t)^2 \right] \left. \right\} + \sum_{\omega_{h_1}} r \bar{h}_1 \times \\ & \max_{0, \tau, t_1, t_1 + \tau} \left[\sum_{k=0}^1 \left((p_z^{(k)})^2 + (p_{rt}^{(k)})^2 \right) \right] + \sum_{\omega_{h_2}} r \bar{h}_2 \max_{0, \tau, t_1, t_1 + \tau} \left[\sum_{k=0}^1 \left((p_r^{(k)})^2 + (p_{rz}^{(k)})^2 \right) \right] + \\ & \sum_{t'=\tau}^{t_1} \left\{ \sum_{\omega_{h_1}} r \bar{h}_1 \left[\sum_{k=0}^1 \left(|(p_z^{(k)})_t|^2 + |(p_{rt}^{(k)})_t|^2 \right) \right] + \sum_{\omega_{h_2}} r \bar{h}_2 \left[\sum_{k=0}^1 \left(|(p_r^{(k)})_t|^2 + \right. \right. \right. \\ & \left. \left. \left. |(p_{rz}^{(k)})_t|^2 \right) \right] \} + \sum_{\omega_h^+} \bar{r} h_1 h_2 (\bar{\lambda}(0))^2 + \sum_{t'=\tau}^{t_1} \tau \sum_{\omega_h} r \bar{h}_1 \bar{h}_2 (f_1^2 + f_2^2) \right\}. \end{aligned} \quad (4.13)$$

We recall that the estimate (4.13) has been obtained under the requirement of non-negativity of the discrete analogue of the energy functional $\bar{\mathcal{E}}(t)$. Such a requirement leads to the stability conditions for the difference scheme (3.5)–(3.8) [16]. Indeed, applying the technique of the derivation of estimate (4.12), we have

$$\begin{aligned} -\epsilon_{33} \sum_{\omega_h^+} \bar{r} h_1 h_2 ((\bar{E}_r)_t)^2 - \epsilon_{11} \sum_{\omega_h^+} \bar{r} h_1 h_2 ((\bar{E}_z)_t)^2 \geq \frac{1}{\epsilon^M} \left[\frac{e_{33}^2}{\epsilon_{33}} \sum_{\omega_h^+} \bar{r} h_1 h_2 ((\bar{\epsilon}_r)_t)^2 + \right. \\ \left. \frac{e_{13}}{\epsilon_{33}} \sum_{\omega_h^+} \bar{r} h_1 h_2 ((\bar{\epsilon}_\theta + \bar{\epsilon}_z)_t)^2 + \frac{4e_{15}^2}{\epsilon_{33}} \sum_{\omega_h^+} \bar{r} h_1 h_2 ((\bar{\epsilon}_{rz})_t)^2 \right], \end{aligned} \quad (4.14)$$

where $\epsilon^M = \min\{\frac{1}{2}, \frac{\epsilon_{11}}{\epsilon_{33}}\}$. Hence, taking into account condition (1.7) and estimate (4.14) we obtain the following condition for non-negativity of $\bar{\mathcal{E}}$ (see [16] for the one-dimensional case)

$$\begin{aligned} \rho \sum_{\omega_h} r \bar{h}_1 \bar{h}_2 ((y_t)^2 + (g_t)^2) - \tau^2 \left(\frac{c_{33}}{4} + \frac{e_{33}^2}{4\epsilon_{33}\epsilon^M} \right) \sum_{\omega_h^+} \bar{r} h_1 h_2 ((\bar{\epsilon}_r)_t)^2 - \\ \tau^2 \left[\frac{c_{11}}{4} + \frac{e_{13}^2}{4\epsilon_{33}\epsilon^M} \right] \sum_{\omega_h^+} \bar{r} h_1 h_2 ((\bar{\epsilon}_\theta + \bar{\epsilon}_z)_t)^2 - \tau^2 \left[\frac{2c_{44}}{4} + \frac{(2e_{15})^2}{4\epsilon_{33}\epsilon^M} \right] \times \\ \sum_{\omega_h^+} \bar{r} h_1 h_2 ((\bar{\epsilon}_{rz})_t)^2 - \frac{\tau^2}{2} c_{13} \sum_{\omega_h^+} \bar{r} h_1 h_2 (\bar{\epsilon}_r)_t ((\bar{\epsilon}_\theta)_t + (\bar{\epsilon}_z)_t) - \frac{\tau^2}{2} c_{12} \times \end{aligned}$$

$$\sum_{\omega_h^+} \bar{r} h_1 h_2 (\bar{\epsilon}_z)_t (\bar{\epsilon}_\theta)_t \geq 0. \quad (4.15)$$

Using the easily proved inequalities

$$\sum_{\omega_h^+} \bar{r} h_1 h_2 (\bar{\epsilon}_r)^2 \leq \frac{4}{h_1^2} \left(1 + \frac{h_1}{2R_0}\right) \sum_{\omega_h} r \bar{h}_1 \bar{h}_2 y^2, \quad \sum_{\omega_h^+} \bar{r} h_1 h_2 (\bar{\epsilon}_\theta)^2 \leq \frac{1}{R_0^2} \sum_{\omega_h} r \bar{h}_1 \bar{h}_2 y^2,$$

$$\sum_{\omega_h^+} \bar{r} h_1 h_2 (\bar{\epsilon}_z)^2 \leq \frac{4}{h_2^2} \left(1 + \frac{h_1}{2R_0}\right) \sum_{\omega_h} r \bar{h}_1 \bar{h}_2 g^2, \quad \sum_{\omega_h^+} \bar{r} h_1 h_2 \bar{\epsilon}_r \bar{\epsilon}_\theta \leq \frac{1}{2R_0 h_1} \sum_{\omega_h} r \bar{h}_1 \bar{h}_2 y^2,$$

$$\sum_{\omega_h^+} \bar{r} h_1 h_2 (\bar{\epsilon}_{rz})^2 \leq \frac{1}{2} \left[\frac{4}{h_2^2} \sum_{\omega_h} r \bar{h}_1 \bar{h}_2 y^2 \left(1 + \frac{h_1}{2R_0}\right) + \frac{4}{h_1^2} \sum_{\omega_h} r \bar{h}_1 \bar{h}_2 g^2 \left(1 + \frac{h_1}{2R_0}\right) \right],$$

$$\sum_{\omega_h^+} \bar{r} h_1 h_2 \bar{\epsilon}_z \bar{\epsilon}_\theta \leq \frac{1}{4R_0 h_2} \sum_{\omega_h} r \bar{h}_1 \bar{h}_2 (g^2 + y^2), \quad \sum_{\omega_h^+} \bar{r} h_1 h_2 \bar{\epsilon}_r \bar{\epsilon}_z \leq \frac{1}{2h_1 h_2} \sum_{\omega_h} r \bar{h}_1 \bar{h}_2 (y^2 + g^2),$$

it can be shown that (4.15) will be satisfied if the inequality

$$\begin{aligned} & \sum_{\omega_h} r \bar{h}_1 \bar{h}_2 (y_t)^2 \left\{ \rho - \tau^2 \left[\frac{4}{h_1^2} \left(1 + \frac{h_1}{2R_0}\right) \left(\frac{c_{33}}{4} + \frac{e_{33}^2}{4\epsilon_{33}\epsilon^M}\right) + \frac{1}{R_0^2} \left(\frac{c_{11}}{4} + \frac{e_{13}^2}{4\epsilon_{33}\epsilon^M}\right) + \right. \right. \\ & \left. \left. \frac{4}{h_2^2} \left(1 + \frac{h_1}{2R_0}\right) \left(\frac{c_{44}}{4} + \frac{e_{15}^2}{2\epsilon_{33}\epsilon^M}\right) + c_{13} \frac{1}{4h_1 R_0} + c_{13} \frac{1}{4h_1 h_2} + c_{12} \frac{1}{8h_2 R_0} \right] \right\} + \\ & \sum_{\omega_h} r \bar{h}_1 \bar{h}_2 (g_t)^2 \left\{ \rho - \tau^2 \left[\frac{4}{h_2^2} \left(1 + \frac{h_1}{2R_0}\right) \left(\frac{c_{11}}{4} + \frac{e_{13}^2}{4\epsilon_{33}\epsilon^M}\right) + \right. \right. \\ & \left. \left. \frac{4}{h_1^2} \left(1 + \frac{h_1}{2R_0}\right) \left(\frac{c_{44}}{4} + \frac{e_{15}^2}{2\epsilon_{33}\epsilon^M}\right) + \frac{c_{13}}{4h_1 h_2} + \frac{c_{12}}{8h_2 R_0} \right] \right\} \geq \epsilon, \end{aligned}$$

holds for $\epsilon > 0$.

The latter is satisfied if the following two inequalities

$$\begin{cases} \rho - \epsilon_1^0 \geq \frac{\tau^2}{h_1^2} \left[\left(1 + \frac{h_1}{2R_0}\right) \left(c_{33} + \frac{e_{33}^2}{\epsilon_{33}\epsilon^M}\right) + \frac{c_{13}}{8} \frac{h_1}{h_2} + \frac{c_{13}}{4R_0} h_1 + \frac{1}{4R_0^2} \times \right. \\ \left. \left(c_{11} + \frac{e_{13}^2}{\epsilon_{33}\epsilon^M} \right) h_1^2 \right] + \frac{\tau^2}{h_2^2} \left[\left(1 + \frac{h_1}{2R_0}\right) \left(c_{44} + \frac{2e_{15}^2}{\epsilon_{33}\epsilon^M}\right) + \frac{c_{12}}{8R_0} h_2 + \frac{c_{13}}{8} \frac{h_2}{h_1} \right], \\ \rho - \epsilon_2^0 \geq \frac{\tau^2}{h_2^2} \left[\left(1 + \frac{h_1}{2R_0}\right) \left(c_{11} + \frac{e_{13}^2}{\epsilon_{33}\epsilon^M}\right) + \frac{c_{12}}{8R_0} h_2 + \frac{c_{13}}{8} \frac{h_2}{h_1} \right] + \\ + \frac{\tau^2}{h_1^2} \left[\left(1 + \frac{h_1}{2R_0}\right) \left(c_{44} + \frac{2e_{15}^2}{\epsilon_{33}\epsilon^M}\right) + \frac{c_{13}}{8} \frac{h_1}{h_2} \right], \end{cases} \quad (4.16)$$

are satisfied simultaneously for $\epsilon_i^0 > 0$, $i = 1, 2$.

It is known [2, 6] that in the general case in an anisotropic electro-elastic medium there are three plane waves, namely quasi-longitudinal and two quasi-transverse (the latter are usually propagated slower than quasi-longitudinal). Therefore we introduce three quantities

$$c_1 = \sqrt{\frac{c_{33}(1+K_1)}{\rho}}, \quad c_2 = \sqrt{\frac{c_{44}(1+K_2)}{\rho}}, \quad c_3 = \sqrt{\frac{c_{11}(1+K_3)}{\rho}}, \quad (4.17)$$

that characterise velocities of these waves, where

$$K_1 = \frac{e_{33}^2}{\epsilon_{33} c_{33}}, \quad K_2 = \frac{e_{15}^2}{\epsilon_{11} c_{44}}, \quad K_3 = \frac{e_{13}^2}{\epsilon_{11} c_{11}} \quad (4.18)$$

are constants of electromechanical coupling. Finally from (4.16) we can obtain stability conditions for the difference scheme (3.5)–(3.8)

$$\left\{ \begin{array}{l} \frac{\tau^2}{h_1^2} c_1^2 \left[\left(1 + \frac{h_1}{2R_0}\right) \frac{1+K_1/\epsilon^M}{1+K_1} + \frac{c_{13}}{8c_{33}(1+K_1)} \frac{h_1}{h_2} + \frac{c_{13}h_1}{4R_0c_{33}(1+K_1)} + \right. \\ \left. \frac{1}{4R_0^2c_{33}(1+K_1)} \left(c_{11} + \frac{e_{13}^2}{\epsilon_{33}\epsilon^M}\right) h_1^2 \right] + \frac{\tau^2}{h_2^2} c_2^2 \left[\left(1 + \frac{h_1}{2R_0}\right) \times \right. \\ \left. \frac{1+2K_2/\epsilon^M}{1+K_2} + \frac{c_{13}}{8c_{44}(1+K_2)} \frac{h_2}{h_1} + \frac{c_{12}}{8R_0c_{44}(1+K_2)} h_2 \right] \leq 1 - \epsilon_1, \\ \frac{\tau^2}{h_2^2} c_3^2 \left[\left(1 + \frac{h_1}{2R_0}\right) \frac{1+K_3/\epsilon^M}{1+K_3} + \frac{c_{13}}{8c_{11}(1+K_3)} \frac{h_2}{h_1} + \frac{c_{12}h_2}{8R_0c_{11}(1+K_1)} \right] + \\ \frac{\tau^2}{h_1^2} c_2^2 \left[\left(1 + \frac{h_1}{2R_0}\right) \frac{1+2K_2/\epsilon^M}{1+K_2} + \frac{c_{13}}{8c_{44}(1+K_2)} \frac{h_1}{h_2} \right] \leq 1 - \epsilon_2, \end{array} \right. \quad (4.19)$$

where $\epsilon, \epsilon_i, \epsilon_i^0, i=1,2$ are positive constants that do not depend on steps τ, h_1 and h_2 .

As a result we have proved the following theorem.

Theorem 4.1 *If conditions (4.19) are satisfied for the solution of the discrete model (3.5)–(3.8) then the estimate (4.13) with the discrete energy function defined by (4.3) holds for arbitrary $t_1 > 0$.*

Remark 4.1 *The conditions (4.19) play the same role in 2D dynamic electroelasticity as the Courant-Friederichs-Lowy stability conditions do in the classical theory of hyperbolic equations. They connect steps τ, h_1 and h_2 with the velocity of mixed electro-elastic waves.*

Remark 4.2 *In the case of isotropic solid in the absence of electric field [6] we have that*

$$c_{11} = c_{33} = \tilde{\lambda} + 2\mu, \quad c_{12} = c_{13} = \tilde{\lambda}, \quad c_{44} = \tilde{\mu},$$

$$K_1 = K_2 = K_3 = 0,$$

where $\tilde{\lambda}$ and $\tilde{\mu}$ are the Lame coefficients.

Therefore from (4.19) it is straightforward to obtain stability conditions in the principal part

$$\frac{\tau^2}{h_1^2} \left(\tilde{\lambda} + 2\tilde{\mu} + \frac{\tilde{\lambda}}{8} - \frac{h_1}{h_2} \right) + \frac{\tau^2}{h_2^2} \left(\tilde{\mu} + \frac{\lambda}{8} - \frac{h_2}{h_1} \right) \leq 1 - \epsilon_1, \quad (4.20)$$

$$\frac{\tau^2}{h_2^2} \left(\tilde{\lambda} + 2\tilde{\mu} + \frac{\tilde{\lambda}}{8} - \frac{h_2}{h_1} \right) + \frac{\tau^2}{h_1^2} \left(\tilde{\mu} + \frac{\lambda}{8} - \frac{h_1}{h_2} \right) \leq 1 - \epsilon_2. \quad (4.21)$$

Summing inequalities (4.20) and (4.21) and taking into account

$$\frac{\tau^2}{h_1 h_2} \leq \frac{1}{2} \left(\frac{\tau^2}{h_1^2} + \frac{\tau^2}{h_2^2} \right),$$

we obtain

$$(\gamma_1^2 + \gamma_2^2) \left(\tilde{\lambda} + 3\tilde{\mu} + \frac{\lambda}{4} \right) \leq 1 - \epsilon, \quad \epsilon > 0, \quad \gamma_i = \tau/h_i, \quad i = 1, 2. \quad (4.22)$$

The condition (4.22) coincides with the stability condition obtained in [21] for pure elastic problems.

5 Numerical Results

In this section we present some typical results of computation for thin hollow finite-length PZT-piezoceramic cylinders poled radially and circularly. In addition to the low cost and relatively easy process of fabrication PZT-based materials have many advantages including the resistance to mechanical and electrical stress-induced depolarisation and moderate operating temperature range. Large piezoelectric coefficients of PZT-ceramic allow the extensive use of these materials in industrial applications.

Our interest in hollow piezoceramic structures is inspired by the existing and potential applications of miniaturised piezoceramic transducers in acoustics, biomedical imaging, sensor and hydrophone design, consumer electronics and other areas [7, 20]. Many challenges of modelling piezoceramic structures during recent years have also been connected with “smart” properties of materials. The word *smart* reflects a striking resemblance in the behaviour of some piezoceramic materials to the behaviour of biological systems. For example, the mechanism by which fishes sense vibrations can be mimicked by piezoelectric hydrophones. Some piezoelectric materials are able to “remember” their original shape whereas others can even “learn” as environmental properties change (see [24] and references therein).

In our computational experiments we use numerical scheme (3.5)–(3.8) which satisfy the conditions (4.19). We consider finite-length ($H=1$) cylinders with thicknesses $l = 0.2$ and $l = 0.05$. In our earlier paper (see [16] for details) we performed similar experiments with a one-dimensional model for infinite-length cylinders of the same thicknesses.

Figure 1 presents the time dependency of radial displacements on the external surface of cylinders. Compared to the infinite-length cylinder case discussed in [16], we observe a noticeable decrease in the magnitude of displacements for cylinders poled radially. In contrast we observe a certain increase in the magnitude of displacements for finite-length cylinders poled circularly. However, in the latter case such an increase is not quantitatively essential. Displacements and stresses at a fixed moment in time $t = 10$ in finite-time cylinders poled radially are shown in Figures 2 and 3.

We conclude that in the case of cylinders poled radially, the accounting for the coupling of electric and elastic fields in anisotropic piezoceramic materials is an important prerequisite for adequate modelling of piezoelectric structures.

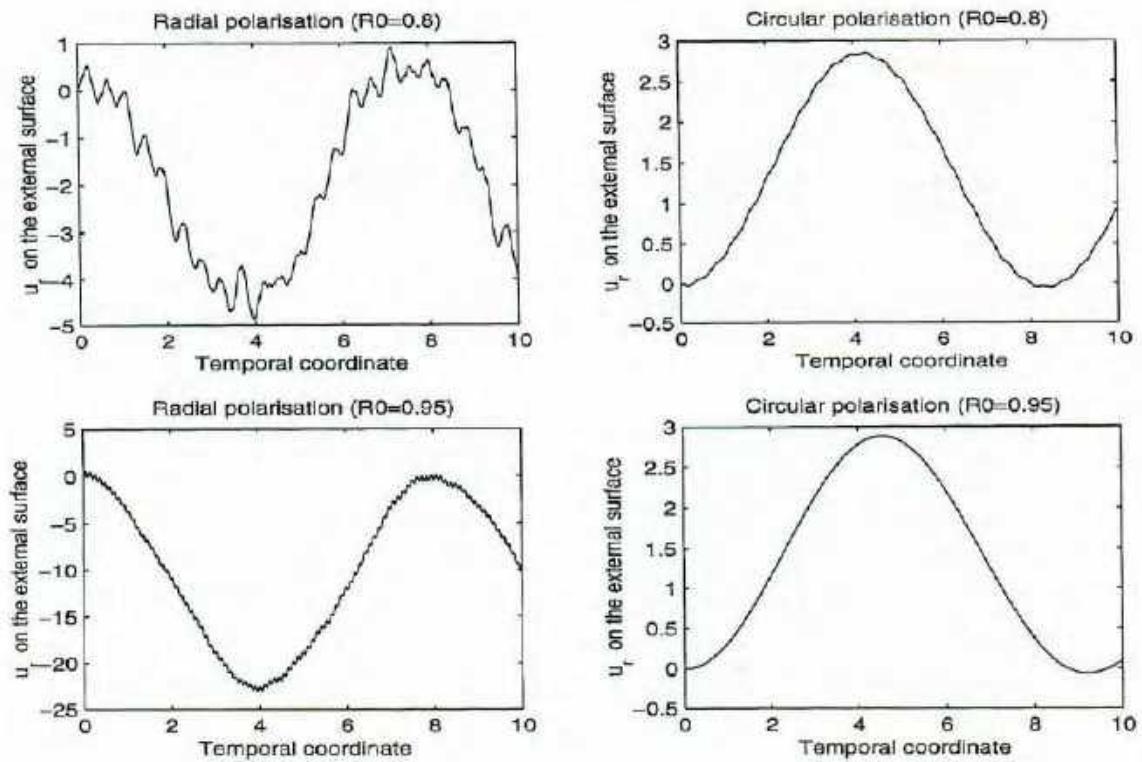


Figure 1: Dynamics of radial displacements on the external surface.

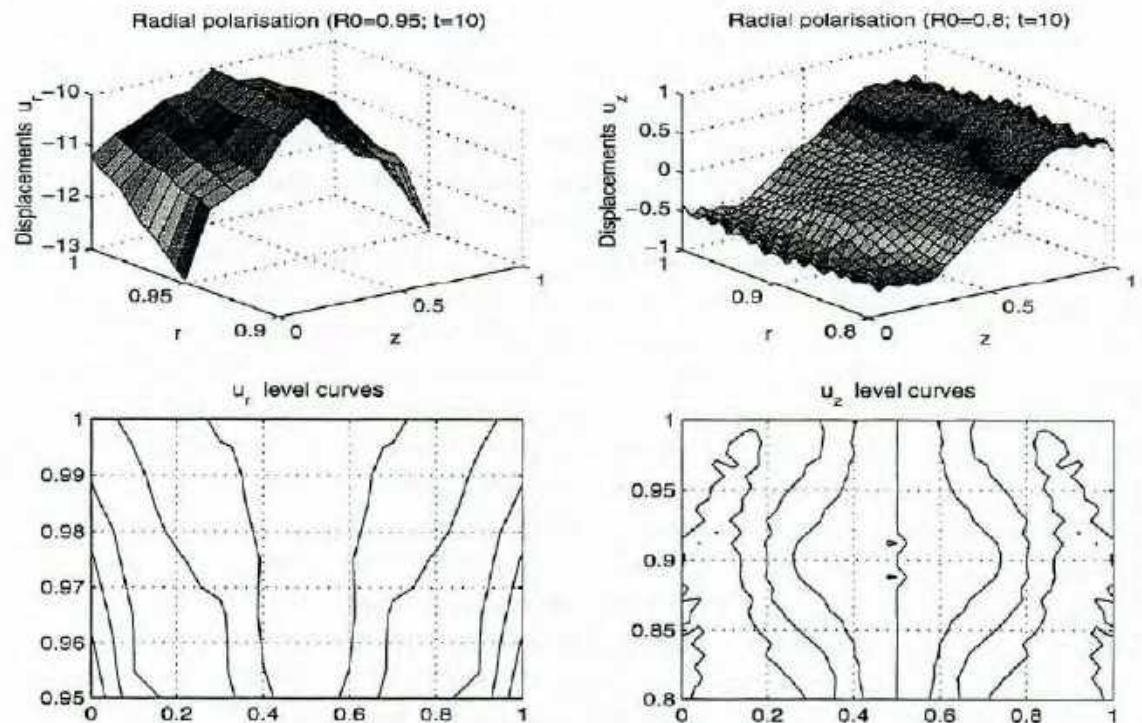


Figure 2: Radial and axial displacements at $t=10$.

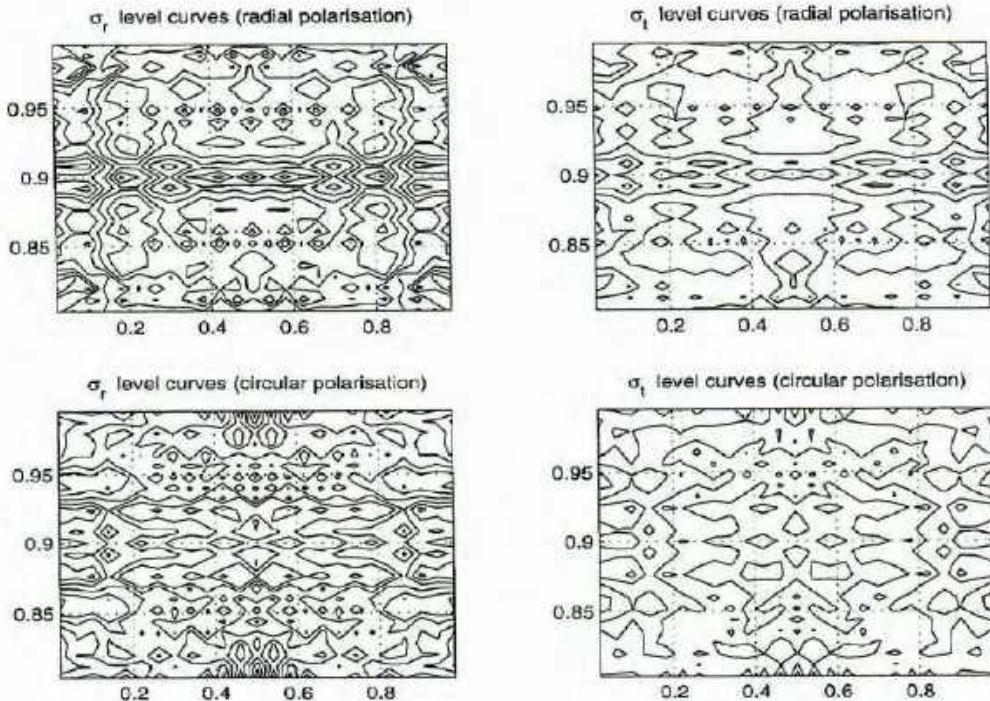


Figure 3: The distribution of stresses in piezoceramic cylinders ($l=0.2, t=10$).

6 Concluding remarks.

A major part of mathematical models for the description of phenomena of continuum physics are based on the local equilibrium hypothesis. A departure from such a hypothesis necessitated by many physical applications [11, 22] means taking into account hyperbolic features of the dynamics of physical phenomena. As a result, mathematical models that describe such phenomena have to include either a hyperbolic-type operator or a mixed operator that contains a hyperbolic mode. Typically such models have to be treated numerically.

In their paper published in *Mathematische Annalen*, Vol. 100 in 1928 R. Courant, K. Friedrichs and H. Lewy wrote that “we can expect convergence of the sequence of mesh functions to the solution of the differential equation only provided the pyramid of determination contains the cone of determination of the differential equation in its interior” (see English translation by P. Fox [4]). The difficulties with the actual proof of such expectations lie with the fact that before we construct a grid we have to know some information of “the cone of determination”. As a result of the necessity for such *a priori* knowledge some “rough” information will inevitably be implemented into the mathematical model [17]. The measure of quality of such informational “roughness” with respect to the real-world applications cannot be estimated with exclusively mathematical tools. In the understanding of many physical, biological and chemical phenomena we have to include more information in mathematical models, consider additional effects, and take into consideration the inter-influence of different fields on each other. The quality of algorithms derived from these models is eventually determined by the consistency of the models with the real-world application. As a result, the quality of the physical parameterisation of mathematical models has been playing an increasing role. Moreover, being an integral part of

the model construction the physical parameterisation may decisively influence the complexity of numerical algorithms. Under such circumstances the investigation of the stability of discrete models becomes the main issue in the success of the whole mathematical modelling enterprise.

Acknowledgements

The work was supported by grant USQ-PTRP 17989. The authors are grateful to Michael Simpson for his helpful assistance at the final stage of preparation of this paper.

References

- [1] Ballato, A., Piezoelectricity: Old Effect, New Thrusts, *IEEE Trans. on Ultrasonics, Ferroelectrics, and Frequency Control*, **42**, No. 5, 916, 1995.
- [2] Berlincourt, D.A., Curran, D.R., and Jaffe, H. *Piezoelectric and Piezomagnetic Materials and Their Function in Transducers*, in "Physical Acoustics", Vol. 1A, Ed. W.P. Mason, New York and London: Academic Press, 1964, 204-236.
- [3] *Ceramic Materials for Electronics: Processing, Properties, and Applications*, Ed. R.C. Buchanan, Marcel Dekker, 1991.
- [4] Courant, R., Friedrichs, and H. Lewy, On partial differential equations of mathematical physics, *Technical Report NYO-7689, AEC Computing Facility, Institute of Mathematical Sciences, New York University, September, 1956*.
- [5] Crawley, E.F., Intelligent structures for aerospace: a technology overview and assessment, *AIAA Journal*, **32**, 1689-1699.
- [6] Dieulesaint, E., Royer, D., *Elastic Waves in Solids: Applications to Signal Processing*, Chichester; N.Y.: J.Wiley, 1980.
- [7] Fielding, J.T. et al, Characterization of PZT Hollow-Sphere Transducers, *Proceedings of the IX IEEE International Symposium on Applications of Ferroelectrics*, 1994, 202-205.
- [8] Fukada, E., Poiseuille Medal Award Lecture: Piezoelectricity of biopolymers, *Biorheology*, **32**, 593, 1995.
- [9] Gururaja, T.R., Piezoelectric Transducers for Medical Ultrasonic Imaging, *American Ceramic Society Bulletin*, Vol. 73, No.5, 1994, 50-55.
- [10] Ikeda, T., *Fundamentals of Piezoelectricity*, Oxford: Oxford University Press, 1990.
- [11] Jou, D., Casas-Vazquez, J., Lebon, G., *Extended Irreversible Thermodynamics*, Springer-Verlag, Berlin, 1993.
- [12] Kagawa, Y., Tsuchiya, T., Kawashima, T., Finite Element Simulation of Piezoelectric Vibrator Gyroscopes, *IEEE Trans. Ultrason., Ferroelect., Freq. Contr.*, Vol. 43, No. 4, 1996, 509-520.
- [13] Lee, J.S., Boundary Element Method for Electroelastic Interaction in Piezoceramics, *Engineering Analysis with Boundary Elements*, Vol. 15, No. 4, 1995, 321-328.

- [14] Lu, P., Mahrenholtz, O., A Variational Boundary Element Formulation for Piezoelectricity, *Mechanics Research Communications*, Vol. 21, No.6, 1994, 605–615.
- [15] Melnik, R.V.N., Moskalkov, M.N., Difference Schemes for and Analysis of Approximate Solutions of Two-Dimensional Nonstationary Problems in Coupled Electroelasticity, *Differential Equations*, Vol. 27, No. 7, 1991, 1220–1230 (by Plenum Publishing Corporation/Consultants Bureau, N.Y., 1992, 860–867).
- [16] Melnik, R.V.N., The stability condition and energy estimate for non-stationary problems of coupled electroelasticity, *Mathematics and Mechanics of Solids*, Vol. 2, No. 2, 1997 153–180.
- [17] Melnik, R.V.N., On consistent regularities of control and value functions, *Numer. Funct. Anal. and Optimiz.*, 18(3&4), 401–426, 1997.
- [18] Melnik, R.V.N., Generalised solutions, discrete models and energy estimates for a 2D problem of coupled field theory, *TR 1997/12, Centre for Industrial and Applied Mathematics, School of Mathematics, University of South Australia, 1997, submitted*.
- [19] Melnik, R.V.N., Convergence of the operator-difference scheme to generalized solutions of a coupled field theory problem, to appear in *Journal of Difference Equations and Applications*, 1998.
- [20] Meyer, R. Jr. et al, Lead zirconate titanate hollow-sphere transducers, *J. Am. Cer. Soc.*, 77, 1994, 1669–1672.
- [21] Moskalkov, M.N., Utebaev, D., On the convergence of centered difference schemes for dynamic problems of elasticity theory, *Differential Equations*, Vol. 21, No. 7, 1985, 1238–1246.
- [22] Muller, I., Ruggeri, T., *Extended Thermodynamics*, Springer-Verlag, New York, 1993.
- [23] Nam, C., Kim, Y., and Weisshaar, T.A., Optimal sizing and placement of piezo-actuators for active flutter suppression, *Smart Mater. Struct.*, 5, 1996, 216–224.
- [24] Newnham, R.E., Markowski, K.A., Composite transducer and actuators, *Proceedings of the IX IEEE International Symposium on Applications of Ferroelectrics*, 1994, 705–708.
- [25] Rahmoune, M., Latour, M., Application of mechanical waves induced by piezofilms to marine fouling protection of oceanographic sensors, *Smart Mater. Struct.*, 4, 1995, 195–201.
- [26] *Piezoelectricity*, Eds. C.Z. Rosen, B.V. Hiresmath, R. Newnham, N.Y. American Institute of Physics, 1992.
- [27] Samarskii, A.A., *Theorie der Differenzenverfahren*, (Series: Mathematik und ihre Anwendungen in Physik und Technik), Aufl. Leipzig: Academische Verlagsgesellschaft Geest & Portig, 1984.
- [28] Samarskii, A.A., Nikolaev, E.S., *Numerical Methods for Grid Equations*, Basel, Boston: Birkhauser Verlag, 1989.
- [29] Shashkov, M., Steinberg, S., *Conservative Finite Difference Methods on General Grids*, Boca Raton: CRC Press, 1995
- [30] Uchino, K., Piezoelectric actuators/ultrasonic motors: their developments and markets, *Proceedings of the IX IEEE International Symposium on Applications of Ferroelectrics*, 1994, 319–324.

USQ



TOOWOOMBA

**APPLICATION OF ALTPACK TO THE
SOLUTION OF NONLINEAR PDEs**

Roderick V Nicholas Melnik

Department of Mathematics & Computing, USQ

Faculty of Sciences Working Paper Series

SC-MC-9706

November 1997

THE UNIVERSITY OF
S O U T H E R N
Q U E E N S L A N D

Working Papers

FACULTY
OF
SCIENCES

APPLICATION OF ALTPACK TO THE SOLUTION OF NONLINEAR PDEs

Roderick V Nicholas Melnik

Department of Mathematics & Computing, USQ

Faculty of Sciences Working Paper Series

SC-MC-9706

November 1997

APPLICATION OF ALTPACK TO THE SOLUTION OF NONLINEAR PDEs

Roderick V. Nicholas Melnik

Department of Mathematics and Computing,
University of Southern Queensland, QLD 4350, Australia
E-mail: melnik@usq.edu.au

Abstract

In this article the author applies the alternating-triangular method (ATM) to the solution of a number of 2D initial-boundary value problems that can be modelled by non-linear parabolic equations with a source term. Different types of non-linearities are considered and the procedure of finding approximate solutions is explained. The main features of the ALTPACK package are outlined and the results of computational experiments are presented.

Key words: non-linear parabolic equations with source terms, alternating-triangular method.

1 Introduction.

The application of mathematical models based on linear and nonlinear elliptic and parabolic equations is extremely wide [2, 6, 10, 15, 23, 9, 16]. It includes such areas as heat conduction, diffusion, electron and ion thermoconductivity, combustion and porous media modelling, chemical kinetics, semiconductor device modelling, biophysics etc. Such models differ from each other only by types of dependencies of coefficients and source terms. From the point of view of constructive investigations it is exactly such types that define the “fingerprint” of a specific model. In some cases such “fingerprints” may be treated using quite general techniques, while in other they require individual approaches [21].

For a start let us consider the linear equation of thermoconductivity. It is known that the fundamental solution of the operator

$$\mathcal{L}(\Phi) = \frac{\partial \Phi}{\partial t} - c^2 \Delta \Phi, \quad (1.1)$$

in the space-time region $\mathbb{R}^n \times \mathbb{R}_+$, i.e. the solution of the equation

$$\mathcal{L}(\Phi) \equiv \frac{\partial \Phi}{\partial t} - c^2 \Delta \Phi = \delta(\mathbf{x}, t), \quad (1.2)$$

can be expressed as

$$\Phi(\mathbf{x}, t) = \frac{H(t)}{(2c\sqrt{\pi t})^n} e^{-\frac{|\mathbf{x}|^2}{4c^2 t}}, \quad (1.3)$$

where $H(t)$ is the Heaviside function, $\mathbf{x} = (x_1, \dots, x_n)$ and c is the thermoconductivity coefficient (see, for example, [23]). Theoretically the solution (1.3) gives the distribution of temperature from the point instant source $\delta(\mathbf{x})\delta(t)$. However since $\Phi(\mathbf{x}, t) > 0 \forall t > 0, \mathbf{x} \in R^n$ one observes that such a solution might be an adequate approximation of the heat transfer process only under certain constraints (such as large values of \mathbf{x} and small values of t). Since such constraints may not conform to many applied problems in a number of situations it is reasonable either

- to use a nonlinear thermoconductivity equation, or
- to apply integro-differential transport equation

in order to describe heat processes more precisely. In the later case such an integro-differential equation may incorporate non-locality and memory effects, and gives a parabolic equation only in the limit of small deviations from the equilibrium state [11]. It is also possible to consider a hyperbolic heat equation [1, 17] and/or to take into account the influence of non-thermal physical fields [13].

In its essence all of the above options can be cast in an equation of the following type

$$u_t = A(u, f), \quad (1.4)$$

where $A(u, f)$ is a differential or integro-differential operator, and the functional f represents source-related terms. In all applied problems the definition of such a functional f is the subject of an approximation. For a specific choice f there may exist a stationary solution u^* such that

$$A(u^*, f) = 0, \quad (1.5)$$

and in this case one may expect that

$$u(\cdot, t) \rightarrow u^* \text{ when } t \rightarrow \infty. \quad (1.6)$$

However, in the general case there are such approximations f that certain measures of the $u(\cdot, t)$ tends to infinity when time approaches a finite critical value. In the latter case it is reasonable to seek such an approximate solution(s) of the equation (1.4) that will be a stationary solution y^* of another equation

$$y_t = \tilde{A}(y, \tilde{f}), \quad (1.7)$$

i.e. such that $u \approx y^*$, where $\tilde{A}(y^*, \tilde{f}) = 0$. For example, there exist a large class of problems (1.4) for which the equation (1.7) is the Hamilton-Jacobi-Bellman-type equation [3, 5].

The equation (1.7) may be seen as an asymptotic approximation of (1.4) with respect to a certain small parameter ϵ which is included into the original equation implicitly (or in some cases explicitly) by the nonlinearities of sources and/or equation coefficients. Although the asymptotic behaviour of (1.4) may be controlled by stationary solutions of (1.7), the efficiency of such a control is essentially dependent on ϵ . In other words, ϵ may be small, but it is always positive, and as a result, y may appear to be a very rough approximation to u for fairly large intervals of time.

Ideally, investigations of parabolic models have to blend asymptotic and numerical techniques for their solution. In many cases asymptotic approximations may give an important qualitative perception of the solution and provide certain guidelines for the obtaining of quantitative picture by a numerical method. The major difficulties with asymptotic approaches lie with an approximate nature of the functional f in the model (1.4). Considering real physical processes in non-linear dissipative media one has to specify the law of energy loss in order to make the model relevant to the real-world applications. In many cases such specifications may be incorporated approximately through the boundary conditions of the problem (for example,

in the combustion theory we often have to take into account the loss of heat on the boundaries of the flame region). Conceptually, approximations (incorporated into the model through nonlinearities, initial and boundary conditions) lead to a continuum set of possibilities in the simulation of the evolution of dissipative processes. In each specific case such approximations may give rise to discrete space-time scales [21], which can be often grasped numerically with a required accuracy [4]. Therefore, in order to reveal the structure of dissipative processes numerical methods will remain a major and the most natural tool in investigations of the evolutions of such processes in non-homogeneous dissipative media.

In this paper we apply an effective numerical method to a number of nonlinear problems that recently attracted a lot of attention of researchers [21, 5, 8]. In a vast majority of cases such attention was devoted to the Cauchy problem for non-linear parabolic equations. Moreover, the application of asymptotic techniques to the investigation of space-time architecture of dissipative structures were often limited to the one dimensional case. In this paper we consider initial-boundary value problems in the two dimensional case. Main examples concern model problems that arise from complex mathematical models in semiconductor device modelling, porous media, combustion theory, reaction-diffusion applications and other applications [16, 9, 14, 10, 6, 2, 15].

The remaining part of the paper is organised as follows.

- In Section 2 we introduce notation that is used throughout the paper.
- Section 3 is devoted to the formulation of the mathematical model that is being investigated numerically.
- Section 4 describes with the main steps of the integro-interpolational approach applied for the construction of difference schemes for the problem.
- Section 5 deals with the alternating-triangular method applied to the solution of discrete problems that arise at each time-layer.
- Section 6 is devoted to the choice of accelerating parameters for the ATM.
- Algorithmic aspects of the alternating-triangular method are discussed in Section 7.
- In Section 8 we present results of computational experiments on the application of the described procedure.

2 Notation

The following notations are used throughout the article.

- $G = \{x = (x_1, x_2) : 0 \leq x_i \leq L_i, i = 1, 2\}$ denotes the spatial region of interest;
- $\bar{G} = G \cup \Gamma$ is the closed region, where Γ denotes its boundaries;
- $\bar{T} = \{t : 0 \leq t \leq T_f\}$ denotes the temporal interval of consideration with the final time denoted by T_f ;
- $\hat{\omega}_h = \hat{\omega}_1 \times \hat{\omega}_2$, denotes a nonuniform grid that covers the region \bar{G} ;
- $\hat{\omega}_1 = \{x_{1,i_1} \in [0, L_1] : x_{1,i_1} = x_{1,i_1-1} + h_{1,i_1}^-, i_1 = 0, \dots, N_1, x_{1,0} = 0, x_{1,N_1} = L_1\}$;
- $\hat{\omega}_2 = \{x_{2,i_2} \in [0, L_2] : x_{2,i_2} = x_{2,i_2-1} + h_{2,i_2}^-, i_2 = 0, \dots, N_2, x_{2,0} = 0, x_{2,N_2} = L_2\}$;
- $\hat{\omega}_{h\tau} = \bar{\omega}_\tau \times \hat{\omega}_h$, where $\bar{\omega}_\tau = \{t_k = k\tau, k = 0, \dots, L, L\tau = T_f\}$;
- ω_i denotes the set of all inner points of the grid $\hat{\omega}_i$, $i=1,2$, and $\omega_h = \omega_1 \times \omega_2$;
- γ_i - is the set of all boundary points in $x_j \in [0, L_j]$ when $x_i = 0$ and $x_i = L_i$, $i = 1, 2$, $j = 3 - i$ (hence the boundary of the grid $\hat{\omega}_h$ is defined by $\gamma = \gamma_1 \cup \gamma_2$ and $\hat{\omega}_h = \omega_h \cup \gamma$);

- $\omega_i(x_j)$ are points of the grid ω_h for fixed x_j , and $\hat{\omega}_i(x_j)$ denotes the set of points of the grid $\hat{\omega}_h$ for fixed x_j ;
- $\omega_i^+(x_j)$ - is the set of points of the grid $\omega_i(x_j)$ plus the node $x_i = L_i$;
- $\omega_i^-(x_j)$ - the set of points of the grid $\omega_i(x_j)$ plus the node $x_i = 0$;
- $h_{1,i_1}^+ = h_{1,i_1+1}^-, h_{2,i_2}^+ = h_{2,i_2+1}^-$;
- $\hbar_{1,i_1} = 0.5(h_{1,i_1}^- + h_{1,i_1}^+)$, $\hbar_{2,i_2} = 0.5(h_{2,i_2}^- + h_{2,i_2}^+)$;
- $\hbar_{i,0} = 0.5h_{i,0}^+, \hbar_{i,N_i} = 0.5h_{i,N_i}^-, i = 1, 2$;
- $x_{1,i_1 \pm 0.5} = x_{1,i_1} \pm 0.5h_{1,i_1}^\pm, x_{2,i_2 \pm 0.5} = x_{2,i_2} \pm 0.5h_{2,i_2}^\pm$.

Let further $y = y(x_1, x_2, t) \equiv y(i_1, i_2, k)$ be a function of discrete arguments $\mathbf{x} = (x_1, x_2)$ and t defined on a non-uniform grid $\hat{\omega}_{ht}$. Then we use the standard notation for difference derivatives of y [19, 20, 14, 22]. For example for a function of \tilde{y} of one variable defined on $\hat{\omega}_1$

- the backward difference derivative is

$$\tilde{y}_{\tilde{x}_1} = \frac{\tilde{y}_{i_1} - \tilde{y}_{i_1-1}}{h_{1,i_1}},$$

- the forward difference derivative is

$$\tilde{y}_{x_1} = \frac{\tilde{y}_{i_1+1} - \tilde{y}_{i_1}}{h_{1,i_1+1}},$$

- and the second (“central”) difference derivative is

$$\tilde{y}_{\tilde{x}_1 \tilde{x}_1} = \frac{1}{\hbar_{i_1}} \left[\frac{\tilde{y}_{i_1+1} - \tilde{y}_{i_1}}{h_{i_1+1}} - \frac{\tilde{y}_{i_1} - \tilde{y}_{i_1-1}}{h_{i_1}} \right].$$

By \mathcal{H} we denote the space of grid functions that are defined on $\hat{\omega}_h$ with the scalar product and the norm defined as follows

$$(u, v) = \sum_{\hat{\omega}} u(x)v(x)\hbar_1\hbar_2, \|u\| = \sqrt{(u, u)}, \quad \forall u(x), v(x) \in \mathcal{H}.$$

In an analogous way we define the scalar products on $\omega_i^+(x_j)$, $\hat{\omega}_i(x_j)$ and $\omega_i^+ \times \hat{\omega}_j$

$$(u, v)_{\omega_i^+(x_j)} = \sum_{\omega_i^+(x_j)} u(x)v(x)\hbar_i^-, (u, v)_{\hat{\omega}_i(x_j)} = \sum_{\hat{\omega}_i(x_j)} u(x)v(x)\hbar_i,$$

$$\text{and } (u, v)_i = \sum_{\omega_i^+ \times \hat{\omega}_j} u(x)v(x)\hbar_i^-\hbar_j, \quad i = 1, 2, \quad j = 3 - i.$$

Other notation used in this paper are standard in theory of difference schemes [22, 19, 14, 20].

3 Mathematical model

In the space-time region $\tilde{Q}_T = \tilde{G} \times \tilde{T}$ we consider a general type of nonlinear partial differential equation

$$a^0(\mathbf{x}, t, u) \frac{\partial u}{\partial t} = \sum_{i=1}^2 \frac{\partial}{\partial x_i} \left[k_i(\mathbf{x}, t, u) \frac{\partial u}{\partial x_i} \right] - q^0(\mathbf{x}, t, u)u + f^0(\mathbf{x}, t, u), \quad (3.1)$$

where $\mathbf{x} = (x_1, x_2)$, and $a^0, q^0, f^0, k_i, i = 1, 2$ are given functions of \mathbf{x}, t and u . Equation (3.1) is supplemented by the initial

$$u(x_1, x_2, 0) = u_0(x_1, x_2) \quad (3.2)$$

and boundary conditions

$$\lambda_i^{(1)} \frac{\partial u}{\partial x_i} = \kappa_i^{(1)}(x_j, t, u)u - g_i^{(1)}(x_j, t, u), \quad x_i = 0, \quad (3.3)$$

$$-\lambda_i^{(2)} \frac{\partial u}{\partial x_i} = \kappa_i^{(2)}(x_j, t, u)u - g_i^{(2)}(x_j, t, u), \quad x_i = L_i, \quad i = 1, 2, \quad j = 3 - i. \quad (3.4)$$

In (3.3) and (3.4) values of $\lambda_i^{(1)}$ and $\lambda_i^{(2)}$ are equal 1 if on the corresponding boundaries of the spatial region G we are given the Neumann or the 3rd kind boundary conditions. They equal zero for the Dirichlet boundary conditions.

Mathematical model (3.1)–(3.4) describes a wide range of phenomena from physics, chemistry, biology and economics. For example, many reaction-diffusion problems, models for heat localisation and flame propagation may be cast in the form (3.1)–(3.4). Solutions of such non-linear evolutionary problems even in the simplest one-dimensional case may form finite time singularities, exhibit blow-up or extinction behaviour. Investigation of such a behaviour in dissipative media is the first step in the simulation of evolution of non-equilibrium open thermodynamical systems [21].

In such a general formulation as (3.1)–(3.4) theoretical issues of existence and uniqueness of the solution lie outside of the scope of this paper. The reader has to consult [21, 5, 8] and references therein in order to get more details on existence/nonexistence of arising problems. For the purposes of this paper we limit ourselves to the situation when the coefficients in (3.1)–(3.4) satisfy the following conditions

$$\kappa_i^1(x_j, t, u) \geq 0, \quad \kappa_i^2(x_j, t, u) \geq 0, \quad 0 < c' \leq k_i(\mathbf{x}, t, u) \leq c'', \quad q^0(\mathbf{x}, t, u) \geq 0. \quad (3.5)$$

The condition (3.5) includes the possibility of fast changing coefficients. Problems with such coefficients arise naturally in many practical situations [10, 6, 9, 16]. Such problems, rarely treatable analytically, cause serious mathematical difficulties and require the application of effective numerical methods. In what follows we describe mathematics of the software package ALTPACK, designed for the solution of a general class of problems (3.1)–(3.4), and present the results of computational experiments.

4 Discrete Conservation Laws and Integro-Interpolational Approach

We cover the region \bar{Q}_T with a non-uniform grid $\hat{\omega}_{h\tau}$ and consider an elementary space-time cell, Δ_{st} , defined as $\Delta_{st} = \Delta_s \times \Delta_t$, where $\Delta_t = \{t_k \leq t \leq t_k + \tau\}$ and

$$\Delta_s = \{(x_1, x_2) : x_{1,i_1-0.5} \leq x_1 \leq x_{1,i_1+0.5}, x_{2,i_2-0.5} \leq x_2 \leq x_{2,i_2+0.5}\}.$$

Let $\mathcal{F}^{(i)}(\mathbf{x}, t, u)$ be the flux in the direction of x_i , i.e.

$$\mathcal{F}^{(i)}(\mathbf{x}, t, u) = k_i(\mathbf{x}, t, u) \frac{\partial u}{\partial x_i}. \quad (4.1)$$

The construction of difference schemes for the solution of problem (3.1)–(3.4) is conducted by the integro-interpolational method [19, 20]. Below we explain the main ideas of such a construction.

Let us integrate differential equation (3.1) over the cell Δ_{st} . We obtain the following balance equation

$$\begin{aligned} & \int_{t_k}^{t_k+\tau} \left\{ \int_{x_2, i_2 - 0.5}^{x_2, i_2 + 0.5} [\mathcal{F}_{i_1+0.5}^{(1)}(x_2) - \mathcal{F}_{i_1-0.5}^{(1)}(x_2)] dx_2 dt + \right. \\ & \quad \left. \int_{x_1, i_1 - 0.5}^{x_1, i_1 + 0.5} [\mathcal{F}_{i_2+0.5}^{(2)}(x_1) - \mathcal{F}_{i_2-0.5}^{(2)}(x_1)] dx_1 dt \right\} - \int_{\Delta_{st}} q^0(\mathbf{x}, t, u) u d\mathbf{x} dt = \\ & \quad \int_{\Delta_{st}} a^0(\mathbf{x}, t, u) \frac{\partial u}{\partial t} d\mathbf{x} dt - \int_{\Delta_{st}} f^0(\mathbf{x}, t, u) d\mathbf{x} dt, \end{aligned} \quad (4.2)$$

where $d\mathbf{x} \equiv dx_1 dx_2$ and

$$\mathcal{F}_{i_1 \pm 0.5}^{(1)}(x_2) = \mathcal{F}^{(1)}(x_{1, i_1 \pm 0.5}, x_2), \quad \mathcal{F}_{i_2 \pm 0.5}^{(2)}(x_1) = \mathcal{F}^{(2)}(x_1, x_{2, i_2 \pm 0.5}).$$

First we divide identity (4.1) by $k_i(\mathbf{x}, t, u)$, $i = 1, 2$ and integrate the result over the intervals $[x_{1, i_1}, x_{1, i_1 + 1}]$ and $[x_{2, i_2}, x_{2, i_2 + 1}]$ respectively. We get

$$u(x_{1, i_1 + 1}, x_2, t) - u(x_{1, i_1}, x_2, t) = \int_{x_{1, i_1}}^{x_{1, i_1 + 1}} \frac{\mathcal{F}^{(1)}(\mathbf{x}, t, u)}{k_1(\mathbf{x}, t, u)} dx_1, \quad (4.3)$$

$$u(x_1, x_{2, i_2 + 1}, t) - u(x_1, x_{2, i_2}, t) = \int_{x_{2, i_2}}^{x_{2, i_2 + 1}} \frac{\mathcal{F}^{(2)}(\mathbf{x}, t, u)}{k_2(\mathbf{x}, t, u)} dx_2. \quad (4.4)$$

Then we use the following interpolation formulas

$$\mathcal{F}^{(1)}(\mathbf{x}, t, u) \approx \mathcal{F}_{i_1+0.5}^{(1)}(x_2, t, u), \quad \text{where } x_1 \in [x_{1, i_1}, x_{1, i_1 + 1}] \quad (4.5)$$

and

$$\mathcal{F}^{(2)}(\mathbf{x}, t, u) \approx \mathcal{F}_{i_2+0.5}^{(2)}(x_1, t, u), \quad \text{where } x_2 \in [x_{2, i_2}, x_{2, i_2 + 1}]. \quad (4.6)$$

Substituting (4.5) and (4.6) into (4.3), (4.4) respectively, we come to the following approximate equalities

$$u(x_{1, i_1 + 1}, x_2, t) - u(x_{1, i_1}, x_2, t) \approx \mathcal{F}_{i_1+0.5}^{(1)}(x_2, t, u) \int_{x_{1, i_1}}^{x_{1, i_1 + 1}} \frac{dx_1}{k_1(\mathbf{x}, t, u)}, \quad (4.7)$$

$$u(x_1, x_{2, i_2 + 1}, t) - u(x_1, x_{2, i_2}, t) \approx \mathcal{F}_{i_2+0.5}^{(2)}(x_1, t, u) \int_{x_{2, i_2}}^{x_{2, i_2 + 1}} \frac{dx_2}{k_2(\mathbf{x}, t, u)}. \quad (4.8)$$

The next step of the procedure is to express the values of $\mathcal{F}_{i_1 \pm 0.5}^{(1)}(x_2, t, u)$ and $\mathcal{F}_{i_2 \pm 0.5}^{(2)}(x_1, t, u)$ through the solution of the problem and known functions. It can be easily done from (4.7), (4.8). Indeed, dividing (4.7) and by h_{1, i_1}^+ and then by h_{1, i_1}^- we have

$$\mathcal{F}_{i_1+0.5}^{(1)}(x_2, t, u) \approx u_{x_1, i_1} \left[\frac{1}{h_{1, i_1}^+} \int_{x_{1, i_1}}^{x_{1, i_1 + 1}} \frac{dx_1}{k_1(\mathbf{x}, t, u)} \right]^{-1}, \quad (4.9)$$

$$\mathcal{F}_{i_1-0.5}^{(1)}(x_2, t, u) \approx u_{x_1, i_1} \left[\frac{1}{h_{1, i_1}^-} \int_{x_{1, i_1 - 1}}^{x_{1, i_1}} \frac{dx_1}{k_1(\mathbf{x}, t, u)} \right]^{-1} \quad (4.10)$$

Analogous expressions can be obtained for $\mathcal{F}_{i_2 \pm 0.5}^{(2)}$ from (4.8).

The integro-difference scheme for the solution of the problem (3.1)–(3.4) follows after substitution (4.9), (4.10) (and analogous expressions for $\mathcal{F}_{i_2 \pm 0.5}^{(2)}$) into (4.2). The application of

interpolational formulas for the remaining functions in the integrals in (4.2) leads to a fully discrete approximation of the model (3.1)–(3.4). Such discrete approximations tend to preserve on the grid conservative properties of the process expressed by (4.2), and are usually referred to as conservative difference schemes [19, 22]. For example, in the linear case of elliptic equations we get [20]

$$\sum_{i=1}^2 (\mathcal{K}_i y_{\bar{x}_i})_{\bar{x}_i} - qy = -f, \quad x \in \omega_h, \quad (4.11)$$

i.e. for each inner node (i_1, i_2) we have the following equation

$$\frac{\mathcal{K}_1^+ y_{x_1} - \mathcal{K}_1^- y_{\bar{x}_1}}{h_{1,i_1}} + \frac{\mathcal{K}_2^+ y_{x_2} - \mathcal{K}_2^- y_{\bar{x}_2}}{h_{2,i_2}} - qy = -f, \quad (4.12)$$

where $y \equiv y(i_1, i_2)$ and the coefficients and the RHS in (4.12) are defined as follows

$$q \equiv q(i_1, i_2) = \frac{1}{h_{1,i_1} h_{2,i_2}} \int_{\Delta_s} q^0(\mathbf{x}) d\mathbf{x}, \quad f \equiv f(i_1, i_2) = \frac{1}{h_{1,i_1} h_{2,i_2}} \int_{\Delta_s} f^0(\mathbf{x}) d\mathbf{x}, \quad (4.13)$$

$$\mathcal{K}_1 \equiv \mathcal{K}_1(i_1, i_2) = \frac{1}{h_{2,i_2}} \int_{x_2, i_2 - 0.5}^{x_2, i_2 + 0.5} \left[\frac{1}{h_{1,i_1}^-} \int_{x_1, i_1 - 1}^{x_1, i_1} \frac{dx_1}{k_1(\mathbf{x})} \right]^{-1} dx_2 \quad (4.14)$$

(with an analogous formula for \mathcal{K}_2), $\mathcal{K}_1^+ \equiv \mathcal{K}_1^+(i_1, i_2) = \mathcal{K}_1(i_1 + 1, i_2)$, $\mathcal{K}_2^+ \equiv \mathcal{K}_2^+(i_1, i_2) = \mathcal{K}_2(i_1, i_2 + 1)$.

When the coefficients in (3.1)–(3.4) are smooth functions, then instead of (4.13), (4.14) we may use simplified expressions for the computation of coefficients and the RHS of (4.11) at each inner node (i_1, i_2) . For example, $q = q^0(x_{1,i_1}, x_{2,i_2})$, $f = f^0(x_{1,i_1}, x_{2,i_2})$, $\mathcal{K}_1 = k_1(x_{1,i_1 - 0.5}, x_{2,i_2})$, $\mathcal{K}_2 = k_2(x_{1,i_1}, x_{2,i_2 - 0.5})$.

Let us consider the approximation of boundary conditions of problem (3.1)–(3.4) when $x_i = L_i$, $i = 1, 2$. We apply a 3-step procedure that consists of

- integrating the original differential equation over the cells

$$\begin{aligned} \Delta_s^{(1)} &= \{(x_1, x_2) : L_1 - 0.5h_{1,N_1}^- \leq x_1 \leq L_1, x_{2,i_2 - 0.5} \leq x_2 \leq x_{2,i_2 + 0.5}\}, \\ \Delta_s^{(2)} &= \{(x_1, x_2) : L_2 - 0.5h_{2,N_2}^- \leq x_2 \leq L_2, x_{1,i_1 - 0.5} \leq x_1 \leq x_{1,i_1 + 0.5}\}, \end{aligned}$$

- using approximate formulas (4.9), (4.10) (and analogous expressions for $\mathcal{F}_{i_2 \pm 0.5}^{(2)}$), and
- computing values of $\mathcal{F}_{N_1}^{(1)}(x_2)$, $\mathcal{F}_{N_2}^{(2)}(x_1)$, from boundary condition (3.4) for $x_i = L_i = x_{i,N_i}$, $i = 1, 2$.

As a result of such a procedure we get

$$\frac{1}{h_i} \left(g_i^{(2)} - \kappa_i^{(2)} y - \mathcal{K}_i y_{\bar{x}_i} \right) + (\mathcal{K}_j y_{\bar{x}_j})_{\bar{x}_j} - qy = -f, \quad x \in \omega_j(L_i), \quad i = 1, 2, \quad j = 3 - i, \quad (4.15)$$

where for the grid prototypes of functions $\kappa_i^{(2)}$ and $g_i^{(2)}$ in (3.4) we used the same notation. In a similar way we construct the approximation of the boundary condition (3.3) for $x_i = 0$:

$$\frac{1}{h_i} \left(\mathcal{K}_i^+ y_{x_i} - \kappa_i^{(1)} y + g_i^{(1)} \right) + (\mathcal{K}_j y_{\bar{x}_j})_{\bar{x}_j} - qy = -f, \quad x \in \omega_j(0). \quad (4.16)$$

where, as above, $\kappa_i^{(1)}$ and $g_i^{(1)}$ are discrete prototypes of functions $\kappa_i^{(1)}$ and $g_i^{(1)}$ from (3.3).

5 Alternating-Triangular Method with Ordered Chebyshev Set of Iterative Parameters

The discrete scheme constructed in the previous section can be written in the following generic form

$$\Lambda y \equiv \sum_{i=1}^2 \Lambda_i y = -\varphi, \quad x \in \hat{\omega}_h, \quad (5.1)$$

where $\varphi = f + \varphi_1 + \varphi_2$, $x \in \hat{\omega}_h$, and for $i = 1, 2$

$$\Lambda_i y = \begin{cases} \frac{2}{h_i^+} \mathcal{K}_i^+ y_{x_i} - \left(\frac{2}{h_i^+} \kappa_i^{(1)} + \frac{1}{2} q \right) y, & x_i = 0, \\ (\mathcal{K}_i y_{\bar{x}_i})_{\bar{x}_i} - \frac{1}{2} q y, & x_i \neq 0, L_i, \\ -\frac{2}{h_i^-} \mathcal{K}_i y_{\bar{x}_i} - \left(\frac{2}{h_i^-} \kappa_i^{(2)} + \frac{1}{2} q \right) y, & x_i = L_i, \quad x_j \in \hat{\omega}_j, \end{cases} \quad \varphi_i = \begin{cases} \frac{2}{h_i^+} g_i^{(1)}, & x_i = 0, \\ 0, & x_i \neq 0, L_i, \\ \frac{2}{h_i^-} g_i^{(2)}, & x_i = L_i, \quad x_j \in \hat{\omega}_j. \end{cases}$$

A special case of our scheme is a rectangular uniform grid in (x_1, x_2) when $h_\alpha^- = h_\alpha^+ = h_\alpha$. In such a case the difference scheme (5.1) with coefficients computed by the formulas (4.13), (4.14) will approximate the problem with error $O(|h|^2)$, where $|h|^2 = h_1^2 + h_2^2$ [20].

The scheme (5.1) can be applied for the solution of linear elliptic equations. In the general case of (3.1)–(3.4) on each time layer we can represent the difference scheme constructed by the described methodology as

$$Ay = \varphi, \quad (5.2)$$

with an operator $A : \mathcal{H} \rightarrow \mathcal{H}$ acting in the Hilbert space \mathcal{H} . More precisely, for the general non-stationary case we have to solve (5.2) on each time layer with an operator A which is a modified version of the operator Λ used in (5.1). In its essence we pass from nonstationary solution(s) of (3.1)–(3.4) to a sequence of stationary solutions for a family of problems that can be formally expressed in the form (1.7). The form of operator \tilde{A} and the function \tilde{f} in (1.7) are now well defined for each element of such a sequence of stationary solutions.

For the solution of problem (5.2) we apply an implicit two-layer scheme of the following general form

$$B(y_{k+1} - y_k)/\tau_{k+1} + Ay_k = \varphi, \quad k = 0, 1, 2, \dots, \forall y_0 \in \mathcal{H}. \quad (5.3)$$

Majority of known iterative procedures (such as the Zeidel method or upper-relaxation algorithm) are just partial cases of the scheme (5.3). They immediately follow from (5.3) if an appropriate choice of sequence $\{\tau_{k+1}\}$ is made and operator B is constructed. In order to make the method effective one has to choose the set of $\{\tau_{k+1}\}$ in some optimal way whereas B has to be constructed as a product of easily invertible operators. In this paper we use the alternating-triangular method with the optimal Chebyshev choice of parameters [7, 20]. We choose the operator B in the form

$$B = (D + \omega_0 A_1) D^{-1} (D + \omega_0 A_2), \quad \omega_0 > 0, \quad (5.4)$$

where

$$Dy = d(x)y, \quad d(x) > 0, \quad x \in \hat{\omega}_h, \quad A = A_1 + A_2, \quad A_1^* = A_2. \quad (5.5)$$

We assume that inequalities

$$\delta D \leq A, \quad A_1 D^{-1} A_2 \leq \frac{\Delta}{4} A, \quad \delta > 0, \quad \Delta > 0 \quad (5.6)$$

are satisfied with certain constants δ and Δ . Then we define parameters τ_{k+1} and ω_0 in the following way

$$\omega_0 = 2\sqrt{\delta\Delta}, \quad \tau_k = \frac{\tau_0}{1 + \rho_0\sigma_k^*}, \quad (5.7)$$

where

$$\tau_0 = \frac{2}{\gamma_1^* + \gamma_2^*}, \quad \gamma_1^* = \frac{\delta}{2(1 + \sqrt{\eta})}, \quad \gamma_2^* = \frac{\delta}{4\sqrt{\eta}}, \quad \rho = \frac{1 - \psi}{1 + \psi}, \quad \psi = \frac{\gamma_1^*}{\gamma_2^*}, \quad \eta = \frac{\delta}{\Delta},$$

and $\sigma_k \in \Sigma_{n_0}^*$, $k = 1, \dots, n_0$. The set

$$\Sigma_{n_0}^* = \left\{ \cos \frac{2i-1}{2n_0}\pi, \quad 1 \leq i \leq n_0 \right\} \quad (5.8)$$

is a specially ordered set of the roots of Chebyshev polynomial degree n_0 . For more details on the ordering procedure the reader may consult [19, 20].

Remark 5.1 In order to keep a trade-off between the accuracy of the approximate solution of (3.1)–(3.4) and the computational complexity of the procedure, we use an estimate of the number of ATM iterations for the given accuracy. If the error of the ATM has to be decreased by the factor of ε it is sufficient to perform n_0 iterations, where

$$n_0 \geq \frac{\ln(2/\varepsilon)}{2\sqrt{2}\sqrt[4]{\eta}}.$$

In the next section we define the operators A_i , $i = 1, 2$ in the case of linear elliptic problems and show how the choice of δ and Δ can be made in the general case.

6 The Choice of Accelerating Parameters δ and Δ

In order to compute the iterative parameter ω_0 in (5.7) we follow the procedure described in [19, 20] and proved to be effective in a number of applications (see, for example, [10, 6]).

Using the first Green's formula [19] for the special case of (5.1) where $\lambda_i^{(1)} = \lambda_i^{(2)} = 1$, we have

$$(Ay, y) = \sum_{i=1}^2 \left((\mathcal{K}_i y_{x_i}^2, 1)_i + (\kappa_i^{(1)} y^2, 1)_{\omega_i(0)} + (\kappa_i^{(2)} y^2, 1)_{\hat{\omega}_j(L_i)} \right) + (qy^2, 1), \quad j = 3 - i. \quad (6.1)$$

We define the operator A in (5.2) as the sum of operators A_1 and A_2 , where

$$A_i = A_i^{(1)} + A_i^{(2)}, \quad i = 1, 2,$$

and

$$A_1^{(i)} y = \begin{cases} \frac{1}{2} \left[\frac{2}{h_i^+} \left(\frac{\kappa_i^+}{h_i^+} + \kappa_i^{(1)} \right) + \frac{1}{2} q \right] y, & x_i = 0, \\ \frac{\kappa_i}{h_i} y_{x_i} + \frac{1}{2} \left[\frac{1}{h_i} \left(\frac{\kappa_i^+}{h_i^+} - \frac{\kappa_i}{h_i^-} \right) + \frac{1}{2} q \right] y, & x_i \neq 0, L_i, \\ \frac{2}{h_i^-} \kappa_i y_{x_i} + \frac{1}{2} \left[\frac{2}{h_i^-} \left(\kappa_i^{(2)} - \frac{\kappa_i}{h_i^-} \right) + \frac{1}{2} q \right] y, & x_i = L_i, \quad x_j \in \hat{\omega}_j \end{cases}$$

$$A_2^{(i)} y = \begin{cases} -\frac{2}{h_i^+} \kappa_i y_{x_i} + \frac{1}{2} \left[\frac{2}{h_i^+} \left(\kappa_i^{(1)} - \frac{\kappa_i^+}{h_i^+} \right) + \frac{1}{2} q \right] y, & x_i = 0, \\ -\frac{\kappa_i^+}{h_i} y_{x_i} + \frac{1}{2} \left[\frac{1}{h_i} \left(\frac{\kappa_i}{h_i^-} - \frac{\kappa_i^+}{h_i^+} \right) + \frac{1}{2} q \right] y, & x_i \neq 0, L_i, \\ \frac{1}{2} \left[\frac{2}{h_i^-} \left(\frac{\kappa_i}{h_i^-} + \kappa_i^{(2)} \right) + \frac{1}{2} q \right] y, & x_i = L_i, x_j \in \hat{\omega}_j. \end{cases}$$

It can be verified that $A_1^* = A_2$ and $A = A_1 + A_2$. Therefore

$$(A_1 D^{-1} A_2 y, y) = (D^{-1} A_2 y, A_2 y) = (D^{-1} (A_2 y)^2, 1) = \left(D^{-1} \left[\sum_{i=1}^2 A_2^{(i)} y \right]^2, 1 \right).$$

Assuming that the function $d(\mathbf{x})$ (see (5.5)) has the following form

$$d(\mathbf{x}) = d_1(\mathbf{x}) + d_2(\mathbf{x}), \quad \mathbf{x} \in \hat{\omega}_h,$$

where

$$d_i(\mathbf{x}) = \begin{cases} \left(2\mathcal{K}_i^+ + h_i^+ \Omega_i \left| \kappa_i^{(1)} - \frac{\mathcal{K}_i^+}{h_i^+} + \frac{h_i^+}{4} q \right| \right) \frac{\Theta_i}{(h_i^+)^2}, & x_i = 0, \\ \left(\mathcal{K}_i^+ + \frac{h_i^+ \Omega_i}{2} \left| \frac{\mathcal{K}_i^-}{h_i^-} - \frac{\mathcal{K}_i^+}{h_i^+} + \frac{h_i^-}{2} q \right| \right) \frac{\Theta_i}{h_i^+ h_i^-}, & x_i \neq 0, L_i, \\ \left| \kappa_i^{(2)} + \frac{\mathcal{K}_i^-}{h_i^-} + \frac{h_i^-}{4} q \right| \frac{\Omega_i \Theta_i}{h_i^-}, & x_i = L_i, \quad x_j \in \hat{\omega}_j, \end{cases}$$

and $\Omega_i = \Omega_i(x_j)$, $\Theta_i = \Theta_i(x_j)$ are certain grid functions that are positive on $\hat{\omega}_h$ (precise definitions are given below). Then applying ϵ -inequality [19, 14] it is straightforward to get

$$(A_1 D^{-1} A_2 y, y) \leq \sum_{i=1}^2 \left[\left(\frac{\mathcal{K}_i}{\Theta_i} y_{x_i}^2, 1 \right)_i + \left(\frac{\nu_2^{(i)}}{\Omega_i \Theta_i} y^2, 1 \right)_i \right], \quad (6.2)$$

where

$$\nu_2^{(i)} = \begin{cases} \frac{1}{h_i^+} \left| \kappa_i^{(1)} - \frac{\mathcal{K}_i^+}{h_i^+} + \frac{h_i^+}{4} q \right|, & x_i = 0, \\ \frac{1}{2h_i} \left| \frac{\mathcal{K}_i^-}{h_i^-} - \frac{\mathcal{K}_i^+}{h_i^+} + \frac{h_i^-}{2} q \right|, & x_i \neq 0, L_i, \\ \frac{1}{h_i^-} |\kappa_i^{(2)} + \frac{\mathcal{K}_i^-}{h_i^-} + \frac{h_i^-}{4} q|, & x_i = L_i, \quad x_j \in \hat{\omega}_j. \end{cases}$$

Hence

$$(Dy, y) = (dy^2, 1) = \sum_{i=1}^2 \left[(\Theta_i \nu_1^{(i)} y^2, 1) + (\Omega_i \Theta_i \nu_2^{(i)} y^2, 1) \right] \quad (6.3)$$

with

$$\nu_1^{(i)} = \begin{cases} \frac{\mathcal{K}_i^+}{h_i^+ \Theta_i}, & x_i \neq L_i, \\ 0, & x_i = L_i, \quad x_j \in \hat{\omega}_j. \end{cases}$$

We need the following lemma.

Lemma 6.1 [20] Let $z(\mathbf{x})$ be the solution of the problem $\Lambda_i z = -\rho$, $x_i \in \hat{\omega}_i$ and $\Psi = 1/\max_{x \in \hat{\omega}_i(x_j)} z(\mathbf{x})$. Then

$$\Psi(\rho y, y)_{\hat{\omega}_i(x_j)} \leq (\mathcal{K}_i y_{x_i}^2, 1)_{\hat{\omega}_i^+(x_j)} + \frac{1}{2} (qy, y)_{\hat{\omega}_i(x_j)} \kappa_i^{(1)} y^2(0, x_j) + \kappa_i^{(2)} y^2(L_i, x_j), \quad \forall x_j \in \hat{\omega}_j.$$

Let

$$\zeta_i^{(1)} = \zeta_i^{(1)}(x_j) = \max_{x_i \in \hat{\omega}_i(x_j)} y_1^{(i)}(x), \quad \zeta_i^{(2)} = \zeta_i^{(2)}(x_j) = \max_{x_i \in \hat{\omega}_i(x_j)} y_2^{(i)}(x), \quad (6.4)$$

where for fixed values of $x_j \in \hat{\omega}_j$ the grid functions $y_1^{(i)}(x)$ and $y_2^{(i)}(x)$ are the solutions of three-point boundary problems

$$\Lambda_i y_1^{(i)} = -\nu_1^{(i)} \quad \text{and} \quad \Lambda_i y_2^{(i)} = -\nu_2^{(i)}$$

respectively.

Since the grid functions Ω_i and Θ_i do not depend on x_i , we get the following inequalities

$$\begin{aligned} (\Omega_i \Theta_i \nu_2^{(i)} y^2, 1) &\leq (\zeta_i^{(2)} \Omega_i \Theta_i \kappa_i y_{x_i}, 1)_i + \frac{1}{2} (\zeta_i^{(2)} \Omega_i \Theta_i q y^2, 1) + \\ &(\zeta_i^{(2)} \Omega_i \Theta_i \kappa_i^{(1)} y^2, 1)_{\hat{\omega}_j(0)} + (\zeta_i^{(2)} \Omega_i \Theta_i \kappa_i^{(2)} y^2, 1)_{\hat{\omega}_j(L_i)} \end{aligned} \quad (6.5)$$

$$\begin{aligned} \left(\frac{\nu_2^{(i)}}{\Omega_i \Theta_i} y^2, 1 \right) &\leq \left(\frac{\zeta_i^{(2)} \kappa_i y_{x_i}^2}{\Omega_i \Theta_i}, 1 \right)_i + \frac{1}{2} \left(\frac{\zeta_i^{(2)} q}{\Omega_i \Theta_i} y^2, 1 \right) + \\ &\left(\frac{\zeta_i^{(2)} \kappa_i^{(1)}}{\Omega_i \Theta_i} y^2, 1 \right)_{\hat{\omega}_j(0)} + \left(\frac{\zeta_i^{(2)} \kappa_i^{(2)}}{\Omega_i \Theta_i} y^2, 1 \right)_{\hat{\omega}_j(L_i)}, \end{aligned} \quad (6.6)$$

$$\begin{aligned} (\Theta_i \nu_1^{(i)} y^2, 1) &\leq (\zeta_i^{(1)} \Theta_i \kappa_i y_{x_i}^2, 1)_i + \frac{1}{2} (\zeta_i^{(1)} \Theta_i q y^2, 1) + \\ &(\zeta_i^{(1)} \Theta_i \kappa_i^{(1)} y^2, 1)_{\hat{\omega}_j(0)} + (\zeta_i^{(1)} \Theta_i \kappa_i^{(2)} y^2, 1)_{\hat{\omega}_j(L_i)}. \end{aligned} \quad (6.7)$$

If we choose Θ_i in the form

$$\Theta_i(x_j) = [\zeta_i^{(2)}(x_j) \Omega_i(x_j) + \zeta_i^{(1)}(x_j)]^{-1}, \quad (6.8)$$

and take into account (6.1), then substituting (6.5) and (6.7) into (6.3) we confirm that

$$(Dy, y) \leq (Ay, y),$$

and hence in inequalities (5.6) we can set

$$\delta = 1. \quad (6.9)$$

Now we shall find the expression for Δ . Substituting (6.6) into (6.2) and taking into account (6.8) we get the following estimate

$$\begin{aligned} (A_1 D^{-1} A_2 y, y) &\leq \sum_{i=1}^2 \left[\left(\left(\frac{\zeta_i^{(2)}}{\Omega_i} + 1 \right) (\zeta_i^{(2)} \Omega_i + \zeta_i^{(1)}) \kappa_i y_{x_i}^2, 1 \right)_i + \frac{1}{2} \left(\frac{\zeta_i^{(2)}}{\Omega_i} (\zeta_i^{(2)} \Omega_i + \zeta_i^{(1)}) q y^2, 1 \right) \right. \\ &+ \left. \left(\frac{\zeta_i^{(2)}}{\Omega_i} (\zeta_i^{(2)} \Omega_i + \zeta_i^{(1)}) \kappa_i^{(1)} y^2, 1 \right)_{\hat{\omega}_j(0)} + \left(\frac{\zeta_i^{(2)}}{\Omega_i} (\zeta_i^{(2)} \Omega_i + \zeta_i^{(1)}) \kappa_i^{(2)} y^2, 1 \right)_{\hat{\omega}_j(L_i)} \right]. \end{aligned} \quad (6.10)$$

Now we choose the optimal Ω_i as the solution of the following minimisation problem

$$\left(\frac{\zeta_i^{(2)}}{\Omega_i} + 1 \right) (\zeta_i^{(2)} \Omega_i + \zeta_i^{(1)}) \rightarrow \min,$$

i.e.

$$\Omega_i(x_j) = \sqrt{\zeta_i^{(1)}(x_j)}. \quad (6.11)$$

Hence from (6.10) we have

$$\begin{aligned} (A_1 D^{-1} A_2 y, y) &\leq \sum_{i=1}^2 \left[\left((\zeta_i^{(2)} + \sqrt{\zeta_i^{(1)}})^2 \kappa_i y_{x_i}^2, 1 \right)_i + \frac{1}{2} \left(\zeta_i^{(2)} (\zeta_i^{(2)} + \sqrt{\zeta_i^{(1)}}) q y^2, 1 \right) + \right. \\ &\left. \left(\zeta_i^{(2)} (\zeta_i^{(2)} + \sqrt{\zeta_i^{(1)}}) \kappa_i^{(1)} y^2, 1 \right)_{\hat{\omega}_j(0)} + \left(\zeta_i^{(2)} (\zeta_i^{(2)} + \sqrt{\zeta_i^{(1)}}) \kappa_i^{(2)} y^2, 1 \right)_{\hat{\omega}_j(L_i)} \right]. \end{aligned}$$

This estimate together with (6.1) confirm that in inequalities (5.6) we may set

$$\Delta = 4 \max_{i=1,2} \left\{ \max_{x_j \in \hat{\omega}_j} \left[\zeta_i^{(2)}(x_j) + \sqrt{\zeta_i^{(1)}(x_j)} \right]^2 \right\}. \quad (6.12)$$

The above estimates obtained for the mixed boundary conditions of the 3rd kind. Only slight modifications are necessary in order to get analogous estimates for Dirichlet or Neumann boundary value problems.

7 Algorithmic Aspects of the Method

The algorithmic procedure for the alternating-triangular method can be derived from (5.3) with the definition of operators B and A given in Sections 5 and 6. We have

- $y_{k+1}(i_1, i_2) = y_k(i_1, i_2) - \tau_{k+1} v_k(i_1, i_2)$, $i_1 = 0, \dots, N_1$; $i_2 = 0, \dots, N_2$;
- $v_k(i_1, i_2) = \beta_1(i_1, i_2) v_k(i_1 + 1, i_2) + \beta_2(i_1, i_2) v_k(i_1, i_2 + 1) + s(i_1, i_2) d(i_1, i_2) u_k(i_1, i_2)$,
 $i_1 = N_1 - 1, \dots, 0$, $i_2 = N_2 - 1, \dots, 0$;
- $u_k(i_1, i_2) = \alpha_1(i_1, i_2) u_k(i_1 - 1, i_2) + \alpha_2(i_1, i_2) u_k(i_1, i_2 - 1) + s(i_1, i_2) r_k(i_1, i_2)$,
 $i_1 = 1, \dots, N_1$, $i_2 = 1, \dots, N_2$;
- $r_k(i_1, i_2) = [r_k^{(1)}(i_1, i_2) + r_k^{(2)}(i_1, i_2)] - \varphi(i_1, i_2)$, $i_1 = 0, \dots, N_1$, $i_2 = 0, \dots, N_2$,

where $\alpha_i = \omega_0 s p_2^{(i)}$, $\beta_i = \omega_0 s \nu_1^{(i)}$, $i = 1, 2$,

$$r_k^{(i)} = p_1^{(i)} y_k - \nu_1^{(i)} y_k^{(+1_i)} - p_2^{(i)} y_k^{(-1_i)}, s = \frac{1}{d + \omega_0(p_1^{(1)} + p_1^{(2)})/2},$$

and

$$p_1^{(i)} = \begin{cases} \frac{2}{h_i^+} \left(\frac{\kappa_i^+}{h_i^+} + \kappa_i^{(1)} \right) + \frac{1}{2}q, & x_i = 0, \\ \frac{1}{h_i^-} \left(\frac{\kappa_i^+}{h_i^+} + \frac{\kappa_i^-}{h_i^-} \right) + \frac{1}{2}q, & x_i \neq 0, L_i, \\ \frac{2}{h_i^-} \left(\frac{\kappa_i^-}{h_i^-} + \kappa_i^{(2)} \right) + \frac{1}{2}q, & x_i = L_i, x_j \in \hat{\omega}_j, \end{cases} \quad p_2^{(i)} = \begin{cases} 0, & x_i = 0, \\ \frac{\kappa_i}{h_i h_i^-}, & x_i = L_i, x_j \in \hat{\omega}_j. \end{cases}$$

Therefore on each time layer the procedure goes as follows

Algorithm 7.1.

- compute the residual

$$r_k = Ay_k - \varphi; \quad (7.1)$$

- find the auxiliary function u_k from the solution of the system of equations

$$(D + \omega_0 A_1) u_k = r_k; \quad (7.2)$$

- find the auxiliary function v_k from the system of equations

$$(D + \omega_0 A_2) v_k; \quad (7.3)$$

- compute a new approximation

$$y_{k+1} = y_k - \tau_{k+1} v_k; \quad (7.4)$$

- if $\max_{\hat{\omega}_h} |y_{k+1} - y_k| > \varepsilon$, then set $y_k = y_{k+1}$ and repeat the procedure.

We terminate the algorithm if either in all nodes of the grid $\hat{\omega}_h$ the inequality

$$|y_{k+1} - y_k| \leq \epsilon$$

holds, or n_0 iterations (see Remark 5.1) have been performed. As an initial approximation we may choose any function y_0 from the space \mathcal{H} .

If the solution is found we may compute fluxes from (4.1) using the following difference derivatives of $\partial u / \partial x_i$

- for inner nodes:

$$\frac{\partial u}{\partial x_i} \approx \frac{h_i^+(y - y^{(-1)})/h_i^- + h_i^-(y^{(+1)} - y)/h_i^+}{2h_i},$$

- for the left boundary node:

$$\frac{\partial u}{\partial x_i} \approx \frac{4u^{(+1)} - 3u - u^{(+2)}}{2h_i^+},$$

- for the right boundary node:

$$\frac{\partial u}{\partial x_i} \approx \frac{3u - 4u^{(-1)} + u^{(-2)}}{2h_i^-}.$$

Then the flux through the boundary Γ is defined as follows [10]

$$\mathcal{F}_i|_{x_i=0} = \sum_{\hat{\omega}_j} \mathcal{F}^{(i)} h_j, \quad \mathcal{F}_i|_{x_i=L_i} = - \sum_{\hat{\omega}_j} \mathcal{F}^{(i)} h_j. \quad (7.5)$$

The structure of the software package is given on Figure 1. It consists of the *MAIN* program, the control program *ALTPACK* and five programs that are called from *ALTPACK* in the sequence from left to right as shown on Figure 1. Details on these programs are given in the Appendix.

8 Numerical Experiments.

The method described in the previous sections was applied to the solution of a number of problems that present interest from both theoretical and practical points of views.

- **Example 1.** We start from a linear problem in which coefficients of the equation are fast-changing functions. Such problems are typical in modelling of semiconductor devices, chemical kinetics, problems of porous media and in many other applications. Consider problem (3.1)–(3.4) in the spatial unit-square region with the Dirichlet boundary conditions ($\lambda_i^{(i)} = 0$, $\kappa_i^{(i)} = 1$, $i = 1, 2$),

$$\begin{aligned} g_1^{(1)} &= \sin(4x_2), \quad g_1^{(2)} = \sin(4x_2 + 4) + t^2 x_2, \\ g_2^{(1)} &= \sin(4x_1), \quad g_2^{(2)} = \sin(4x_1 + 4) + t^2 x_1. \end{aligned}$$

and the initial condition given in the form

$$u(x_1, x_2, 0) = \sin 4(x_1 + x_2).$$

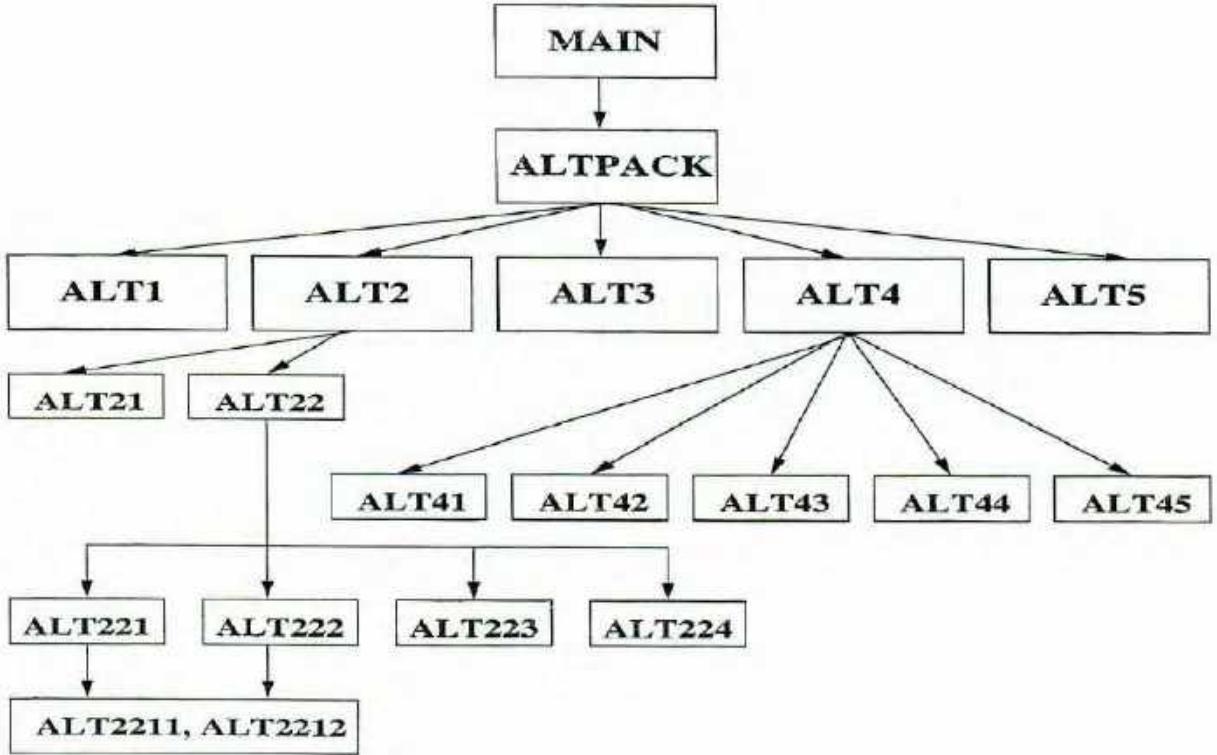


Figure 1:

Let coefficients of the equation and its right-hand side be defined as follows

$$k_1 = 1 + cx_1(1 + x_2), \quad k_2 = 1 + cx_2(1 + x_1), \quad q^0 = 0,$$

where c is a constant defined below, and

$$f^0 = 16 \sin 4(x_1 + x_2) [cx_1(1 + x_2) + cx_2(1 + x_1) + 2] - 4c \cos 4(x_1 + x_2) [x_1 + x_2 + 2] - ct^2 [x_2(1 + x_2) + x_1(1 + x_1)] + 2tx_1x_2.$$

From the choice of the coefficients $k_i(x)$ follows that

$$c' = \min_x k_i(x) = 1 \quad \text{and} \quad c'' = \max_x k_i(x) = 1 + 2c.$$

We investigated the dependency of the number of ATM iterations on the ratio $\mathcal{C} = c''/c'$. In this example we used a uniform grid with $N_1 = N_2 = 80$. The results are given in the table

below.

\mathcal{C}	Δ (DB)	$n_0(\epsilon = 1.0 \cdot 10^{-5})$	$n_0(\epsilon = 1.0 \cdot 10^{-7})$
2^0	138.922	15	21
2^1	197.437	17	23
2^2	309.474	19	25
2^3	557.979	21	29
2^4	1017.03	25	34
2^5	1836.22	29	39
2^6	3191.87	33	45
2^7	5131.17	37	51
2^8	7358.03	40	56

It is clear that the number of iterations for the ATM, n_0 , weakly dependent on the ratio $\mathcal{C} = c''/c'$ for the values of ϵ within the range of practical applications. We computed the solution for $\mathcal{C} = 32$. It can be verified that the function

$$u(x_1, x_2, t) = \sin 4(x_1 + x_2) + t^2 x_1 x_2 \quad (8.1)$$

provides the analytical solution to the above problem. Our results do not deviate from the exact solution (8.1) for more than $\sim 10^{-4}$. On Figure 2 we present the solution for $t = 0.01$, $t = 1.2$ and $t = 5$. We see that the “upper tongue” of the solution grows steadily that indicates an increasing contribution of the second term from (8.1). Finally this term becomes dominant that is confirmed by the lower left plot. The typical structure of the solution error is shown on the lower right plot of Figure 2.

The situation when coefficients of the differential equation change locally very rapidly is typical in many applications such as semiconductor device modelling and chemical kinetics [15]. Although spectra boundaries may not change drastically, large values of the ratio $\mathcal{C} = c''/c'$ cause serious mathematical difficulties for the solution of such problems. The proposed method allows us to overcome some such difficulties due to a weak dependency of its quality on the increase of such a ratio.

- **Example 2.** In this example we consider a power-type of coefficient nonlinearities in (3.1)

$$\frac{\partial u}{\partial t} = \sum_{i=1}^2 \frac{\partial}{\partial x_i} \left(k_i(u) \frac{\partial u}{\partial x_i} \right), \quad (8.2)$$

with $k_1(u) = k_2(u) = u^\sigma$. Equation (8.2) is a two-dimensional analogue of self-similar S -regime considered in [21]. We choose the initial condition in the “peaking” form

$$u(x_1, x_2, 0) = 0.5 T_f^{-1/\sigma} \left(1 - \sum_{i=1}^2 \frac{\xi_i x_i}{\Xi} \right)_+^{2/\sigma}, \quad (8.3)$$

where

$$(z)_+ = \max(z, 0), \quad \Xi = \sqrt{2(0.5)^\sigma \frac{\sigma+2}{\sigma}}.$$

For $x_1 = 0$ and $x_2 = 0$ the boundary conditions were defined from the following majorant

$$u(x_1, x_2, t) = 0.5 (T_f - t)^{-1/\sigma} \left(1 - \sum_{i=1}^2 \frac{\xi_i x_i}{\Xi} \right)_+^{2/\sigma}. \quad (8.4)$$

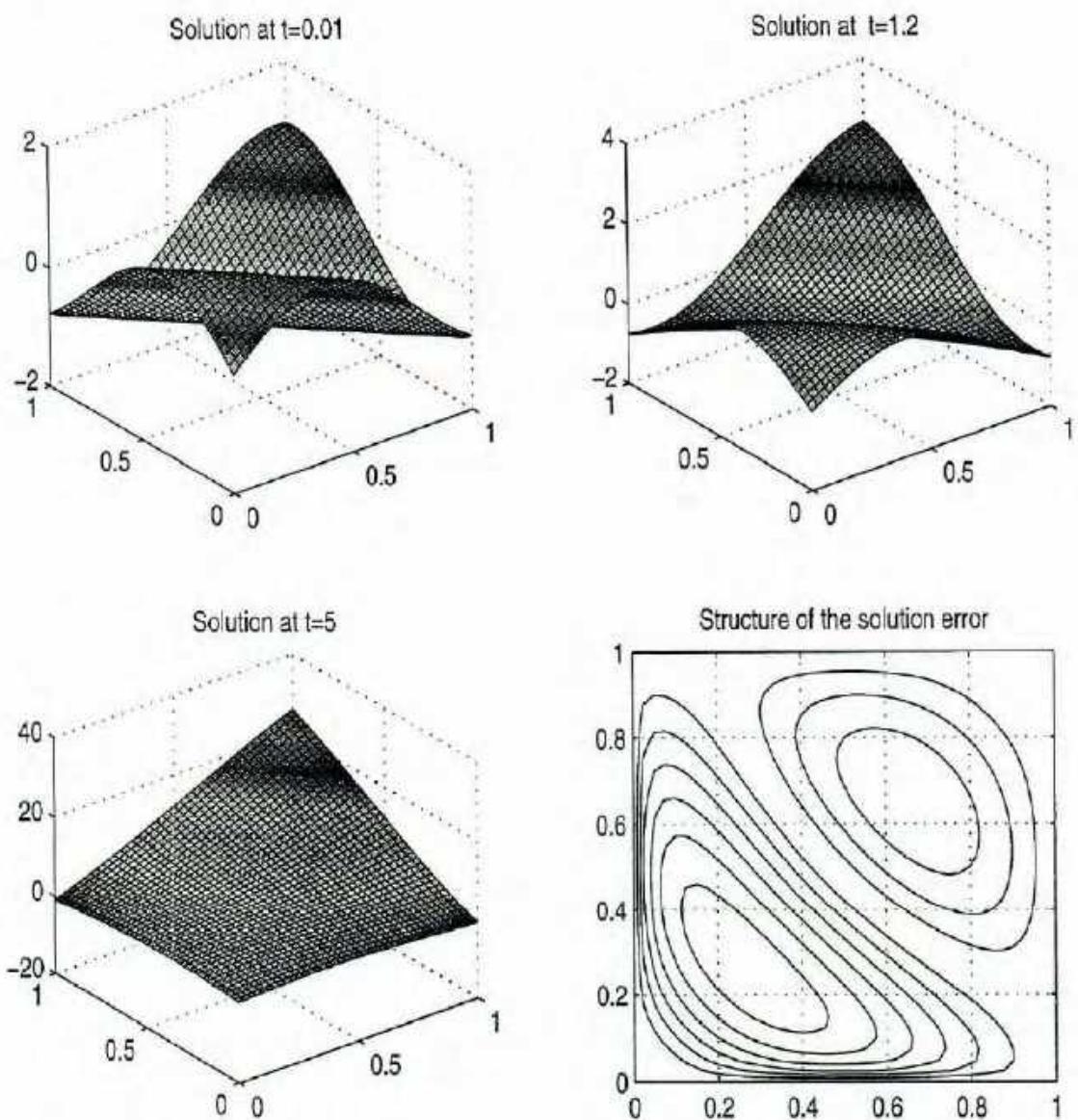


Figure 2:

Namely, the computation was conducted for

$$u(x_1, 0, t) = 0.5(T_f - t)^n \left(1 - \frac{\xi_1 x_1}{\Xi}\right)_+^{2/\sigma}, \quad (8.5)$$

$$u(0, x_2, t) = \frac{1}{2}(T_f - t)^n \left(1 - \frac{\xi_2 x_2}{\Xi}\right)_+^{2/\sigma} \quad (8.6)$$

with the constants $T_f = 1$, $\Xi = 1$, $\xi_1 = \xi_2 = 1/\sqrt{2}$. On other two sides of the region boundary zero-flux conditions were assumed.

We note (see details in [21]) that the majorant function (8.4) for boundary conditions (8.5) and (8.6) defines the boundary temperature on the axes x_1 and x_2 as zero whenever $x_1 \geq \Xi/\xi_1$, $x_2 \geq \Xi/\xi_2$. The temperature increases with peaking when $0 \leq x_1 \leq \Xi/\xi_1$, $0 \leq x_2 \leq \Xi/\xi_2$.

First we consider the case when $n < -1/\sigma$ (HS-regime according to the classification given in [21]). In our experiment we have chosen $n = -1$ and $\sigma = 2$. From the initial moment of time to a certain finite time qualitative picture of the process remains the same. We observe a two-dimensional standing thermal wave shown on the left upper plot of Figure 3. The thermal energy of such a wave is in a finite region of heat localisation defined by a triangle with vertices $(0, 0)$, $(\sqrt{2}, 0)$, and $(0, \sqrt{2})$. In other parts of the region temperature remains zero during a finite period of time. When such time elapses, the profile of the standing wave starts to change gradually, indicating the movement of the thermal wave. The interface between the region of heat “localisation” and the region with zero-temperature distorts (see upper right plot), and we finally observe the solution blow-up in finite time when $t \rightarrow T_f^-$. On lower plots of Figure 3 the solution and its contour lines in $(x_1 x_2)$ plane are presented for the moment $t = 0.99$.

In order to obtain this solution on the mesh 41×41 it was necessary to conduct 62 ATM iteration. The computation required 10 non-linear iterations and Δ was equal 11497.5.

For the case $n > -1/\sigma$ (LS-regime according to the classification given in [21]) the qualitative behaviour of the solution remains similar to the situation described above. However, the change of the temperature profile and deterioration of the “localisation” region begins essentially later compared to the case $n < -1/\sigma$. The results are presented on Figure 4. For $n = -0.25$ and $\sigma = 2$ it was necessary to perform 36 ATM iterations and the same number of nonlinear iterations in order to obtain the solution on the mesh 41×41 . In this case Δ was equal 1280.65.

- **Example 3.** In this example we investigate the behavior of “thermal crystals” produced by thermal energy localised in certain space regions [21]. In the region Q_T we consider equation (8.2) with the initial condition (8.3). As above, the initial temperature distribution is localised in a region

$$\alpha_1|x_1| + \alpha_2|x_2| \leq \Xi, \text{ where } \alpha_1^2 + \alpha_2^2 = 1. \quad (8.7)$$

However now we assume zero-flux conditions on the boundary of the region G , i.e. in (3.3), (3.4) we set

$$\lambda_i^{(i)} = 1, \quad \kappa_i^{(i)} = 0, \quad g_i^{(i)} = 0, \quad i = 1, 2. \quad (8.8)$$

Figure 5 demonstrates some results of our investigations. Initially thermal energy is localised in the square region $[-\sqrt{2}, \sqrt{2}] \times [-\sqrt{2}, \sqrt{2}]$ (upper left plot). When time progresses the temperature profile becomes more and more convex (upper right plot). When the thermal wave starts moving, the region of heat localisation is also transformed acquiring oval outlines

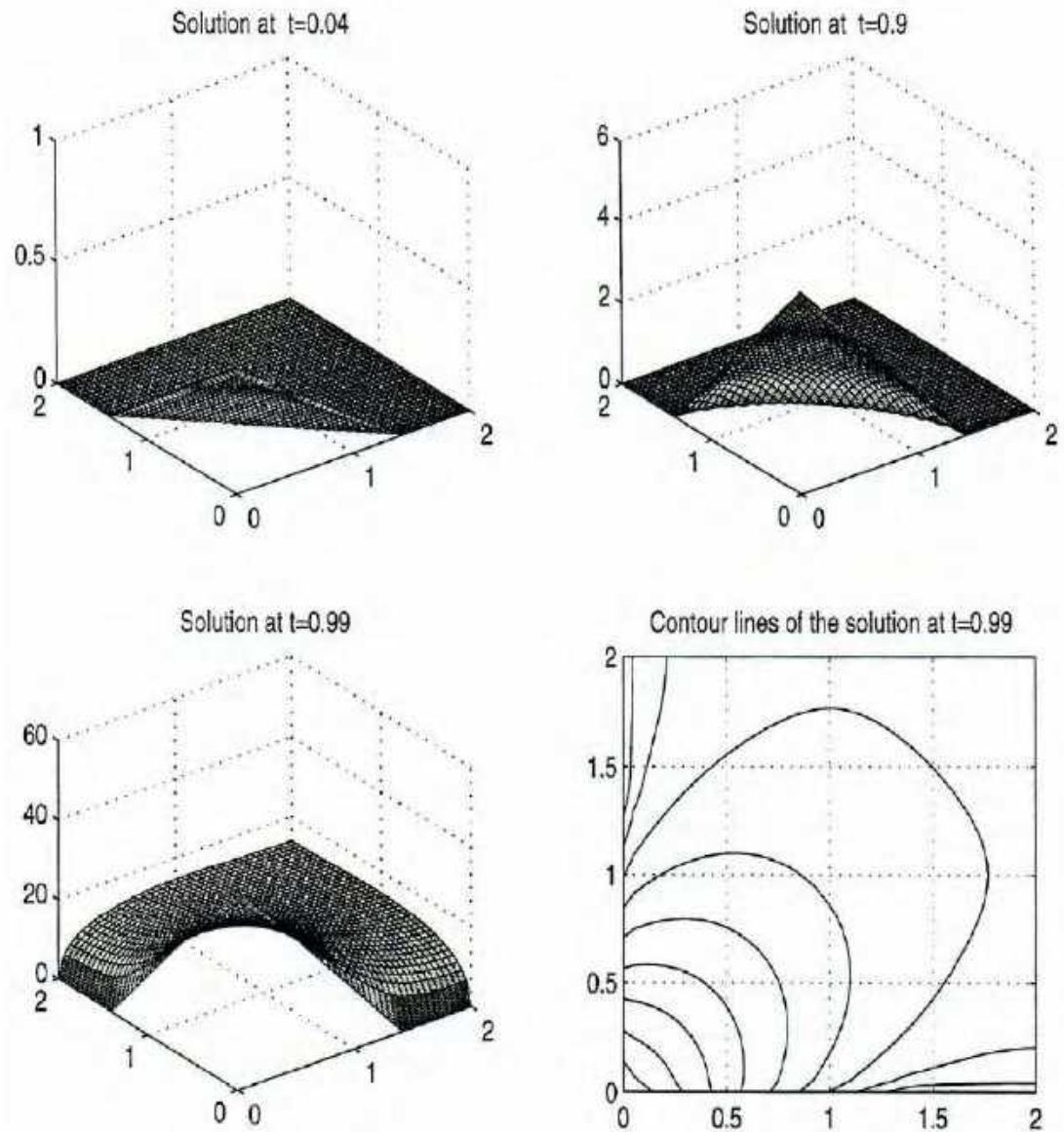


Figure 3:

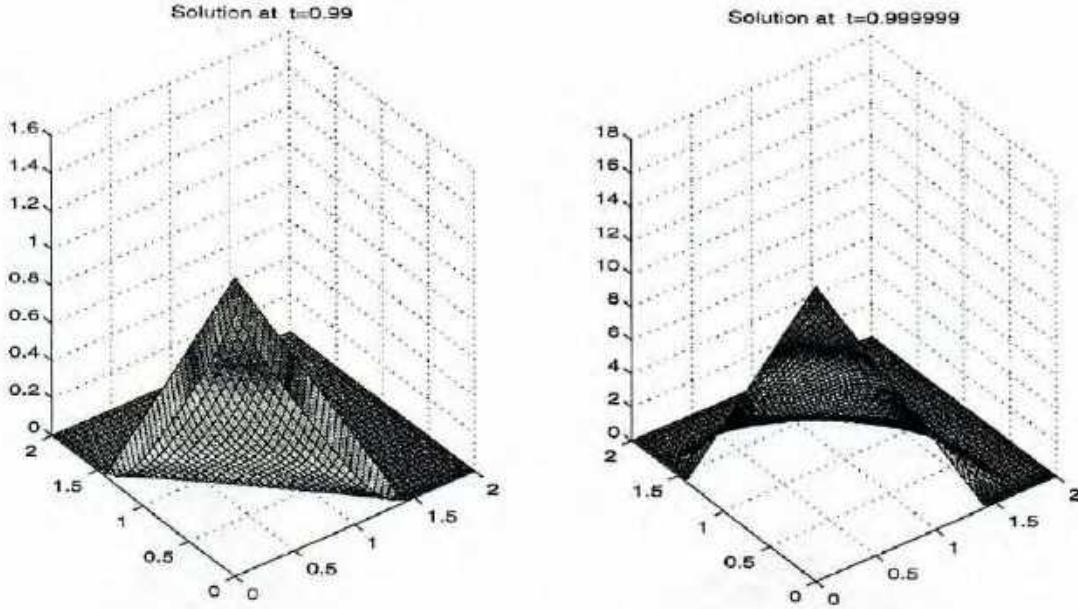


Figure 4:

(see lower right plot). The time of existence of such thermal crystals may be estimated from below [21]

$$t_{loc} \geq \frac{\sigma \Xi^2}{2(\sigma + 2)(0.5)^\sigma}. \quad (8.9)$$

For $\sigma = 2$ and $\Xi = 1$ the estimate (8.9) gives $t_{loc} \geq 1$. When $t = 1$ we needed 9 ATM iterations. In this case Δ equal 3.39866.

- **Example 4.** In the concluding example we consider the equation

$$\frac{\partial u}{\partial t} = \Delta u^m \pm u^p. \quad (8.10)$$

First we consider problem (8.10) with a positive source where $m = 2$, $p = 2$. The problem is supplemented by the initial condition

$$u(x_1, x_2, 0) = (x_1 - 1)^2 + (x_2 - 1)^2. \quad (8.11)$$

Boundary conditions for (8.10) coincide with the boundary conditions of Example 2.

Figure 6 presents the results on the evolution of the thermal-wave solution which exhibits the blow-up behaviour when $t \rightarrow T_f^-$. We note that the computational complexity is essentially increases when $t \rightarrow T_f^-$. For example, to compute the solution on the mesh 41×41 for $t = 0.01$ we needed only 14 ATM iterations ($\Delta = 27.9721$). When $t = 0.97$ the number of ATM iterations increases to 70 ($\Delta = 18856.5$) and the process requires 25 nonlinear iterations.

We also considered the problem (8.10) with absorption where $m = 2$, $p = 0.5$. The initial condition was defined by (8.11) and no-flux boundary conditions were assumed.

Figure 7 demonstrates the solution quenching in finite time. The theoretical investigation of such a solution and the conditions on the existence of its non-trivial continuation after the onset of extinction are given in [5].

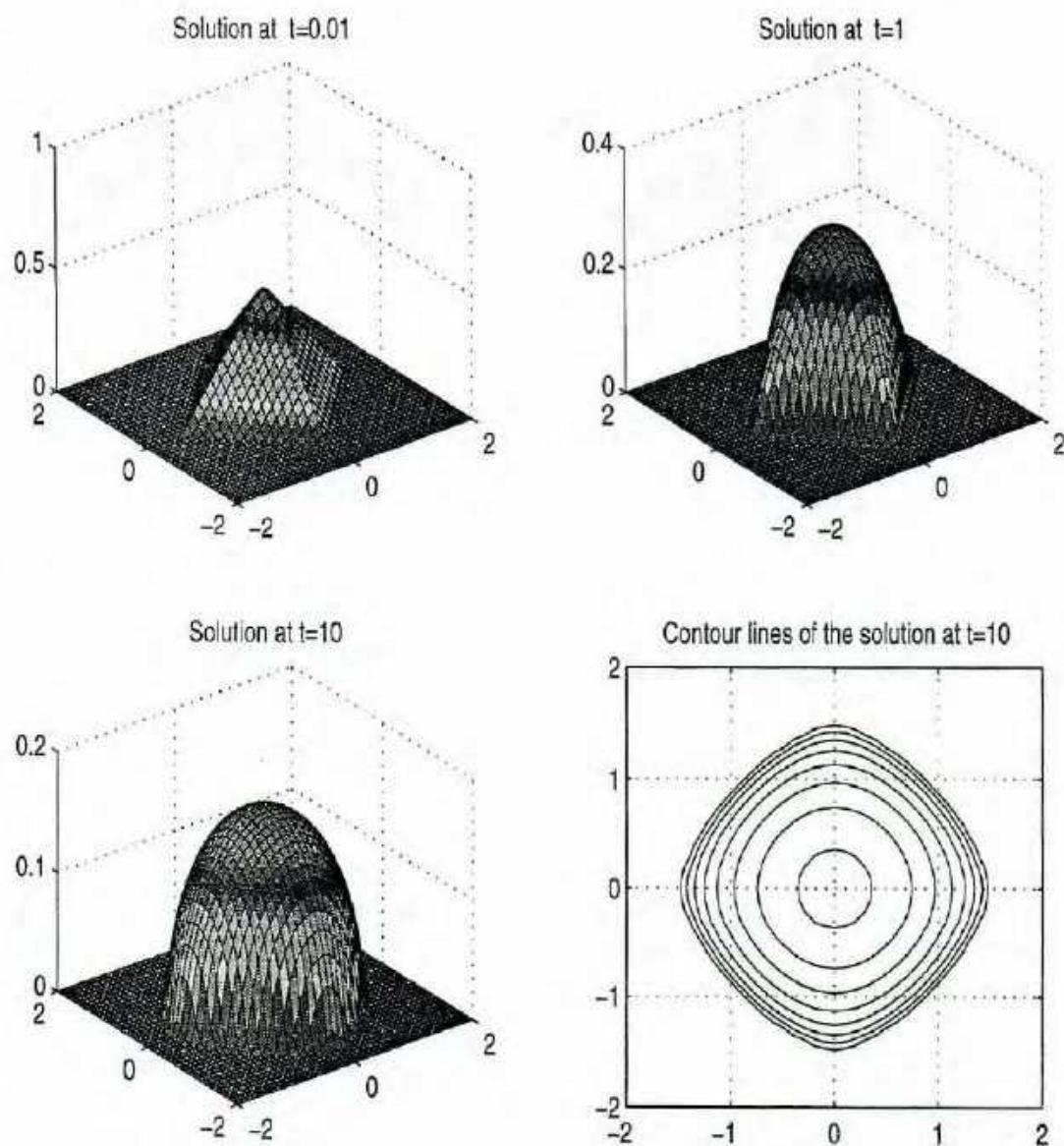


Figure 5:

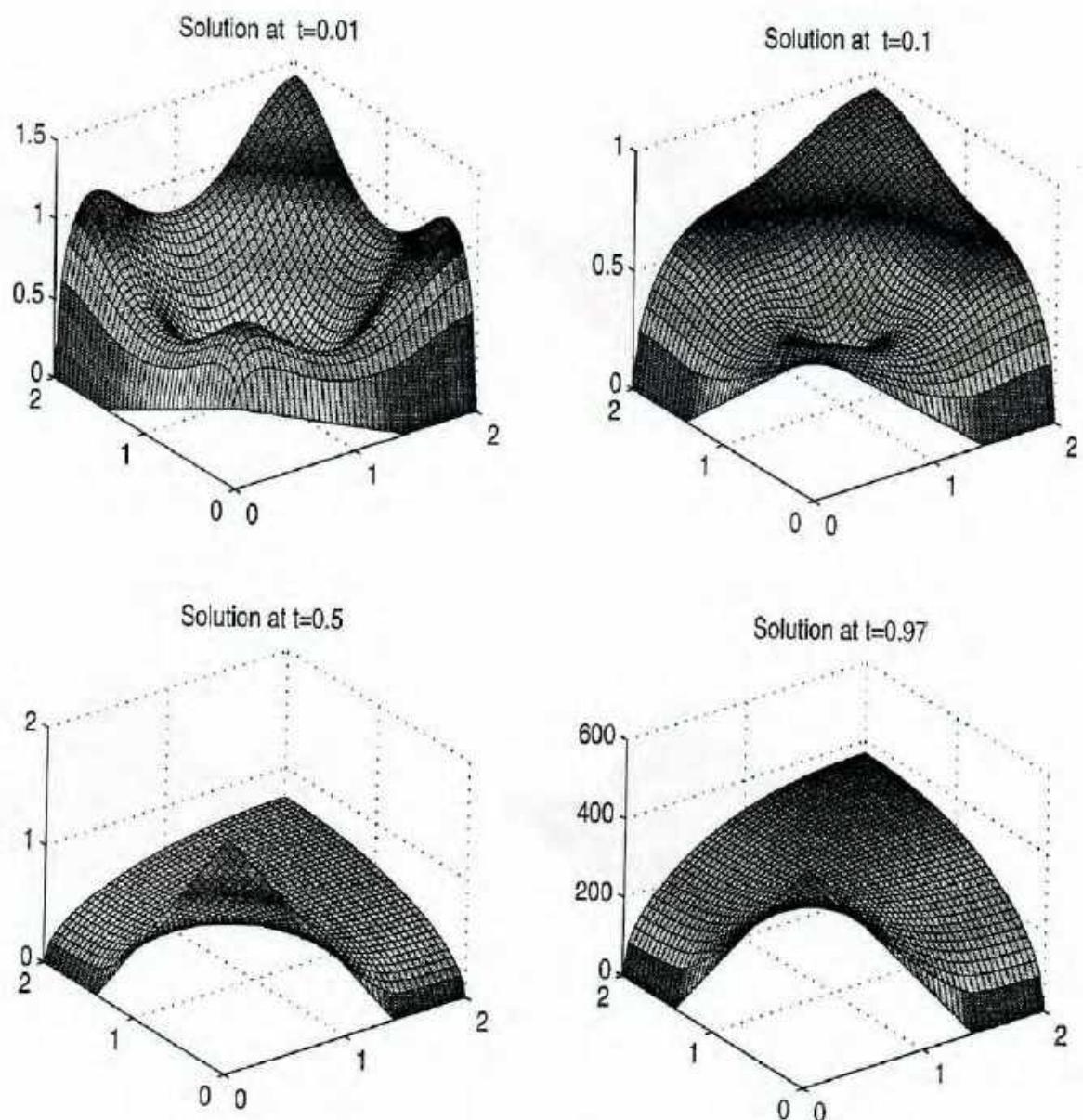


Figure 6:

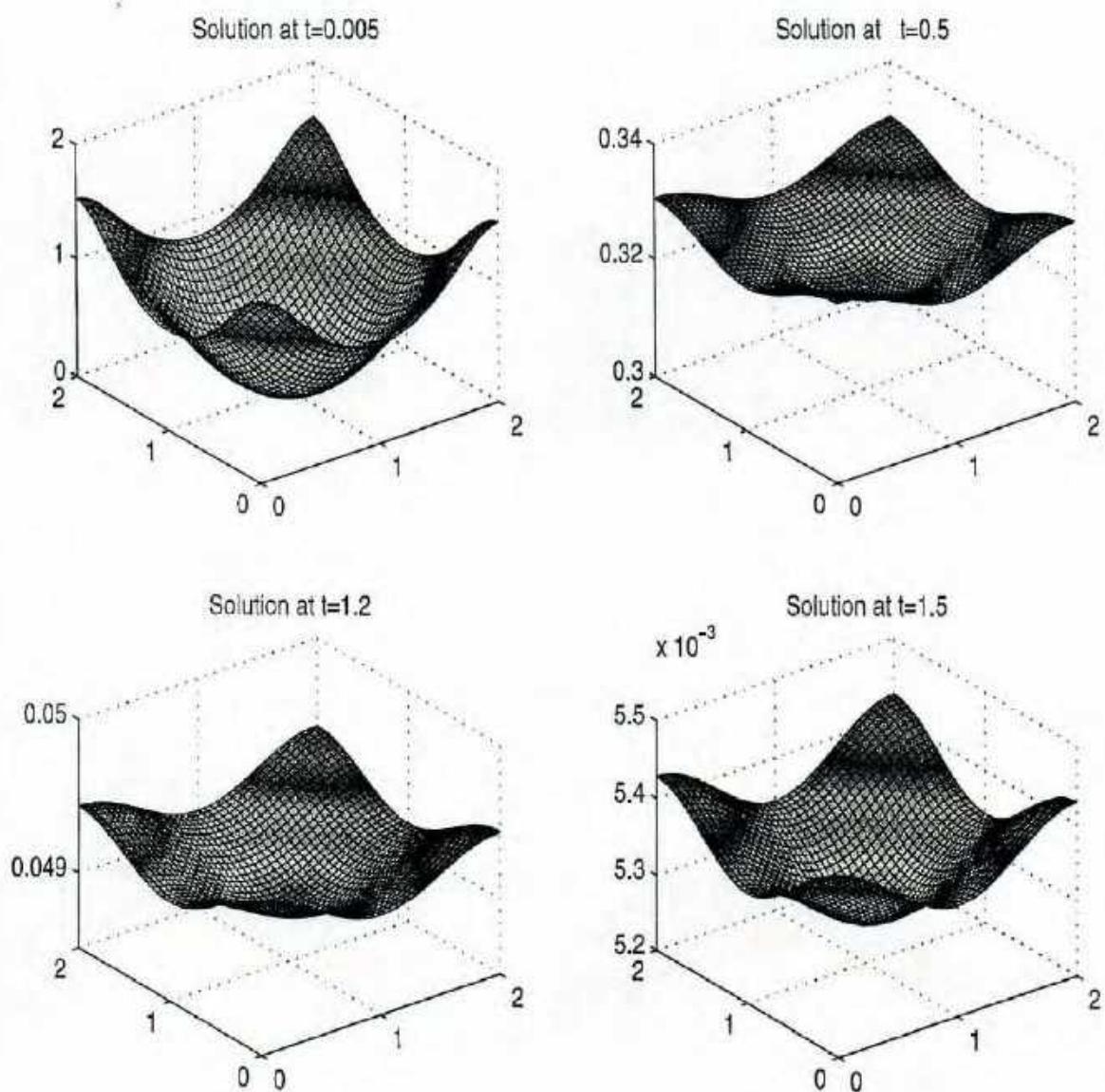


Figure 7:

9 Conclusion

In this paper we described an effective procedure for the solution of nonlinear (and linear) parabolic and elliptic equations in the two-dimensional case. We showed the effectiveness of the method for the problems with fast-changing coefficients as well as for the problems with different types of nonlinearities. Blow-up and quenching phenomena were investigated using the software package ALTPACK that implements the alternating-triangular method.

In the program ALTPACK we applied the procedure of ordering for Chebyshev iterative parameters. This allowed us to minimise the influence of round-off errors and to eliminate large intermediate values (that are dependent on the number of iterations $n(\epsilon)$) in the computational procedure. For a number of non-linear problems ALTPACK shows better results than the classical implicit methods of alternating directions [12, 18]. It is also known that for model problems [19, 20] both methods, ATM and ADI, require the same number of iterations but on average the ATM requires less arithmetic operations. Finally the ATM is easily implemented for parallel and vector computation.

Acknowledgements

The work was supported by grant USQ-PTRP 179389. The author is grateful to Ross Darnell and Harry Butler for their helpful assistance at the final stage of preparation of this paper.

References

- [1] Bai, C., Lavin, A.S., On hyperbolic heat conduction and the second law of thermodynamics, *ASME J. of Heat Transfer*, 117, 1995, 256–263.
- [2] Bebernes, J., Eberly, D., *Mathematical Problems from Combustion Theory*, Springer-Verlag, 1989.
- [3] Fleming, W. and Soner, H., *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, 1993.
- [4] Fujita, H., and Suzuki, T. Evolution problems, in "Handbook of Numerical Analysis", Ed. P.G. Ciarlet & J.L Lions, North-Holland, Amsterdam, New York, Oxford, Tokyo, 1991, Vol. II, 789–923.
- [5] Galaktionov, V.A., Vazquez, J.L., Continuation of Blowup Solutions of Nonlinear Heat Equations in Several Space Dimensions, *Comm. on Pure and Appl. Math.*, Vol. L, 1997, 1–67.
- [6] Goncharenko, V.M., Lychman, V.V., Numerical Solution of a Boundary-Value Problem of Elasticity Theory for a Body with an Inclusion, *J. Math. Sci.*, 72, No. 3, 1994, 3116–3119.
- [7] Hageman, L.A., Young, D.M., *Applied Iterative Methods*, Academic Press, 1981.
- [8] Levine, H., The role of critical exponents in blowup theorems, *SIAM Review*, Vol. 32, No. 2, 1990, 262–288.
- [9] Logan, J.D., *An Introduction to Nonlinear Partial Differential Equations*, Wiley & Sons, 1994.
- [10] Lychman, V.V., Mistetskij, G.E., Numerical Solution of Subsurface Moisture Transport Problems, *Journal of S. Math.*, 54, No. 2, 1991, 824–831.
- [11] Makarenko, A.S., Mathematical Modeling of Heat-Distribution Processes using Generalized Equations for Thermal Conductivity, *J. of Math. Sciences*, Vol. 70, No. 1, 1994, 1529–1533.
- [12] Melnik, R.V.N., Correction for Nonstationarity and Internal Nonlinearity in the Analysis of Integrated-Circuit Thermal Parameters, *Radioelectronics & Communications Systems (Alerton Press)*, Vol. 34, No. 6, 1991, 84–86.
- [13] Melnik, R.V.N., Discrete Models of Coupled Dynamic Thermoelasticity for Stress-Temperature Formulations, submitted to *Mathematical Problems in Engineering: TMA* (Preprint SC-MC-9703, Dept. of Math. & Comp., University of Southern Queensland).
- [14] Melnik, R.V.N., The Stability Condition and Energy Estimate for Nonstationary Problems of Coupled Electroelasticity, *Mathematics and Mechanics of Solids*, 2, 1997, 153–180.

- [15] Melnik, R.V.N., Melnik, K.N., Modelling of Nonlocal Physical Effects in Semiconductor Plasma Using Quasi-Hydrodynamic Models, *Computational Techniques and Applications: CTAC97*, World Scientific, 1998.
- [16] Morton, K.W., *Numerical Solution of Convection-Diffusion Problems*, Chapman & Hall, 1996.
- [17] Muller, I., Ruggeri, T., *Extended Thermodynamics*, Springer-Verlag, 1993.
- [18] Ortega, J.M., *Introduction to Parallel and Vector Solution of Linear Systems*, Plenum Press, New York, 1988.
- [19] Samarskii, A.A., *Theorie der Differenzenverfahren*, (Series: Mathematik und ihre Anwendungen in Physik und Technik), Aufl. Leipzig: Akademische Verlagsgesellschaft Geest & Portig, 1984.
- [20] Samarskii, A.A., Nikolaev, E.S., *Numerical Methods for Grid Equations*, Basel, Boston: Birkhauser Verlag, 1989.
- [21] Samarskii, A.A. et al. *Blow-up in Quasilinear Parabolic Equations*, Vol. 19, de Gruyter Expositions in Mathematics, de Gruyter, Berlin and Hawthorne, NY, 1995.
- [22] Shashkov, M., Steinberg, S., Conservative Finite-Difference Methods on General Grids, *Boca Raton: CRC Press*, 1995.
- [23] Vladimirov, V.S., *Equations of Mathematical Physics*, Moscow, Mir, 1984.

Appendix

Subroutines	Calling Source	Functions/Description
ALTPACK	MAIN	Solution of time-dependent (or stationary) problems for nonlinear (or linear) PDEs
ALT1	ALTPACK	Preparation of initial information for the region discretisation
ALT2	ALTPACK	Preparation of initial information for the ATM; call ALT21, ALT22
ALT3	ALTPACK	Construction of the optimal ordered set of Chebyshev iterative parameters
ALT4	ALTPACK	Solution of the discrete problem by the ATM; call ALT41, ALT42, ALT43, ALT44, ALT45
ALT5	ALTPACK	Print the results for graphics post-processing
ALT21	ALT2	Computation of grid functions for the discrete problem
ALT22	ALT2	Computation of <i>a priori</i> information for the ATM (call ALT221, ALT222, ALT223, ALT224)
ALT41	ALT4	Formation of the discrete problem (construction of the grid functions in the left hand side of the system)
ALT42	ALT4	Construction of the right hand side of the system of grid equations
ALT43	ALT4	Solution of the linear system of discretised equations
ALT44	ALT4	Verification of the convergence of nonlinear iterations
ALT45	ALT4	Verification of stopping criteria
ALT221	ALT22	Preparation of the information for the sweeping algorithm in the x_1 -direction
ALT222	ALT22	Preparation of the information for the sweeping algorithm in the x_2 -direction
ALT223	ALT22	Computation of operator bounds required as <i>a priori</i> information for the problem
ALT224	ALT22	Computation of grid functions required for the definition of the system of discretised equations (such functions are used in ALT41)
ALT2211	ALT221, ALT222	Solution of discretised equations by the left sweeping algorithm
ALT2212	ALT221, ALT222	Solution of the discretised equations by the right sweeping algorithm



TOOWOOMBA

DISCRETE MODELS OF COUPLED
DYNAMIC THERMOELASTICITY FOR
STRESS-TEMPERATURE
FORMULATIONS

Roderick Melnik

Department of Mathematics & Computing, USQ
Faculty of Sciences Working Paper Series

THE UNIVERSITY OF
S O U T H E R N
Q U E E N S L A N D

Working Papers

FACULTY
OF
SCIENCES

**DISCRETE MODELS OF COUPLED
DYNAMIC THERMOELASTICITY FOR
STRESS-TEMPERATURE
FORMULATIONS**

Roderick Melnik

Department of Mathematics & Computing, USQ

Faculty of Sciences Working Paper Series

SC-MC-9703

October 1997

DISCRETE MODELS OF COUPLED DYNAMIC THERMOELASTICITY FOR STRESS-TEMPERATURE FORMULATIONS*

Roderick V. Nicholas Melnik

Department of Mathematics and Computing,
University of Southern Queensland, QLD 4350, Australia
E-mail: melnik@usq.edu.au

Abstract

In this article, the author studies the properties of discrete approximations for mathematical models of coupled thermoelasticity in the stress-temperature formulation. Since many applied problems deal with steep gradients of thermal fields, the main emphasis is given to the investigation of non-smooth solutions of non-stationary thermoelasticity. Convergence of operator-difference schemes on weak solutions of thermoelasticity is proved, and the dispersion analysis of models is performed. Error estimates and the results of computational experiments are presented.

Key words: hyperbolic-parabolic models, operator-difference schemes for thermoelasticity problems, weak solutions, optimal error control.

1 Mixed Modes in Dynamics Described by Mathematical Models of Coupled Field Theory.

In essence, any mathematical model describes a transformation of different types of energy. The recognition of this fact leads to an integral reformulation of differential models. On the one hand, such a reformulation is a fundamental step in obtaining effective numerical procedures of projective-variational type. On the other hand, such a reformulation allows us to relax the theoretical assumptions about a solution's smoothness typically made for differential problems.

In the final analysis, we do not know *a priori* regularity of the solution when solving a practical problem. Hence it is always necessary to cover a gap between mathematical assumptions on the solution smoothness and the smoothness of the solution in a real problem. Properly organized computational experiments may not only verify the validity of theoretical assumptions and estimates, but also identify new effects and tendencies that may lead to new directions in the development of theory. Of course, a high level of understanding of physical, chemical,

*Submitted to Mathematical Problems in Engineering (Theories, Methods and Applications).

biological processes or even properties of the solution of an abstract evolutionary problem may not be achieved by purely theoretical analysis [32]. Human experience will always play a fundamental role in the validation of theoretical results. The acquisition of new information through such experience leads to the possibility of the inclusion of additional *information* about the process, system or phenomenon into the mathematical model (for example, by an improved physical parameterization or by additional relations between system parameters). In turn, this ultimately leads to a change of solution regularity. The process of model improvement may continue indefinitely, and hence it is important to find a balance between the energetic and informational parts of the model's complexity [35, 20]. Coupling, which is a natural way of reflecting additional information about a process (system, phenomenon) requires the relaxation of traditional regularity assumptions. Otherwise, a-priori estimates of solutions may become meaningless [15, 18].

Non-smoothness of solutions is a typical feature of many important problems in structural mechanics where a structure may be subjected to extreme mechanical and/or thermal loads. A lack of regularity in the solutions of thermoelasticity problems is also typical in many other areas of application. During recent years increasing attention has been paid to the investigation of processes in crystals under sharp impulse heating of the surface, in particular, through the investigation of physical effects in semiconductors that critically influence the characteristics of electronic devices. Thermal perturbations in the crystal result in elastic waves, and the process of their propagation cannot be considered without taking into account the thermal field. The rate of temperature change for such processes may be quite large. Hence, for the investigation of thermal deformations in crystals it is essential to take into account dynamic effects, induced by the motion of particles of solid under a rapid thermal expansion.

The study of weak solutions in thermoelasticity is especially important when the coupling between thermal and mechanical fields is relatively strong and has to be considered in the dynamics. Coupled problems in computational physics and other sciences is a two-way dynamic interaction between physically distinct components [9]. Such components may be mechanical, thermal, electromagnetic, biological in nature, but at least one component of the system as a whole has the *hyperbolic mode*. This leads to mathematical challenges in the investigation of such problems which have to be addressed. Computationally we also have a challenging problem because the states of all components should be considered when integrating over time. It is often inappropriate to approximate the mathematical models in coupled field theory using the arguments of "parabolization" [10]. The latter may lead to a distorted picture in the description of real objects by mathematical models. While for coupled mathematical models L^2 -type estimates may provide an important characterization of unknown solutions, after the "parabolic" approximation such estimates may be completely inadequate for grasping changes in the solution behaviour [18]. Since the time of Fourier the heat equation has played a special role in mathematical physics, but despite of its wide applicability it does not provide an adequate description of thermal processes since it predicts an infinite velocity of heat propagation. This fact led to different formulations of hyperbolic equations for heat conduction starting from the publication of well-known paper by Cattaneo that following on Peshkov's experiments [27].

The key to the nature of heat propagation lies in the understanding of the fact that the thermal field never acts on its own. It is always coupled to some other physical fields, such as elastic and/or electromagnetic fields. The effects of such fields may not be negligible. As a result of the interaction with other fields the dynamics of heat propagation acquires mixed hyperbolic-parabolic modes. The competition between these modes is at the heart of the physical process itself. Such a competition is fundamental for many other mathematical models including the convection-diffusion-reaction equation. One realises that the inclusion of

additional information into such equations may continue indefinitely subject to the specification of a reaction law, time scale of the processes involved, and coupling with other processes. A side effect of the inclusion of additional information into the mathematical model is a change in the solution regularity. As a result of this effect when the smoothness of the solution decreases methods based on the model parabolization may become inappropriate. Hence, ideally we have to have error estimates that automatically react to the solution smoothness. Since the solution smoothness is *a priori* unknown we have to balance between *a priori* and *a posteriori* information [23]. Such a balance reflects the essence of the adaptive property of computational algorithms. In this paper we use the Steklov's operator technique combined with the application of the Bramble-Hilbert lemma [22, 23] in order to get *a priori* estimates depending on the solution's smoothness that may be predicted using *a posteriori* information from computational experiments.

The physical, chemical and biological processes that involve the thermal field are dissipative in their nature. Since mathematical models for such processes can be treated analytically only in exceptional cases, it is important to attain a better correspondence between dispersion properties of discrete and continuous problems. Moreover, it appears that using a dispersion relationship it is possible to obtain stability conditions for the discrete problem.

The remaining part of the paper is organized as follows.

- Section 2 provides the reader with the basic mathematical and numerical models that are investigated in the paper.
- In Section 3 we consider a more general operator-difference scheme, prove its stability and obtain a new estimate for its solution. Convergence results for the classical case follow easily from our consideration.
- Section 4 deals with the case of generalized solutions. Using Steklov's operator technique and the Bramble-Hilbert lemma, we prove the convergence result when the solution of the problem is from the Sobolev class $W_2^2(Q_T)$.
- In Section 5 we perform the dispersion analysis of continuous and discrete models of thermoelasticity and derive stability conditions using the Cayley transform.
- In Section 6 we present results of computational experiments on the investigation of planar nonstationary waves in a thermoelastic layer under instantaneous action of surface forces.
- Future directions of present work are addressed in Section 6.

2 Mathematical Models of Coupled Dynamic Thermoelasticity.

Thermoelasticity is one of the first areas in coupled field theory that attracted the attention of mathematicians. Nevertheless there are still many problems in this field that have to be addressed. At present, the development of theory is driven by at least two interconnected features: the non-local nature of heat propagation and its non-equilibrium character. Difficulties arise from the need to take into account the dispersive nature of thermoelastic wave propagation [3, 4]. One possibility is to use non-Fourier type models with time relaxations for heat fluxes (i.e. to allow time for acceleration of the heat flow in response to an applied temperature gradient). Secondly, if we consider a uniform asymptotic expansion with respect to the inertial constant for linear thermoelasticity, the usual coupled quasi-static approximation of the temperature, displacement, and stress (which sets the inertial constant to zero) is not uniformly asymptotically correct even to the lowest order [7]. This result underlines the importance of the role of energy dissipation in thermoelasticity and confirms that adequate

models cannot ignore dissipative processes. Such models may be discrete and non-local in principle because of the discrete characteristics of the medium itself (for example, in applications involved phonon dispersion in solids, in fracture mechanics, in electromagnetic solids etc). Moreover, there is strong experimental evidence indicating that thermally-induced transformations due to the relaxation of elastic strain energy proceed through a series of discrete steps between metastable equilibrium states [29]. In such cases (for example, when high-rate heat sources such as lasers or microwave devices with short duration and/or high frequency are applied) there is not enough time to reach thermodynamic equilibrium, and hence the transport of heat may be better approximated by a wave propagation process rather than diffusion. Although hyperbolic modes in the dynamics become important (or even dominant), the system of partial differential equations for coupled non-stationary thermoelasticity cannot be treated as strictly hyperbolic (see also [3, 36]). The fact of competition between different modes in dynamics causes major difficulties in the construction and investigation of discrete mathematical models.

Exact solutions of initial-boundary value problems in thermoelasticity, obtainable by analytic technique, are known for a quite narrow class of problems, mainly for problems in uncoupled thermoelasticity. In general, the application of analytic procedures to nonstationary thermoelasticity problems and the necessity of taking coupling phenomenon into consideration lead to serious mathematical difficulties. As a result, the development of numerical methods for coupled nonstationary problems of thermoelasticity is a fruitful area of investigation abounding in new mathematical ideas.

We limit ourselves in this paper to the investigation of a one-dimensional system of coupled thermoelasticity for a homogeneous isotropic body. This system after an appropriate scaling has the following form

$$\begin{cases} \frac{\partial^2 s}{\partial t^2} - \frac{\partial^2 s}{\partial x^2} + \frac{\partial^2 \Theta}{\partial t^2} = f_1(x, t), \\ (1 + \epsilon) \frac{\partial \Theta}{\partial t} - a \frac{\partial^2 \Theta}{\partial x^2} + \epsilon \frac{\partial s}{\partial t} = f_2(x, t). \end{cases} \quad (2.1)$$

Here s denotes stress, Θ denotes temperature, ϵ is the coupling parameter between thermal and mechanical fields, and a is the coefficient of thermal conductivity [28, 16]. The system (2.1) holds in the space-time domain $Q_T = \{(x, t) : 0 < x < 1, 0 < t \leq \bar{T}\}$, thus the problem is considered in $\bar{Q}_T = Q_T \cup \Gamma \cup \bar{\Gamma}$, where $\Gamma = \{(x, t) : x = 0, x = 1; 0 < t \leq \bar{T}\}$, $\bar{\Gamma} = \{(x, t) : 0 \leq x \leq 1, t = 0\}$. The system (2.1) is supplemented by the initial and boundary conditions

$$\Theta(x, 0) = \Theta_0(x), \quad s(x, 0) = s_0(x), \quad \frac{\partial s(x, 0)}{\partial t} = \bar{s}_0(x), \quad \text{for } t = 0, \quad (2.2)$$

$$s(x_i, t) = s_i(t), \quad \frac{\partial \Theta(x_i, t)}{\partial x} = \Theta_i(t), \quad \text{for } x_i \in \Gamma, \quad (2.3)$$

where $i = 0, 1$, $x_0 = 0$, $x_1 = 1$.

For many applied problems the values of the stress are the principal unknowns of practical importance, even though mathematically we solve the problem in displacements. But in considering the problem of thermoelasticity in displacements, we have to perform the operation of differentiation (which is ill-posed numerically). This leads to a reduction of the order accuracy [6]. Hence, numerical methods for coupled problems of dynamic thermoelasticity formulated for stress-temperature are preferred in many situations. However, they have not been adequately explored.

The problem in stresses (2.1) – (2.3) is equivalent to the problem in deformations. The former can be reduced to the latter by the change of variables $r = s + \Theta$

$$\begin{cases} \frac{\partial^2 r}{\partial t^2} = \frac{\partial^2 r}{\partial x^2} - \frac{\partial^2 \Theta}{\partial t^2} = f_1(x, t), \\ \frac{\partial \Theta}{\partial t} + \epsilon \frac{\partial r}{\partial t} - a \frac{\partial^2 \Theta}{\partial x^2} = f_2(x, t). \end{cases} \quad (2.4)$$

with the corresponding initial and boundary conditions. The variable r in the system (2.4) denotes deformations, which at the initial moment of time will be denoted by r_0 . The rate of change of deformations at the initial moment of time will be denoted by \bar{r}_0 (defined more precisely below) and other notation will be as for the problem (2.1)–(2.3). The main reason for the introduction of the model (2.4) lies in the fact that numerical schemes for this model are more easily realizable algorithmically than numerical schemes for the model (2.1).

We note that the above models of thermoelasticity do not belong to any of the classical type of partial differential equations. Such a situation is typical in coupled field theory. We have mixed equations which contain different modes: hyperbolic, parabolic, elliptic. Such mathematical models are important for both mathematical theory and application [23, 8]. In fact, in the case of thermoelasticity we can see that one of the equations of the system contains a hyperbolic type operator, whereas the other contains a parabolic type operator. There is a coupling effect between these equations by the parameter ϵ , which in the case of the formulation in deformations is amplified by non-homogeneous boundary conditions. In the general case coupling effects between thermal and elastic fields may lead to the appearance of boundary layers. Indeed when the material moduli depend both on the temperature and the stretch, their effects can either reinforce or mitigate one another. As a result we may observe the accentuation or annihilation of the boundary layer structure [31]. On the other hand, if we formally set $\epsilon = 0$ one may apply a splitting technique based on the solution of the parabolic equation and substitution the temperature function into the hyperbolic equation. However, in all practical applications the value of ϵ may be small but still positive. Such a coupling between parabolic and hyperbolic modes of thermoelastic waves is an essential prerequisite in an adequate description of underlying processes.

Using the basic principles of the construction of difference schemes (see, for example [33] and references therein), let us approximate the problem in deformations. We introduce the grid $\bar{\omega}_{h\tau} = \bar{\omega}_h \times \bar{\omega}_\tau$ in Q_T where

$$\bar{\omega}_h = \{x_i = ih, h = 1/N, i = 1, \dots, N\},$$

$$\bar{\omega}_\tau = \{t_j = j\tau, \tau = T/M, j = 1, \dots, M\}.$$

Let $(y, \eta) \equiv (y_i^j, \eta_i^j)$ be the difference approximation on this grid for the solution (r, Θ) of the problem in deformations. Then we construct the following difference scheme with respect to (y, η) (without the loss of generality we assume here that $s_i = \Theta_i = 0$, $i = 1, 2$)

$$\begin{cases} y_{it} = \Lambda y - \Lambda \eta + \varphi_1, & (x, t) \in \omega_h \times \omega_\tau, \\ \eta_t = \tilde{\Lambda} \eta^{(\sigma)} - \epsilon y_t + \varphi_2, & (x, t) \in \bar{\omega}_h \times \bar{\omega}_\tau, \end{cases} \quad (2.5)$$

$$y = \eta, x \in \Gamma; y = r_0(x), \eta = \Theta_0(x), y_t = r_1(x), x \in \omega_h, t = 0. \quad (2.6)$$

In the scheme (2.5)–(2.6) we use the following notation:

- φ_1 and φ_2 denote approximations to the functions f_1 and f_2 respectively;

- $r_0(x) = s_0(x) + \Theta_0(x)$;
- $r_1(x) = \bar{r}_0(x) + \tau[r_0'' + \Theta_0'' + f_1(x, 0)]/2$;
- $\bar{r}_0(x) = [f_2(x, 0) + \bar{s}_0(x) + a\Theta_0'']/(1 + \epsilon)$.

The operator Λ denotes the second difference derivatives in space, that is $\Lambda y = y_{xx}$, and

$$\bar{\Lambda}\eta = \begin{cases} \Lambda\eta, & \text{when } x \in \omega_h, \\ 2\eta_x/h, & \text{when } x = 0, \\ -2\eta_x/h, & \text{when } x = 1. \end{cases}$$

We use the difference scheme with weights σ ($0 \leq \sigma \leq 1$) in order to achieve the second order of approximation in space and time (in the general case we have $\Lambda y^{(\sigma)} = \sigma\Lambda\hat{y} + (1 - \sigma)\Lambda y$, where the hat denotes the value of y taken from the upper time-level). The approximations to f_1 and f_2 are chosen by analogous reasoning, for example, $\varphi_1 = f_1(x_i, t_j)$, $\varphi_2 = f_2(x_i, t_{j+0.5})$ (another way to approximate the functions f_1 and f_2 will be given in Section 4).

In addition to the consistency of difference scheme (2.5), (2.6) we need a stability result for our discrete approximation to ensure that the error of such an approximation is small. However, due to non-homogeneity of the Dirichlet boundary conditions for difference approximations of deformations (which are computed using approximate values of temperature!) analysis of the stability of difference scheme (2.5)–(2.6) is hampered. To overcome this difficulty we propose to return to the formulation in stresses for the discrete rather than for the continuous problem. This idea is intrinsic in the analysis of stability and in obtaining a new a-priori estimate in the next section. The accuracy of a-posteriori error estimates will depend on the accuracy of the recovered stresses which are computed directly from a difference scheme that we propose in the next section. The implicit balance between a-priori and a-posteriori estimates provide a foundation for effective adaptive computational procedures [15].

3 Stable Operator-Difference Scheme for Problems of Thermoelasticity.

So far we put into correspondence to the continuous mathematical model of thermoelasticity its discrete analogue. Now let us introduce a discrete function $v = y - \eta$ that gives an approximation to the function of stresses. Then the system of difference equations (2.5)–(2.6) may be rewritten in terms of (v, η) as follows

$$\begin{cases} v_t = \Lambda v - \eta_t + \varphi_1, \\ (1 + \epsilon)\eta_t = \bar{\Lambda}\eta^{(\sigma)} - \epsilon v_t + \varphi_2, \end{cases} \quad (3.1)$$

$$v = 0, \quad x \in \Gamma; \quad v = s_0(x), \quad v_t = s_1(x), \quad \eta = \Theta_0(x), \quad t = 0, \quad (3.2)$$

where

$$s_1(x) = (1 - \epsilon)r_1(x) + \bar{\Lambda}\eta^{(\sigma)} \Big|_{t=0} + \varphi_2(x, 0).$$

We introduce two sets of discrete functions

$$H_1 = \{v(x) : x \in \bar{\omega}_h\}, \quad H_2 = \{\eta(x) : x \in \bar{\omega}_h\}$$

with the scalar product

$$(y, z) = \sum_{x \in \bar{\omega}_h} h y(x) z(x)$$

($y \in H_i$, and $z \in H_i$, for $i = 1, 2$ respectively), where

$$\hbar = \begin{cases} h, & \text{when } x \in \omega_h, \\ h/2, & \text{when } x \in \bar{\omega}_h/\omega_h. \end{cases}$$

Instead of the difference scheme (3.1)–(3.2) we shall investigate a more general operator-difference scheme. All results for the scheme (3.1)–(3.2) will follow as a special case of our construction.

Let us define operators from

$$H_1^0 = v(x) : x \in \bar{\omega}_h; v = 0 \text{ at } x = 0, 1$$

into H_1 as follows

$$D_1 = I - \frac{\sigma_1 + \sigma_2}{2} \tau^2 \Lambda, \quad B_1 = -(\sigma_1 - \sigma_2) \tau \Lambda, \quad A_1 = -\Lambda, \quad C_1 = I, \quad (3.3)$$

and the operators from H_2 into H_2 by the formulas

$$B_2 = (1 + \epsilon)I + (1 - \sigma_3)\tau \bar{\Lambda}, \quad A_2 = -\bar{\Lambda}, \quad C_2 = \epsilon I, \quad (3.4)$$

where σ_i , $i = 1, 2, 3$ are weights taking values between zero and one, and I is the identity operator. Let us further assume that

$$\sigma_3 \geq \frac{1 + \alpha}{2} - \frac{h^2(1 - \beta)}{4a\tau}, \quad \frac{\sigma_1 + \sigma_2}{2} \geq \frac{1 + \gamma}{4} - \frac{h^2}{4\tau^2}, \quad (3.5)$$

where

$$\alpha, \gamma > 0, \quad 0 < \beta < 1, \quad 0 \leq \sigma_i \leq 1, \quad i = 1, 2, 3.$$

We note that if the conditions (3.5) are satisfied then the operators defined by the formulas (3.3), (3.4) exist, they are positive definite and selfadjoint. Now we are in the position to consider the following operator-difference generalization of the scheme (3.1)–(3.2)

$$\begin{cases} D_1 v_{tt} + B_1 v_t + A_1 v + C_1 \eta_t = \varphi_1, \\ B_2 \eta_t + A_2 \hat{\eta} + C_2 v_t = \varphi_2, \end{cases} \quad (3.6)$$

where v_t denotes the central difference derivative (that is $v_t = (v_i^{j+1} - v_i^{j-1})/(2\tau)$). For norms of our discrete functions we introduce the following notation

$$\|y(t)\|^2 = (y(t), y(t)), \quad \|y(t)\|_{A_1}^2 = (A_1 y(t), y(t)),$$

$$\|y(t)\|_{(1)}^2 = \|y(t)\|^2, \quad \|y(t)\|_{(2)}^2 = \left\| \sum_{t'=0}^t A_2 y(t') \right\|^2.$$

We formulate the main result of this section as follows

Theorem 3.1. *The operator-difference scheme (3.6), (3.2) is stable with respect to initial data and right-hand side if the conditions (3.5) are satisfied. For the discrete approximation of the problem (2.1)–(2.3) by the scheme (3.6), (3.2) the following estimate*

$$\begin{aligned} \|v(t_1)\|_{(1)} + \|\eta(t_1)\|_{(2)} &\leq M \{ \|v(0)\|_{D_1} + \|A_2 v_t(0)\| + \|A_2 \eta(0)\| + \\ &\quad \sum_{t'=0}^T \tau \|A_2(\kappa_1 + \kappa_2)\| + \sum_{t'=0}^T \tau \|\xi_1 + \xi_2\| + \sum_{t'=\tau}^T \tau \|(\kappa_1 + \kappa_2)_{t'}\| \}, \end{aligned} \quad (3.7)$$

holds, where

$$\bar{D}_1 = (B_2 + \frac{\tau}{2} A_2) D_1 - C_2, \quad \varphi_1 = (\xi_1)_t + (\xi_2)_x, \quad \varphi_2 = (\kappa_1)_t + (\kappa_2)_x.$$

Proof. First, we apply the operator $\bar{B}_2 = B_2 + \tau A_2 / 2$ to both parts of the first equation of the system (3.6). Then, using the expression for the second difference derivative $\eta_{\bar{t}}$ from the second equation of the system (3.6), and substituting its value into the resulting first equation, we have

$$\bar{D}_1 v_{\bar{t}} + \bar{B}_2 B_1 v_0 + \bar{B}_2 A_1 v - A_2 \eta_0 = \bar{B}_2 \varphi_1 - (\varphi_2)_{\bar{t}}, \quad (3.8)$$

where $\bar{D}_1 = \bar{B}_2 D_1 - C_2$.

Now we apply

- the operator C_2 to both parts of the equation (3.8), and
- the operator A_2 to the following consequence of the second equation in (3.6)

$$B_2 \eta_0 + A_2(\hat{\eta} + \eta)/2 + C_2 v_0 = \varphi_2 + \check{\varphi}_2,$$

where the above check denotes values taken from the previous time layer. As a result we obtain the following system of operator-difference equations

$$\begin{cases} C_2 \bar{D}_1 v_{\bar{t}} + C_2 \bar{B}_2 B_1 v_0 + C_2 \bar{B}_2 A_1 v - C_2 A_2 \eta_0 = \\ C_2 \bar{B}_2 \varphi_1 - C_2 (\varphi_2)_{\bar{t}}, \\ A_2 B_2 \eta_0 + (A_2)^2 (\hat{\eta} + \eta)/2 + A_2 C_2 v_0 = A_2 (\varphi_2 + \check{\varphi}_2). \end{cases} \quad (3.9)$$

We note that the operators A_2 and C_2 are commutative. Then we perform the following two operations:

- we multiply both parts of the first equation in equation (3.9) by the function of the discrete argument $t \in \omega_\tau$

$$w(t) = \sum_{t'=\bar{t}+\tau}^{t_1} \tau [v(t') + v(t' - \tau)],$$

which has the following properties [26]

$$w_{\bar{t}} = -(v + \check{v})/2, \quad w(t) = 0, \quad t \geq t_1,$$

and

- multiply both parts of the second equation in equation (3.9) by the function of the discrete argument $t \in \omega_\tau$

$$\zeta(t) = \sum_{t'=\bar{t}+\tau}^{t_1} \tau [\eta(t') + \eta(t' - \tau)].$$

Then the resulting equalities are added together, the result is multiplied by 2τ , and is summed up in t from τ to t_1 . Using a technique developed for hyperbolic equations (see [26]) we can get an energy identity. Assuming that

$$\varphi_1 = (\xi_1)_t + (\xi_2)_x, \quad \varphi_2 = (\kappa_1)_t + (\kappa_2)_x,$$

we use the conditions (3.5) and apply Cauchy-Schwarz inequality and the discrete analogue of the Gronwall lemma [11] to ensure the validity of the estimate (3.7). ■

Convergence results under classical assumptions on solution smoothness follow easily from Theorem 3.1. For example, from the analysis of approximation error using the Taylor formula and the a-priori estimate (3.7) we can readily come to the following conclusion.

Corollary 3.1. *If conditions (3.5) are satisfied, then the solution (v, η) of the scheme (3.6), (3.2) converges to the solution of the problem (2.1)–(2.3) $(s, \Theta) \in C^{4,4}(\bar{Q}_T) \times C^{4,3}(\bar{Q}_T)$, and the error of the scheme is characterized by the following estimate*

$$\|v^j - s^j\|_{(1)} \leq M_1 \Phi_1(h, \tau), \quad \|\eta^j - \Theta^j\|_{(1)} \leq M_2 \Phi_1(h, \tau),$$

where j is the index of the current time layer, M_1 and M_2 are constants that do not depend on h and τ , and $\Phi_1(h, \tau) = h^2 + \tau^2 + (\sigma_1 - \sigma_2)\tau + (\sigma_3 - 0.5)\tau$.

In the next section Theorem 3.1 enables us to prove convergence of difference approximations under relaxed smoothness assumptions that are typical in many applications of thermoelasticity.

4 Convergence of the Operator-Difference Scheme on Weak Solutions of Thermoelasticity Problems.

It is well-known that the analysis of error approximation using the technique of the classical Taylor formula leads to excessive requirements on the smoothness of the sought-for solution [23, 30]. Below we shall show how such requirements can be relaxed using the Steklov operators.

Let us assume that conditions under which the solution of the problem (2.1)–(2.3) belongs to the class $W_2^2(Q_T)$ are satisfied (see [14, 17] and references therein). That is, we assume that $(s, \Theta) \in W_2^2(Q_T) \times W_2^2(Q_T)$.

We consider the operator-difference scheme (3.6), (3.2) with the following source terms

$$\varphi_1 = S^x \otimes S^{t_1} f_1(x, t), \quad \varphi_2 = S^x \otimes S^{t_2} f_2(x, t),$$

where $S^x \otimes S^{t_i}$, $i = 1, 2$ denote the composition of averaging Steklov's operator acting in space and time. The operators S^x and S^{t_i} acting on a function $u(x, t)$ are introduced as follows

$$S^x u(x, t) = \begin{cases} 2 \int_0^{h/2} u(\xi, t) d\xi / h, & \text{when } x = 0, \\ \int_{x-h/2}^{x+h/2} u(\xi, t) d\xi / h & \text{when } 0 < x < 1, \\ 2 \int_{1-h/2}^1 u(\xi, t) d\xi / h & \text{when } x = 1, \end{cases}$$

$$S^{t_1} u(x, t) = \begin{cases} \int_{t-\tau/2}^{t+\tau/2} u(x, \mu) d\mu / \tau, & \text{when } t > 0, \\ 2 \int_0^{\tau/2} u(x, \mu) d\mu / h & \text{when } t = 0, \end{cases}$$

$$S^{t_2}u(x,t) = \begin{cases} \int_t^{t+\tau} u(x,\mu)d\mu/\tau, & \text{when } t > 0, \\ 2 \int_0^{\tau/2} u(x,\mu)d\mu/h & t = 0. \end{cases}$$

Then the scheme error

$$z_1 = v - s, \quad z_2 = \eta - \Theta$$

is the solution of the following problem

$$\begin{cases} D_1(z_1)_{\bar{t}\bar{t}} + B_1(z_1)_0 + A_1z_1 + C_1(z_2)_{\bar{t}\bar{t}} = \psi_1, \\ \bar{B}_2(z_2)_t + A_2(\bar{z}_2 + z_2)/2 + C_2(z_1)_t = \psi_2, \\ z_1 = 0, \text{ when } x = 0, 1; \quad z_1 = 0, (z_1)_t = \psi_1 \text{ when } t = 0, \end{cases} \quad (4.1)$$

where

$$\begin{aligned} \psi_1 &= \varphi_1 - [D_1s_{\bar{t}\bar{t}} + B_1s_0 + A_1s + C_1\Theta_{\bar{t}\bar{t}}], \text{ if } t \in \omega_\tau, \\ \psi_2 &= \varphi_2 - [\bar{B}_2\Theta_t + A_2(\hat{\Theta} + \Theta)/2 + C_2s_t], \text{ if } t \in \omega_\tau, \text{ and} \\ \psi_1 &= \bar{r}_0 + \tau[r_0'' - \Theta_0'' + \varphi_1(x, 0)]/2 - s_t(x, 0), \text{ if } t = 0. \end{aligned}$$

We apply the composition of the operators $S^x \otimes S^{t_1}$ and $S^x \otimes S^{t_2}$ to the first and the second equations of the system (2.1) respectively. Then we use the basic properties of Steklov's operators as follows

$$\begin{aligned} S^x \frac{\partial u}{\partial x} &= \frac{1}{h}[u(x + h/2, t) - u(x - h/2, t)] = (u^{(-0.5)})_x, \\ S^{t_1} \frac{\partial u}{\partial t} &= (\bar{u})_t, \quad S^{t_2} \frac{\partial u}{\partial t} = u_t, \end{aligned}$$

where $\bar{u} = u(x, t + \tau/2)$. In words, the above formulas allow us to transform derivatives that may not exist in the classical sense into difference derivatives. As a result, such a transformation naturally leads to a discrete problem which is constructed with respect to the smoothness requirements on the unknown solution.

It is straightforward to deduce a representation for the scheme error, for example, for internal nodes we have

$$\psi_1 = (\eta_1)_t + \frac{\sigma_1 + \sigma_2}{2}\tau^2(\eta_2)_t + (\eta_3)_x + (\sigma_1 - \sigma_2)\tau(\eta_5)_t,$$

where the functionals η_i , $i = 1, \dots, 5$ are defined as follows

$$\begin{aligned} \eta_1 &= S^x \left(\frac{\partial \bar{s}}{\partial t} \right), \quad \eta_2 = \Lambda s_{\bar{t}}, \quad \eta_3 = s_x - S^{t_1} \left(\left(\frac{\partial s}{\partial x} \right)^{(-0.5)} \right), \\ \eta_4 &= S^x \left(\frac{\partial \bar{\Theta}}{\partial t} \right) - \Theta_{\bar{t}}, \quad \eta_5 = \Lambda(s - \tau s_{\bar{t}}/2), \end{aligned}$$

and similarly,

$$\psi_2 = (1 + \epsilon)(\eta_6)_t + \epsilon(\eta_7)_t - (\sigma_3 - 0.5)\tau(\eta_8)_t + a(\eta_9)_x,$$

where the functionals η_i , $i = 6, 7, 8, 9$ are defined by the formulas

$$\eta_6 = S^x \Theta - \Theta, \quad \eta_7 = S^x s - s, \quad \eta_8 = A_2 \Theta,$$

$$\eta_9 = \left(\frac{\hat{\Theta} + \Theta}{2} \right)_x - S^{t_2} \left(\left(\frac{\partial \Theta}{\partial x} \right)^{(-0.5)} \right).$$

Analogous functions are present in the representations of error approximation of

- boundary conditions for temperature, and
- initial conditions of the problem.

Functionals η_i , $i = 1, \dots, 9$ are estimated using the Bramble-Hilbert lemma (see [23] and references therein). For example, it is easy to see that the linear functional η_3 is bounded in the space $W_2^2(Q_T)$, therewith

$$|\eta_3| \leq M h^{(-1)} \|s\|_{W_2^2(e)},$$

where

$$e = \{(x', t') : x - h < x' < x, t - \tau/2 < t' < t + \tau/2\}.$$

By a standard linear change of variables we can pass from the domain e to the domain

$$E = \{(u_1, u_2) : -1 < u_1 < 0, -0.5 < u_2 < 0.5\}.$$

Since a linear change of variables does not change the class of functions, we have

$$|\eta_3| \leq M h^{(-1)} \|s\|_{W_2^2(E)}.$$

Now it is easy to verify that the functional

$$\eta_3 = \frac{1}{2h} \{ \bar{s}(0, 0) - \bar{s}(-1, 0) - \int_{-0.5}^{0.5} \frac{\partial \bar{s}(-0.5, u_2)}{\partial u_1} du_2 \}$$

(where $\bar{s}(u) = s(x(\xi_1), t(\xi_2))$) is zero for polynomials up to first degree inclusive. Therefore, from the Bramble-Hilbert lemma we have

$$|\eta_3| \leq M h^{-1} |\bar{s}|_{W_2^2(E)},$$

and passing to the variables (x, t) we finally get

$$|\eta_3| \leq M \frac{h^2 + \tau^2}{h} (h\tau)^{(-1/2)} |s|_{W_2^2(E)}.$$

Using the technique of the Bramble-Hilbert lemma for other functionals , and applying the estimate (3.7), we come to the following result

Theorem 4.1. *The solution of the operator-difference scheme (3.6), (3.2) with $\varphi_i = S^x \otimes S^{t_i} f_i$, $i = 1, 2$ converges to the generalized solution of the problem (2.1)–(2.3) $(s, \Theta) \in W_2^2(Q_T) \times W_2^2(Q_T)$ if the stability conditions (3.5) are satisfied. Therewith the following accuracy estimate*

$$\|v^j - s^j\|_{(1)} \leq M_1 \Phi_2(h, \tau), \quad \|\eta^j - \Theta^j\|_{(1)} \leq M_2 \Phi_2(h, \tau), \quad (4.2)$$

where j is the index of the current time layer, M_1 and M_2 are constants that do not depend on h and τ , and $\Phi_2(h, \tau) = h + \tau$. holds.

Remark 4.1. In the case of uncoupled thermoelasticity (when $\epsilon = 0$) and homogeneous boundary conditions for deformations the accuracy estimate (4.2) can be improved using the result of Theorem 3.1.

Remark 4.2. In the coupled case the second order accuracy estimate may be obtained in some special cases if we use a weaker than L^2 metric for the thermoelastic field (see [22] and references therein).

5 Dispersion Analysis.

Impulses of arbitrary shape may be decomposed in a Fourier integral or a Fourier series as a superposition of harmonic waves of the form [34]

$$u(x, t) = \exp[(\kappa x - \omega t)i],$$

where ω is the wave frequency, $\kappa = 2\pi/\lambda$ is the wave number, and λ is the wave length. Due to the dependence of the phase velocity of wave propagation on the wave length (i.e. due to the presence of dispersion) harmonic components of the signal are shifted with respect to each other. As a result, we observe a distortion of the impulse profile [34]. In order to find such a dependency for a thermoelastic media we substitute the harmonics of the differential problems (2.4),

$$r = R \exp[i(\kappa x - \omega t)] \text{ and } \theta = T \exp[i(\kappa x - \omega t)], \quad (5.1)$$

into the differential system of thermoelasticity. In (5.1) R and T denote amplitudes of the corresponding harmonics. Expanding the system determinant and equating it to zero, we get the following dispersion relationship for the differential problem:

$$(\omega^2 - \kappa^2)(i\omega - \kappa^2) - i\epsilon\kappa^2\omega = 0. \quad (5.2)$$

To find the phase velocity from equation (5.2) we may use arguments based on the method of a small parameter. Such an approach may be often justified from the physical point of view, since the coupling parameter ϵ is small for a wide range of known materials. From the mathematical point of view such an approach requires $\epsilon \rightarrow 0$, that may not be true in practice. Indeed, the value of ϵ may be very small, but is nevertheless always positive. If we formally substitute the series

$$\omega = \sum_{k=0}^{\infty} \epsilon^k \omega_k$$

into the dispersion relationship (5.2) and neglect terms of the order $O(\epsilon^2)$, we get two equations that have to be satisfied simultaneously:

$$\begin{cases} (\omega_0^2 - \kappa^2)(i\omega_0 - \kappa^2) = 0 \\ i\omega_1(\omega_0^2 - \kappa^2) + 2\omega_0\omega_1(i\omega_0 - \kappa^2) - i\kappa^2\omega_0 = 0. \end{cases} \quad (5.3)$$

From (5.3) we find three distinct modes, that are approximate roots of the dispersion relationship (5.2), $\omega_{1,app}$, $\omega_{2,app}$, $\omega_{3,app}$. They give us the dependencies of the frequency on the wave number, namely

$$\begin{aligned} \omega_{1,app} &= \kappa \left(1 + \frac{\epsilon}{2(1 + \kappa^2)} \right) - i \frac{\epsilon\kappa^2}{2(1 + \kappa^2)}; \quad \omega_{2,app} = -\kappa \left(1 + \frac{\epsilon}{2(1 + \kappa^2)} \right) - \\ &\quad i \frac{\epsilon\kappa^2}{2(1 + \kappa^2)}; \quad \omega_{3,app} = -i\kappa^2 \left(1 - \frac{\epsilon}{1 + \kappa^2} \right). \end{aligned} \quad (5.4)$$

It is well known that the quantity $\Re \omega$ is responsible for the phase velocity of the harmonics, whereas $\Im \omega$ - for the increase or damping of harmonics. Since the phase velocity of propagation of harmonics is defined as

$$v = \Re \omega / \kappa,$$

we can easily define from (5.4) approximate phase velocities of harmonic propagation for the differential problem (2.4). They are

$$v_1 = 1 + \epsilon/[2(1 + \kappa^2)]; \quad v_2 = -v_1; \quad v_3 = 0. \quad (5.5)$$

Furthermore, we know that when approximate solutions of differential problems are sought (for example, using difference schemes), it is important to take into account the dispersion properties of numerical methods used for such approximations. Such properties may be crucial when problems contain a hyperbolic type operator. Indeed, in this case numerical solutions of such problems are typically accompanied by parasitic oscillations that are connected with the dispersion of harmonics of the applied numerical scheme. In other words such oscillations have purely numerical origin [19]. Therefore it is important to achieve a better correspondence of dispersion properties of differential and discrete problems. Ultimately it is this correspondence that determines the quality of the problem solution because in coupled field theory it is rarely the case that a differential problem may be solved analytically.

Analogue of the harmonics (5.1) on the discrete grid $\bar{\omega}_{hr}$ have the following form

$$y = Rq^n \xi^j, \quad \eta = Tq^n \xi^j, \quad (5.6)$$

where $\xi = \exp(i\kappa h)$, and q is the transfer factor of the difference scheme. It is straightforward to get the dispersion relationship for the difference scheme (2.5), namely

$$p^2/\tau^2 - \zeta p - \zeta([(1/\tau - \sigma\zeta) - \zeta]\epsilon\zeta p(p+1)/\tau) = 0, \quad (5.7)$$

where

$$p = \sum_{k=0}^{\infty} \epsilon^k p_k, \quad \zeta = (\xi - 2 + \xi^{-1})/h^2 = 4 \sin^2(\kappa h/2)/h^2.$$

Using the method of a small parameter we can derive the analogue of (5.3) for the discrete problem (2.5) (see details in [19]). Standard requirements for the proper representation of dispersion effects are the conditions of superiority of the dispersion of the original differential problem compared to the numerical dispersion. Such conditions, which were derived in [19], have to be satisfied together with the stability conditions for the difference scheme. Here we sketch an alternative approach to the investigation of stability of difference schemes.

Let us represent the dispersion relationship (5.7) for the discrete problem through a polynomial of the third degree $Q_3(p) = 0$. Under quite general assumptions the question on system stability can be reduced to the question of whether the roots of the characteristic equation of the linearized system are to the left from the imaginary axis. If such a condition is satisfied then such polynomials are called stable.

We consider a stable polynomial given by the equation $Q_3(z) = 0$ (i.e. the stability domain of $Q_3(z)$ is assumed to be the whole left semi-plane). By the Cayley transform

$$q = (z + 1)/(z - 1) \quad (5.8)$$

the domain $\Im z \leq 0$ is mapped into the unit circle in the complex q plane. This domain is the stability domain of a transformed polynomial $\bar{Q}_3(q) = 0$. Now the transformation

$$p = q - 1 = 2/(z - 1) \quad (5.9)$$

will lead us to the polynomial $Q_3(p)$, the stability domain of which is a circle of unit radius with the center in $(-1, 0)$. The described transformations are sketched on Fig. 1. Now we

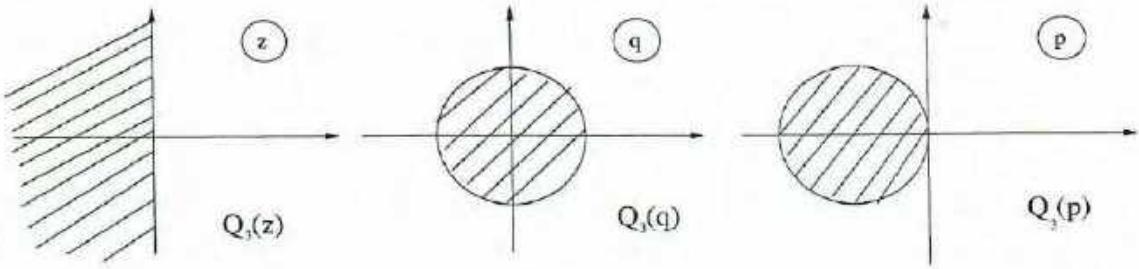


Figure 1:

can substitute into the dispersion equation (5.7) instead of p its expression in terms of z . As a result, we get

$$[-\zeta\tau^2 z^3 + 4\zeta\tau^2][-\zeta\tau z + 2(1 - \sigma\zeta\tau) + \zeta\tau] - 2\epsilon\zeta\tau^2(z^2 - 1) = 0. \quad (5.10)$$

It is easy to see that the equation (5.10) may be rewritten in the form

$$a_0 z^3 + a_1 z^2 + a_2 z + a_3 = 0, \quad (5.11)$$

where

$$\begin{aligned} a_0 &= \zeta^2\tau^3, \quad a_1 = \zeta\tau^2[2(1 - \sigma\zeta\tau) + \zeta\tau + 2\epsilon], \\ a_2 &= -\zeta\tau(4 + \zeta\tau^2), \quad a_3 = (4 + \zeta\tau^2)[2(1 - \sigma\zeta\tau) + \zeta\tau] + 2\epsilon\zeta\tau^2. \end{aligned}$$

A special case of the Routh-Hurwitz theorem on stability of polynomials is the Vyshegradsky criterion that gives necessary and sufficient conditions for stability of third degree polynomials, namely

$$a_i > 0, \quad i = 0, 1, 2, 3; \quad a_1 a_2 > a_0 a_3. \quad (5.12)$$

In our case the conditions $a_0 > 0$ and $a_1 a_2 > a_0 a_3$ will be satisfied for all values of σ, τ, h . The condition $a_2 > 0$ will be satisfied whenever the following inequality

$$\frac{\tau^2}{h^2} \sin^2(\kappa h/2) < 1 \quad (5.13)$$

holds. The conditions $a_1 > 0$ and $a_3 > 0$ will be satisfied if

$$\tau(1 - 2\sigma) < (1 + \epsilon)h^2/[2 \sin^2(\kappa h/2)]. \quad (5.14)$$

We note that the conditions (5.13), (5.14) are equivalent to the stability conditions (3.5) for the difference scheme (2.5)-(2.6).

In conclusion we recall that the main property of the Cayley transform is the preservation of the main global properties of the system. Indeed, the transform (5.8) is a special case of a more general operator Cayley transform

$$T_\gamma = (\gamma I + A)(\gamma I - A)^{-1}, \quad (5.15)$$

where A is a linear operator acting in Banach spaces, $\gamma > 0$. The Cayley transform (5.15) allows us to transform conservative dissipative systems evolving in continuous time into discrete systems with the same global properties [1]. Under quite general assumptions the Cayley transform technique provides a general way for the construction of effective numerical approximations for differential equations in Banach spaces [13]. We will discuss these questions in details elsewhere.

6 Computational Experiments.

In this section we consider an application of the constructed and rigorously justified discrete models to the investigation of planar nonstationary waves in a thermoelastic layer under instantaneous action of surface forces of intensity p_0 . Assuming that surfaces of the layer are thermo-isolated, the subject of investigation is reducible to the following initial-boundary value problem:

$$\begin{cases} \frac{\partial^2 \bar{\sigma}}{\partial \xi^2} - \frac{1}{c^2} \frac{\partial^2 \bar{\sigma}}{\partial \zeta^2} = \frac{1+\nu}{1-\nu} \rho \alpha_T \frac{\partial^2 (T - T_0)}{\partial \zeta^2}, \\ \frac{\partial^2 T}{\partial \xi^2} - \frac{1}{a} (1+\epsilon) \frac{\partial T}{\partial \xi} = \frac{(1+\nu) \alpha_T T_0}{(1-\nu) \lambda_q} \frac{\partial \bar{\sigma}}{\partial \xi}, \end{cases} \quad (6.1)$$

$$T = T_0, \bar{\sigma} = 0, \frac{\partial \bar{\sigma}}{\partial \zeta} = 0 \text{ for } \zeta = 0, \quad (6.2)$$

$$\bar{\sigma} = -p_0 H(\xi), \frac{\partial T}{\partial \xi} = 0 \text{ for } \xi = 0, L. \quad (6.3)$$

The notation is standard, namely

- $\bar{\sigma}$ denotes the stress in the layer;
- T is the temperature at a given moment of time;
- T_0 is the temperature of the layer in the non-deformed state;
- ρ is the density of the material;
- α_T is the averaged coefficient of linear thermal expansion;
- λ_q is the coefficient of heat conductivity;
- ν is the Poisson coefficient;
- c is the velocity of elastic extension wave;
- $H(\zeta)$ is the unit Heaviside function.

After rescaling the problem (6.1)-(6.3) may be cast in the form of the model (2.1)-(2.3) with

$$f_1 = f_2 = 0, \Theta_0 = s_0 = \bar{s}_0 = 0, \Theta_i(t) = 0, s_i = -H(t).$$

Then the discrete model (2.5), (2.6) allows us to compute deformations explicitly. The algorithm goes as follows. Assuming that we have computed

- the values of deformations on the i -th and $(i+1)$ -st time layers (except at the two ends of the layer), and
- the values of temperature on the i -th time layer (for $i=0$ they are known from the initial conditions),

we obtain a system of linear algebraic equation for the determination of the values of temperature in all spatial points of the $(i+1)$ -st time layer. Then we compute the values of deformations at both ends of the layer. This allows us to determine explicitly the values of deformations on the $(i+2)$ -nd time layer (again, except for the ends of the layer) etc.

Figures 1 throughout 4 show the distribution of stress and temperature in a steel layer ($\epsilon = 0.0114, a = 0.2 \times 10^{-8}$). A jump at the beginning of the observation is due to the action of inertia after surface forces were applied. Two thermoelastic waves, propagating from the layer boundaries, have not met yet. As a result, we observe a contractive stress that appears

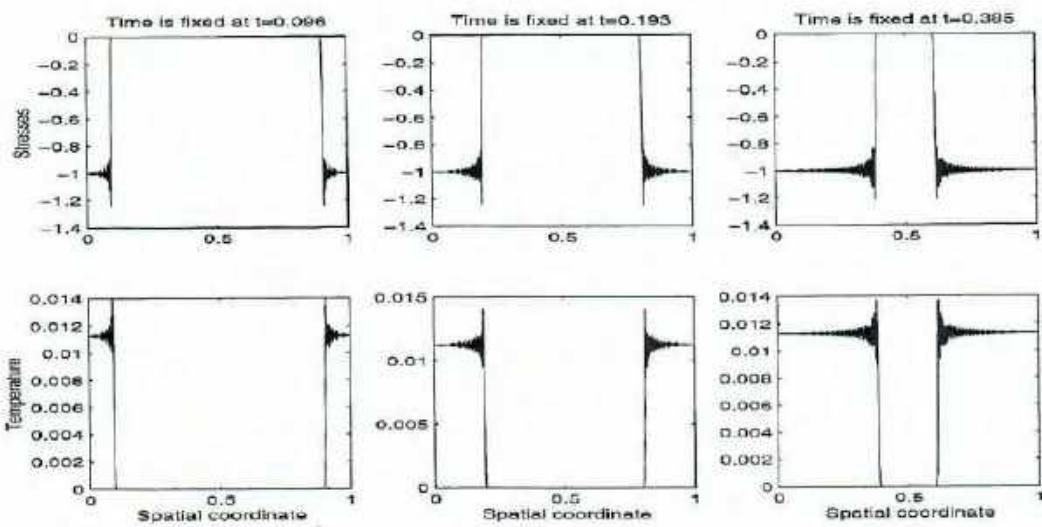


Figure 2:

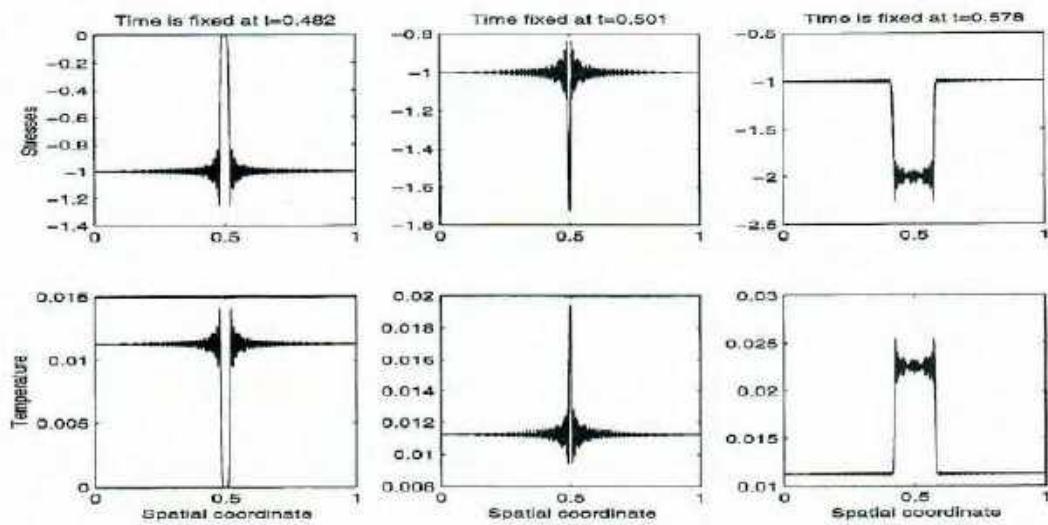


Figure 3:

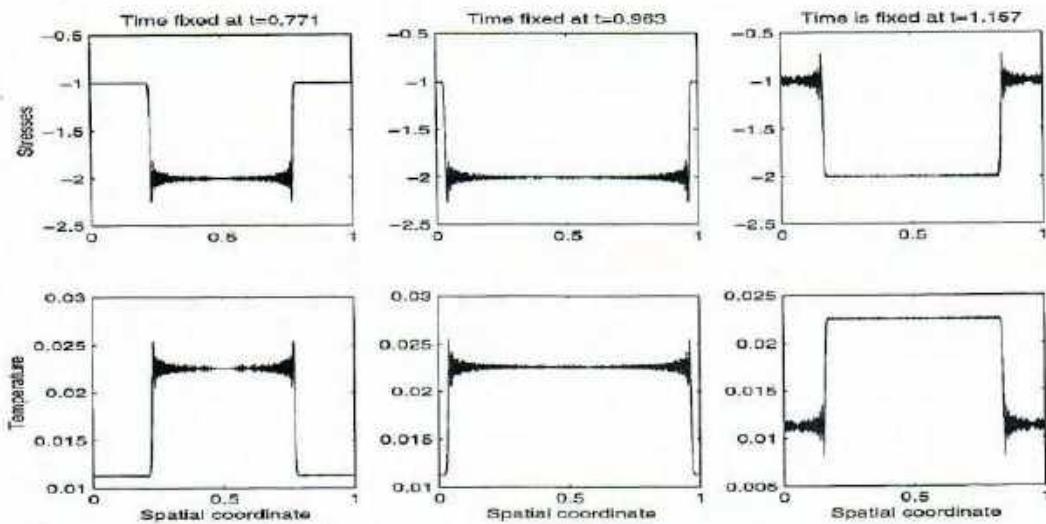


Figure 4:

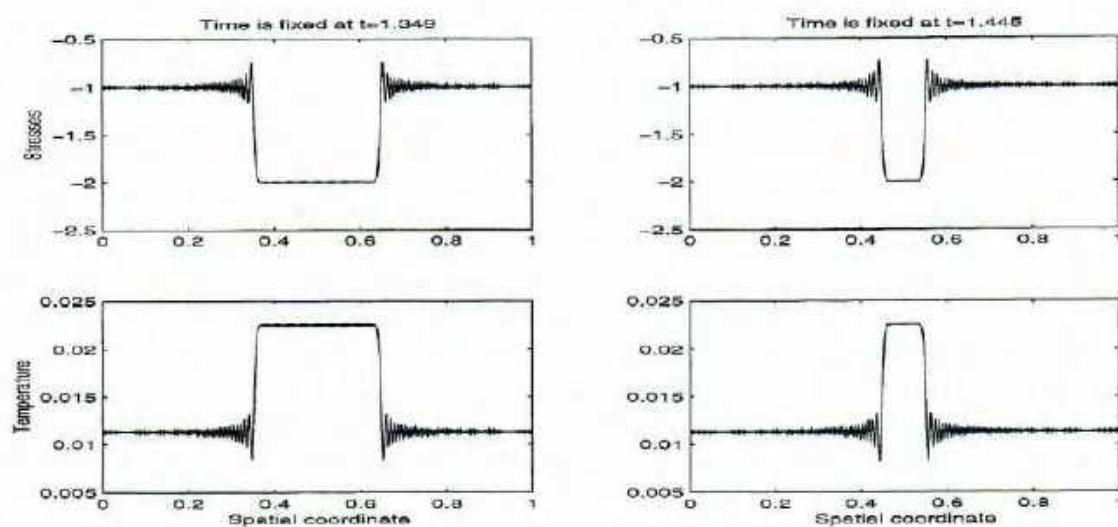


Figure 5:

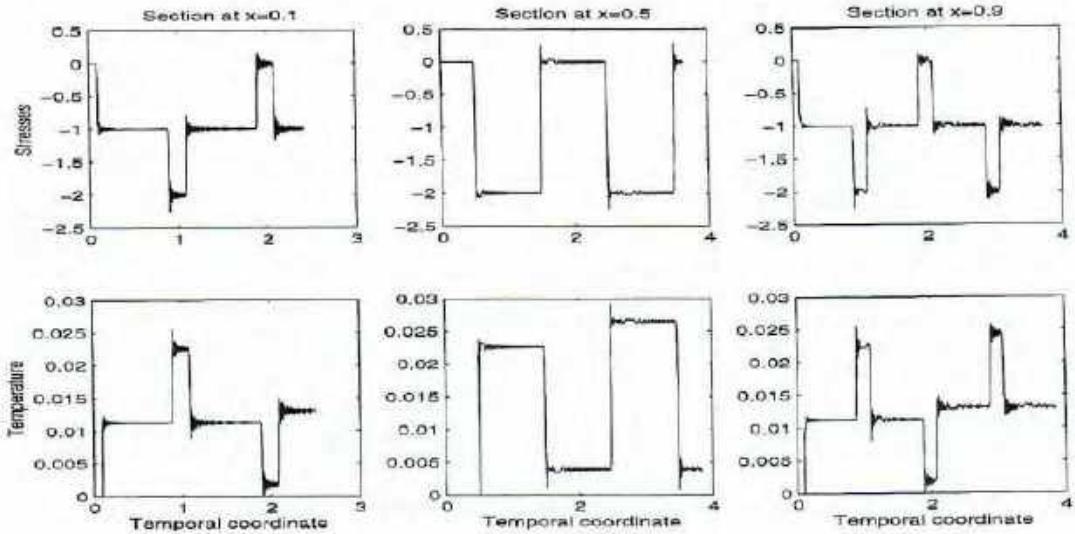


Figure 6:

in the layer. As time goes on, waves meet and we observe the “concentration” of stresses of magnitude $-2p_0$. Such a concentration is expanding with time from the middle to the layer boundaries. This reflects the physical essence of the process. Finally, Figure 5 shows the distribution of stress and temperature with respect to time in different sections of the layer. The problem provides an example where neither spatial nor temporal derivatives exist in the classical sense. However, the developed computational procedures based on discrete models for thermoelasticity allowed us to correctly convey both qualitative and quantitative features of the process under investigation.

7 Future directions.

The effective computational procedures developed for problems in coupled thermoelasticity can be used as building blocks for the development of refined models in other areas of coupled field theory, for example, in coupled thermo-electroelasticity. Interest in piezoelectrics has been revitalized [2] by its significance for smart materials [24] and the importance of piezoelectricity in biopolymers [12]. Mathematical models of dynamic electroelasticity have been extensively studied in the literature (see [23] and references therein). However, notwithstanding the coupled treatment of electro-mechanical fields in such models, for many practical applications it is important to take into consideration thermal effects as temperature has a profound influence on the properties of piezoelectrics.

To be competitive numerical codes in coupled field theory have to be adaptive. One of the major goals of adaptive computational schemes is to control the computational process. The success of this process often hinges on how well the numerical error can be estimated. This implies that in the general case we need a combination of a-priori and a-posteriori estimates [15]. The major difficulty in obtaining such a combination stems from the fact that the error needs to be integrated with respect to time which complicates numerical analysis in the nonstationary case. Computationally we are inevitably led to an optimization problem and the whole computational process can be seen as a problem of *optimal error control* (see [21] and references therein).

Acknowledgements

The author wishes to acknowledge the support of the University of Southern Queensland and to thank Senior Lecturer Walter Spunde for helpful assistance at the final stage of preparation of this paper.

References

- [1] Arov, D.Z. and I.P Gavriljuk, A method for solving initial value problems for linear differential equations in Hilbert space based on the Cayley transform, *Numer. Funct. Anal. and Optimiz.*, 14(5&6), 459-473, 1993.
- [2] Ballato, A., Piezoelectricity: Old Effect, New Thrusts, *IEEE Trans. on Ultrasonics, Ferroelectrics, and Frequency Control*, v. 42, No. 5, 916, 1995.
- [3] Bolliat, G. et al *Recent Mathematical Methods in Nonlinear Wave Propagation*, Springer-Verlag, 1996, Lecture Notes in Mathematics 1640.
- [4] Dhaliwal, R., Wang, J., Some theorems in generalized nonlocal thermoelasticity, *Intern. Journal of Engin. Science*, v. 32, No. 3, 473, 1994.
- [5] Duvaut, G., Lions, J.L., Inequations en thermo-elasticite et magneto-hydrodynamique, *Rat. Mech. Analysis*, 46, 241-279, 1972.
- [6] Erikson, K. et al Introduction to adaptive methods for differential equations, *Acta Numerica*, 105-158, 1995.
- [7] Esham, B.F. and R. Weinacht, Singular Perturbations and the Coupled/Quasi-Static Approximation in Linear Thermoelasticity, *SIAM J. Math. Anal.*, Vol.25, No.6, 1521-1536, 1994.
- [8] Fatemi, E.A., Linear Analysis of the Hydrodynamic Model, *Numer. Funct. Anal. and Optimiz.*, 16(3&4), 303-314, 1995.
- [9] Fellipa, C., Farhat, C., Park, K., Research in grnd challenge coupled problems in computational mechanics, in *Proceedings of the Third World Congress on Computational Mechanics, Chiba/Japan, 1994*, 554-555.
- [10] Fleming, W. and Soner, H., *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, 1993.
- [11] Fujita, H., and Suzuki, T. Evolution problems, in "Handbook of Numerical Analysis", Ed. P.G. Ciarlet & J.L. Lions, North-Holland, Amsterdam, New York, Oxford, Tokyo, 1991, 789-923.
- [12] Fukada, E. , Poiseuille Medal Award Lecture: Piezoelectricity of biopolymers, *Biorheology*, v. 32, No. 6, 593, 1995.
- [13] Gavriljuk, I.P. and V.L. Makarov, Representation and approximation of the solution of an initial value problem for a first order differential equation in Banach spaces, *ZAA (Journal for Analysis and its Applications)*, Vol.15, No. 2, 495-527, 1996.
- [14] Gawinecki, J. , The Faedo-Galerkin method in thermal stresses theory, *Roczn. PTM. Ser. I*, V. 27, No. 1, 83-107, 1987. v. 22, No. 5, 467, 1995
- [15] Johnson, C., A new paradigm for adaptive finite element methods, in "The Mathematics of Finite Elements and Applications", Ed. J.R. Whiteman, John Wiley & Sons, 1994, 105-120.
- [16] Kovalenko, A.D., *Thermoelasticity*, Kiev, Naukova Dumka, 1975.
- [17] Kowalski, L., Existence and uniqueness of the solution of the boundary-initial value problem for linear hyperbolic thermoelasticity equations, *Annales Societatis Mathematicae Polonae*, v. 33, 73, 1993.
- [18] Kurtz, T., Protter, P., Weak convergence of stochastic integrals and differential equations, in *Lecture Notes in Mathematics*, No. 1627, 1996, 197-285.
- [19] Melnik, V.N., Dispersion analysis of a difference scheme for computing of thermotense crystal conditions, *Design Automation in Electronics*, ISSN 0320-6920, No. 46, 71-76, 1992.
- [20] Melnik, V.N., Nonlinear Dynamical Systems: Coupling Information and Energy in Mathematical Models, *40th Conference of the Australian Mathematical Society, Adelaide, July, 1996*, 34-35.
- [21] Melnik, V.N., On Consistent Regularities of Control and Value Functions, *Numer. Funct. Anal. and Optimiz.*, 18(3&4), 401-426, 1997.
- [22] Melnik, R.V.N. The stability condition and energy estimate for nonstationary problems of coupled electroelasticity, *Mathematics and Mechanics of Solids* 2: 153-180, 1997.

- [23] Melnik, V.N., Convergence of the operator-difference scheme to generalized solutions of a coupled field theory problem, *to appear in Journal of Difference Equations and Applications*, 1997.
- [24] Melnik, V.N., Intelligent Structures and Coupling in Mathematical Models: Examples from Dynamic Electroelasticity, *to appear in Proceedings of the IEEE ICPADM'97*, Seoul, Korea, 1997.
- [25] Mihalescu-Suliciu, M., Siliciu, I., Energy estimates and energy control of numerical stability in coupled dynamic thermoelasticity, *Mechanics Research Communications*, v. 22, No. 5, 467, 1995.
- [26] Moskalkov, M.N., On accuracy of difference schemes for approximation of wave equation with piecewise smooth solutions, *Computational Mathematics and Mathematical Physics*, 14, 1974, 390-401.
- [27] Muller, I. and T. Ruggeri, *Extended Thermodynamics*, Springer-Verlag, 1993.
- [28] Nowacki, W. *Dynamic Problems of Thermoelasticity*, Noordhoff International Publishing, Leyden, The Netherland and PWN, 1975.
- [29] Ortin, J. and A. Planes, Thermodynamics and Hysteresis Behaviour of Thermoelastic Martensic Transformations, *Journal de Physique IV, Colloque C4*, Vol. 1, 13-23, 1991.
- [30] Partial Differential Equations with Minimal Smoothness and Applications, *Eds. B. Dahlberg et al, The IMA Volumes in Mathematics and its Applications*, Vol. 42, Springer-Verlag, 1992.
- [31] Rajagopal, K.P., Boundary layers in finite thermoelasticity, *J. of Elasticity*, 36, 271-301, 1995.
- [32] Samarski, A.A. et al *Blow-up in Quasilinear parabolic equations*, de Gruyter, Berlin and Hawthorne, NY, 1995.
- [33] Shashkov, M., Steinberg, S., Conservative Finite-Difference Methods on General Grids, *Boca Raton: CRC Press*, 1995.
- [34] Shokin, Yu.I. and N.N. Yanenko, *The Method of Differential Approximations*, Novosibirsk, Nauka, 1985.
- [35] Traub, J.F. and Wozniakowski, H. *A General Theory of Optimal Algorithms*, New York: Academic Press, 1980.
- [36] Tveito, A. and R. Winther, The solution of nonstrictly hyperbolic conservation laws may be hard to compute, *SIAM J. Sci. Comput.*, Vol. 16, No. 2, 320-329, 1995.

UNIVERSITY OF SOUTH AUSTRALIA



TECHNICAL REPORT SERIES

**Generalised Solutions, Discrete
Models and Energy Estimates for
A 2D Problem of Coupled Field Theory**

by

R V N Melnik

Report No. 1998/1

**CENTRE FOR INDUSTRIAL
AND APPLIED MATHEMATICS
SCHOOL OF MATHEMATICS**

Faculty of Information Technology

The Levels, South Australia 5095, Telephone (08) 8302 3343 Facsimile (08) 8302 5785

TECHNICAL REPORT SERIES

**Generalised Solutions, Discrete
Models and Energy Estimates for
A 2D Problem of Coupled Field Theory**

by

R V N Melnik

Report No. 1998/1

GENERALISED SOLUTIONS, DISCRETE MODELS AND ENERGY ESTIMATES FOR A 2D PROBLEM OF COUPLED FIELD THEORY

R. V. N. Melnik *

School of Mathematics, University of South Australia, SA 5095

Abstract

In this article the author constructs an effective difference scheme for the solution of coupled dynamic electroelasticity problems for finite dimensional solids. Such a construction is based on the energy conservation law for the electromechanical system and the definition of generalised solutions for the corresponding differential problem. Results of computational experiments are presented for hollow piezoceramic cylinders in the two-dimensional case.

Key words: piezoelectrics, discrete conservation laws for coupled systems, electro-mechanical interactions, generalised solutions.

1 Introduction: Piezoceramic Applications and Coupled Field Theory

Piezoelectricity is an example of phenomena where coupling two physical fields of different natures (namely mechanical and electrical fields) is a key factor to be taken into account in a variety of applications. It is just one of many important examples where two theories, originally developed independently of each other (in this case the theory of elasticity and the Maxwell theory of electromagnetic waves), have to be considered in intrinsic correlation. Such examples are usually assigned to the domain of coupled field theory.

Discovered in 1880 by the Curie's, piezoelectricity has essentially contributed to technological advances. Since the time of Langevin's sonar emitter, more and more devices and functional components are being based on piezoelectrics. The development of new piezoelectric material has echoed in the rapid growth in many industrial applications including telecommunications, consumer electronics and many other areas. Piezoelectricity has been put on a rigorous mathematical basis

*Permanent address: Department of Mathematics and Computing, University of Southern Queensland, QLD4350, E-mail: melnik@usq.edu.au

by classical works of W. Voigt, R.D. Mindlin, H.F. Tiersten, W.P. Mason, G. A. Maugin, A.F. Ulitko, W. Nowacki and many other researchers (see references in [2, 3, 10, 13, 26, 27]).

One of the first piezoelectric material that deserved a lot of attention from engineers and designers was the lead-zirconate-titanate (PZT) piezoceramic (and its derivatives). The PZT-family of piezoelectrics has a number of important properties, such as moderate temperature range, resistance to mechanical and electrical stress-induced depolarisation, large piezoelectric coefficients [9]. Of course, it would be proper to say that ceramics have been always a part of human life. However active applications of ceramic materials began just about six decades ago. Later, in the 1950's, piezoceramics started being widely employed for industrial purposes.

By now it is technologically possible (for example, by doping of PZT ceramics) to achieve many useful properties of piezoceramics such as, a high surface coupling coefficient, a low temperature coefficient of delay time and others. As a result, doped piezoceramics have become very important in many high temperature, high frequency device applications. For example, piezoceramic substrates are widely used for surface acoustic wave (SAW) applications including filters and biological/chemical sensors [8]. Piezoceramic is an intrinsic part in many ultrasonic transducer elements for non-destructive material evaluation, infrared sensors, hydrophones, medical diagnostics etc. An important aspect of new piezoceramic applications is that new technologies become more ecologically friendly. For new families of piezoceramics (including those with bismuth perovskites), the atmosphere control over the lead-induced evaporation is either not necessary, or can be minimised. Another advantage of new piezoceramics is due to their properties under a large level of vibrations for high-power piezoelectric devices (such as piezoelectric actuators or ultrasonic motors). Indeed, in many applications, one has to confront the problem of heat generation and subsequent changes in piezoelectric properties which is difficult to solve for classical piezoceramics (for example, due to the lead loss in PZT-based piezoceramics) [31].

In many applications PZT-based ceramics remain the most extensively used piezoelectric material. Amongst many other advantages it is cheap and easy to fabricate. During the last few decades piezoelectric polymers have become an alternative material. There are applications (such as hydrophone design) where it is important to have a higher acoustic match between ceramics and the transmitting media as well as to be able to form curved surfaces [12]. In the latter cases the use of piezoelectric polymers (such as Polyvinylidene Fluoride (PVDF) and its copolymers with trifluoroethylene) is quite successful [4]. In addition, many piezoceramic polymers have a high electrical conductivity and are being widely used in many technical devices, including those in laser and infrared technologies. However, the main problems with polymers lie with, typically, a low dielectric constant, relatively high cost of fabrication and a low electromechanical coupling. Therefore many technological advances are connected with the use of composites based on the PZT-ceramics which have a high thermal stability and many other useful properties. Some piezoelectric ceramic/polymer composites have a great future potential to replace classical piezoceramics and piezopolymers. In many cases such composites are not difficult to fabricate, for example, by embedding of PZT rods in a polymer/copolymer matrix.

Recent interest in smart structures has once again reemphasised the importance of rigorous mathematical tools in a better understanding of the electroelastic behaviour of piezoceramics as an integral part of structures [23].

In a large number of the applications discussed above, it is important to have quantitative characteristics of the electroelastic behaviour of finite dimensional solids in the dynamic, rather

than in the stationary, regime taking into account the influence of electric and mechanical fields onto each other. The quantitative understanding of the effect produced by the coupling of electrical and mechanical field is essential to the successful design of components and devices [6]. However, analytical solutions for coupled problems of piezoelectricity are limited. This is especially true for the nonstationary case. As a consequence, numerical methods become the most natural and effective tool of research in coupled field theory in general, and in piezoelectricity in particular [1, 5, 14, 15, 16, 17, 18, 19, 21, 22, 24, 25, 33].

In this paper we consider a two-dimensional, nonstationary model for the description of coupled electromechanical fields in piezoceramic materials. We give a rigorous mathematical derivation of an effective numerical method for the investigation of such fields in finite-length hollow piezoceramic cylinders. The article is organised in five sections.

- Section 2 is concerned with the main notation used throughout of the paper.
- In Section 3 we consider the mathematical model of coupled dynamic electroelasticity in the 2D case and discuss the main approaches for its treatment.
- Section 4 deals with the energy balance equation for the piezoelectric body. Such an equation is derived from the definition of the generalised solution of the original differential problem.
- In Section 5 we construct the difference scheme from the energy balance equation.
- Section 6 addresses the important issues of boundedness of the energy operator for the differential model.
- The results of computational experiments are presented in Section 7.

2 Notation.

The following notations are used throughout this paper

- $\tilde{Q}_T = \tilde{G} \times \tilde{I}$ where

$$\tilde{G} = \{(r, z) : R_0 \leq r \leq R_1, Z_0 \leq z \leq Z_1\}, \quad \tilde{I} = \{t : 0 \leq t \leq T\}$$

is the space-time region of interest;

- u_r and u_z are components of the displacement vector;
- ϕ is the electric field potential;
- $f_i(r, z, t)$, $i = 1, 2$ are components of the vector of mass forces;
- $f_3(r, z, t)$ is the function of volume charges;
- ρ is the density of the piezoceramic material;
- E_r and E_z are vector components of the stress of electric field;
- D_r and D_z are vector components of electric induction;
- σ_r , σ_θ , σ_z and σ_{rz} are components of the field of stresses;
- c_{kl} tensor components of elastic quantities;
- e_{ij} tensor components of electro-elastic quantities;
- c_{kl} tensor components of electric quantities;

- $\bar{\omega}_{h\tau} = \bar{\omega}_h \times \bar{\omega}_\tau$ is the difference grid that covers the space-time region Q_T ;
- $\bar{\omega}_h = \bar{\omega}_{h_1} \times \bar{\omega}_{h_2}$;
- $\bar{\omega}_{h_1} = \{r_i : r_i = R_0 + ih_1, i = 0, 1, \dots, N, h_1 = (R_1 - R_0)/N\}$ is the r -direction difference grid;
- $\bar{\omega}_{h_2} = \{z_j : z_j = Z_0 + jh_2, j = 0, 1, \dots, M, h_2 = (Z_1 - Z_0)/M\}$ is the z -direction difference grid;
- $\bar{\omega}_\tau = \{t_k : t_k = k\tau, \tau = T/L, k = 0, 1, \dots, L\}$ is the temporal grid;
- “flux” nodes are introduced as

$$\tilde{r} = r - h_1/2, \quad \tilde{z} = z - h_2/2,$$

(in these nodes we define deformations and stresses);

- $\omega_{h_1} = \{r_i = R_0 + ih_1, i = 1, \dots, N-1\}, \omega_{h_1}^+ = \{r_i = R_0 + ih_1, i = 1, \dots, N\}, \omega_{h_1}^- = \{r_i = R_0 + ih_1, i = 0, \dots, N-1\}$ are auxiliary grids (in a similar way we define grids $\omega_{h_2}, \omega_{h_2}^+, \omega_{h_2}^-$);
- $\gamma_1 = \{(r, z) : R_0 < r < R_1, z = Z_0\}, \gamma_2 = \{(r, z) : R_0 < r < R_1, z = Z_1\}, \gamma_3 = \{(r, z) : r = R_0, Z_0 < z < Z_1\}, \gamma_4 = \{(r, z) : r = R_1, Z_0 < z < Z_1\}$ are the boundaries of the spatial region G ;
- $\gamma_{13} = \{r = R_0, z = Z_0\}, \gamma_{23} = \{r = R_0, z = Z_1\}, \gamma_{24} = \{r = R_1, z = Z_1\}, \gamma_{14} = \{r = R_1, z = Z_0\}$ are the vertices (corner points) of the region G .

We use the standard notation for difference derivatives of a function $y(r, z, t)$ defined on the grid $\bar{\omega}_{h\tau}$ [28, 29, 22, 30]. For example,

- $y_r = (y(r+h_1, z, t) - y(r, z, t))/h_1$ denotes the forward difference derivative in the r -direction;
- $y_{\bar{r}} = (y(r, z, t) - y(r-h_1, z, t))/h_1$ denotes the backward difference derivative in the r -direction;
- $y_{rr} = (y(r+h_1, z, t) - 2y(r, z, t) + y(r-h_1, z, t))/h_1^2$ is the second (“central”) difference derivative in the r -direction.

In a similar manner we define difference derivatives in the z -direction and the derivatives with respect to t .

3 Mathematical Models of Coupled Dynamic Electroelasticity.

Mathematically, the process of propagation of axially symmetric electroelastic waves in hollow finite-length piezoceramic cylinders with radial preliminary polarisation, can be described by a coupled system of partial differential equations which include

- the equations of motion of the piezoelectric medium in cylindrical coordinates

$$\rho \frac{\partial^2 u_r}{\partial t^2} = \frac{\partial \sigma_r}{\partial r} + \frac{\partial \sigma_{rz}}{\partial z} + \frac{\sigma_r - \sigma_\theta}{r} + f_1, \quad (3.1)$$

$$\rho \frac{\partial^2 u_z}{\partial t^2} = \frac{\partial \sigma_{rz}}{\partial r} + \frac{\partial \sigma_z}{\partial z} + \frac{\sigma_{rz}}{r} + f_2, \quad (3.2)$$

- the Maxwell equation for piezoelectrics (in the acoustic range of frequencies, it is the forced electrostatic equation of dielectrics)

$$\frac{1}{r} \frac{\partial}{\partial r} (r D_r) + \frac{\partial D_z}{\partial z} = f_3, \quad (3.3)$$

- and, state equations for piezoceramic with radial preliminary polarisation

$$\begin{cases} \sigma_r = c_{33}\epsilon_r + c_{13}(\epsilon_\theta + \epsilon_z) - e_{33}E_r, & \sigma_\theta = c_{13}\epsilon_r + c_{11}\epsilon_\theta + c_{12}\epsilon_z - e_{13}E_r, \\ \sigma_z = c_{13}\epsilon_r + c_{12}\epsilon_\theta + c_{11}\epsilon_z - e_{13}E_r, & \sigma_{rz} = c_{44}\epsilon_{rz} - e_{15}E_z, \\ D_r = e_{33}\epsilon_r + e_{13}(\epsilon_\theta + \epsilon_z) + e_{33}E_r, & D_z = 2e_{15}\epsilon_{rz} + e_{11}E_z \end{cases} \quad (3.4)$$

(or state equations for piezoceramic with circular preliminary polarisation [3, 25]).

The function of electrostatic potential is introduced by the formulae

$$E_r = -\frac{\partial \varphi}{\partial r}, \quad E_z = -\frac{\partial \varphi}{\partial z} \quad (3.5)$$

and the relationship between deformations and displacements (Cauchy relationships) have the form

$$\epsilon_r = \frac{\partial u_r}{\partial r}, \quad \epsilon_\theta = \frac{u_r}{r}, \quad \epsilon_z = \frac{\partial u_z}{\partial z}, \quad \epsilon_{rz} = \frac{1}{2} \left(\frac{\partial u_r}{\partial z} + \frac{\partial u_z}{\partial r} \right). \quad (3.6)$$

The system (3.1)–(3.6) is considered in the space-time region Q_T and is complemented by the following initial conditions

$$u_r(r, z, 0) = u_r^{(0)}(r, z), \quad \frac{\partial u_r(r, z, 0)}{\partial t} = u_r^{(1)}(r, z), \quad (3.7)$$

$$u_z(r, z, 0) = u_z^{(0)}(r, z), \quad \frac{\partial u_z(r, z, 0)}{\partial t} = u_z^{(1)}(r, z) \quad (3.8)$$

with known functions $u_r^{(i)}$, $u_z^{(i)}$, $i = 1, 2$.

Mechanical boundary conditions for this problem have the form

$$\sigma_r(R_i, z, t) = p_r^{(i)}(z, t), \quad \sigma_z(r, Z_i, t) = p_z^{(i)}(r, t), \quad i = 0, 1, \quad (3.9)$$

$$\sigma_{rz}(R_i, z, t) = p_{zt}^{(i)}(z, t), \quad \sigma_{rz}(r, Z_i, t) = p_{rt}^{(i)}(r, t), \quad i = 0, 1, \quad (3.10)$$

where $p_r^{(i)}$, $p_z^{(i)}$, $p_{zt}^{(i)}$, $p_{rt}^{(i)}$, $i = 0, 1$ are given functions.

Finally, the formulation of electric boundary conditions depends on the character of the electric loading and the location of electrodes on the body surface. We assume that lateral surfaces of the cylinder are covered by infinitely thin short circuiting electrodes, and that the dielectric permittivity of the surrounding media is much less than the dielectric permittivity of ceramics (this is of course true for both vacuum and air). Then we have

$$\varphi(R_i, z, t) = 0, \quad D_z(r, Z_i, t) = 0, \quad i = 0, 1. \quad (3.11)$$

The assumption of homogeneity of electric boundary conditions made in (3.11) does not restrict generality of the model. Indeed problems with nonhomogeneous conditions can be reduced to homogeneous ones using the procedure described in [22].

In the general case, the determination of electric field in a piezoelectric body has to be performed simultaneously with the determination of its deformation. In other words, a consistent solution of system (3.1)–(3.3) has to be found with corresponding mechanical and electrical boundary conditions defined by (3.9)–(3.11), state equations (3.4) and initial conditions (3.7), (3.8).

In its essence, the solution of (3.1)–(3.11) represents mixed electroelastic waves. The coupling of electric and elastic components in a unified electroelastic field and the necessity of dealing with nonstationary rather than steady-state situations lead to major difficulties in the rigorous investigation of behaviour of piezoceramic-based solids. For dynamic problems of coupled field theory in general, and for piezoelectricity in particular, numerical methods become the natural and efficient way of problem solution. Amongst the numerical techniques effectively used in this field are; Finite Element Approach (FEM) [1, 5, 14, 15, 17, 18], Boundary Element Method (BEM) [16, 19, 33] and Method of Finite Differences (FDM) [21, 22, 24, 25]. The latter method is very attractive since it can be readily applied to practically any system of differential equations that is especially important in coupled field theory. In addition, FDM enjoys many of the advantages present in FEM and BEM provided the differential problem is reformulated in a variational form. Such a reformulation for mathematical models of piezoelectricity goes back to Mindlin's works (see references in [3, 22]). Of course, the existence of physical conservation laws for a physical quantity is closely connected with the description of a system using variational principles. Therefore, one of the features of models for electroelasticity that is desirable to transfer to a difference scheme, is their original variational properties. An effective way to construct such difference schemes, is the method of approximation of variational functional [28, 29, 30].

From the mathematical point of view, the most natural way of variational reformulation of the differential problem, is based on the concept of generalised solutions. Such a reformulation puts us closer to the real practical situations where solutions may not be smooth in the classical sense. For coupled dynamic piezoelectricity, this idea was first applied by the author of this article and his collaborators in the one-dimensional case (see [22] and references therein). Effective difference schemes for infinite-length hollow piezoceramic cylinders were constructed and investigated in [22, 24]. In [20] a theory of existence and uniqueness for generalised solutions of the associated mathematical models was developed. In this paper we deal with the modelling of finite-length cylinders.

4 Generalised Solutions of Coupled Electroelasticity and the Energy Balance Equation.

The solution of problem (3.1)–(3.11) will be understood in the following sense.

Definition 4.1 [21]. *The triple of functions*

$$(u_r(r, z, t), u_z(r, z, t), \varphi(r, z, t)) \in [W_2^1(Q_T)]^2 \times L_2\left(I, W_2^0(G)\right)$$

such that for $t = 0$ $u_r(r, z, t) = u_r^{(0)}(r, z)$ and $u_z(r, z, t) = u_z^{(0)}(r, z)$ is called the generalised solution of the coupled dynamic problem electroelasticity (3.1)–(3.11) if it satisfies the following

identities

$$\int_{Q_T} r \left(-\rho \frac{\partial u_r}{\partial t} \frac{\partial \eta_1}{\partial t} + \sigma_r \frac{\partial \eta_1}{\partial r} + \frac{\sigma_\theta}{r} \eta_1 + \sigma_{rz} \frac{\partial \eta_1}{\partial z} \right) dr dz dt - \\ \int_G r \rho u_r^{(1)}(r, z) \eta_1(r, z, 0) dr dz = \int_{Q_T} r f_1 \eta_1 dr dz dt \quad \forall \eta_1 \in \hat{W}_2^1(Q_T), \quad (4.1)$$

$$\int_{Q_T} r \left(-\rho \frac{\partial u_z}{\partial t} \frac{\partial \eta_2}{\partial t} + \sigma_{rz} \frac{\partial \eta_2}{\partial r} + \sigma_z \frac{\partial \eta_2}{\partial z} \right) dr dz dt - \\ \int_G r \rho u_z^{(1)}(r, z) \eta_2(r, z, 0) dr dz = \int_{Q_T} r f_2 \eta_2 dr dz dt \quad \forall \eta_2 \in \hat{W}_2^1(Q_T), \quad (4.2)$$

$$\int_G r D_r \frac{\partial \zeta}{\partial r} dr dz - \int_G r \frac{\partial D_z}{\partial z} \frac{\partial \zeta}{\partial z} dr dz = \int_G r f_3 \zeta dr dz \\ \forall \zeta \in \overset{0}{W}_2^1(G) \text{ and for a.e. } t \in (0, T). \quad (4.3)$$

The metric space $\hat{W}_2^1(Q_T)$, used in the definition, consists of such elements of $W_2^1(Q_T)$ which equal zero at $t = T$, and the space $\overset{0}{W}_2^1(G)$ consists of such elements of $W_2^1(G)$ which equal zero for $r = R_0, R_1$.

The existence and uniqueness of generalised solutions for the model (3.1)–(3.11) was investigated in [21] where the following theorem was proved.

Theorem 4.1 If

$$f_i \in W_2^1(I, H), i = 1, 2, \quad \frac{\partial f_3}{\partial t} \in L_2 \left(I, \left(\overset{0}{W}_2^1(G) \right)^* \right), \\ f_3|_{t=0} \in \left(\overset{0}{W}_2^1(G) \right)^*, \quad u_r^{(0)}, u_z^{(0)} \in W_2^1(G), \quad u_r^{(1)}, u_z^{(1)} \in L_2(G),$$

then there exists a unique generalised solution of the problem (3.1)–(3.11) with the following properties

$$\frac{\partial u_r}{\partial t}, \frac{\partial u_z}{\partial t} \in L_2(I, H), \quad \frac{\partial^2 u_r}{\partial t^2}, \frac{\partial^2 u_z}{\partial t^2} \in L_2 \left(I, (W_2^1)^* \right), \quad \frac{\partial \varphi}{\partial t} \in L_2 \left(I, (\overset{0}{W}_2^1)^* \right),$$

where the star denotes a dual space, and $H = L_2(G)$.

This result is central to the construction of effective numerical schemes for the solution of problem (3.1)–(3.11). Indeed, two of the key elements of the proof of Theopiezoelectricity has been put on a rigorous mathematical basis. Theorem 4.1, is the application of the Faedo-Galerkin procedure and the obtaining the representation for the energy operator. The later may be derived from definition 4.1 by choosing appropriate trial functions η_1, η_2 and ζ in (4.1)–(4.3). We will not concentrate here on this procedure, and refer the reader to [22], where such a procedure was explained for one-dimensional models. The final result is as follows

$$\frac{d\mathcal{E}}{dt} = \int \int_{\Omega} r \left[\frac{\partial D_r}{\partial t} E_r + \frac{\partial D_z}{\partial t} E_z \right] d\Omega + \int_{R_0}^{R_1} r \left[\sigma_{rz} \frac{\partial u_r}{\partial t} + \sigma_z \frac{\partial u_z}{\partial t} \right] dr \Big|_{z_0}^{z_1} + \\ \int_{z_0}^{z_1} r \left[\sigma_r \frac{\partial u_r}{\partial t} + \sigma_{rz} \frac{\partial u_z}{\partial t} \right] dz \Big|_{R_0}^{R_1} + \int \int_{\Omega} r \left[f_1 \frac{\partial u_r}{\partial t} + f_2 \frac{\partial u_z}{\partial t} \right] d\Omega, \quad (4.4)$$

where the total inner energy of the system, $\mathcal{E} = K + W + P$, consists of the three coupled parts

- kinetic energy of the system

$$K = \frac{\rho}{2} \int \int_{\Omega} r \left\{ \left(\frac{\partial u_r}{\partial t} \right)^2 \left(\frac{\partial u_z}{\partial t} \right)^2 \right\} d\Omega,$$

- the energy of elastic deformation

$$W = \frac{1}{2} \int \int_{\Omega} r \left\{ c_{33} \epsilon_r^2 + c_{11} (\epsilon_\theta^2 + \epsilon_z^2) + 2c_{13} (\epsilon_\theta \epsilon_r + \epsilon_z \epsilon_r) + 2c_{12} \epsilon_z \epsilon_\theta + 2c_{44} \epsilon_{rz}^2 \right\} d\Omega$$

- and the energy of electric field

$$P = \frac{\epsilon_{33}}{2} \int \int_{\Omega} r E_r^2 d\Omega + \frac{\epsilon_{11}}{2} \int \int_{\Omega} r E_z^2 d\Omega.$$

The energy conservation law (4.4) derived for the coupled system described by the model (3.1)–(3.11), is fundamental to the investigation of system stability. One of the key factors of such investigations includes establishing some bounds on the energy functional $\mathcal{E}(t)$ at any given moment of time. Ultimately, it is such bounds that allow us to guarantee the stability of corresponding difference problems under certain constraints on time and space discretisations.

In the next section we use the energy conservation law (4.4) in order to construct effective difference schemes for the solution of problem (3.1)–(3.11).

5 Derivation of Numerical Schemes From the Conservation Laws for Coupled Systems.

Let $\tilde{u}_1 \equiv \tilde{u}_1(r, z, t)$, $\tilde{u}_2 \equiv \tilde{u}_2(r, z, t)$, $\tilde{\varphi} \equiv \tilde{\varphi}(r, z, t)$ be the functions of discrete variables $r \in \bar{\omega}_{h_1}$, $z \in \bar{\omega}_{h_2}$ and continuous variable $t \in \bar{I} \equiv \{t : 0 \leq t \leq T\}$, that give approximations to u_r , u_z and φ respectively. Then we denote

$$\begin{aligned} \tilde{u}_1^{(\pm 1_r)} &= \tilde{u}_1(r \pm h_1, z, t), \quad \tilde{u}_2^{(\pm 1_z)} = \tilde{u}_2(r, z \pm h_2, t), \quad \tilde{u}^{(\pm 1, \pm 1)} = \tilde{u}(r \pm h_1, z \pm h_2, t), \\ \tilde{\epsilon}_r &= \frac{1}{2} ((\tilde{u}_1)_r + (\tilde{u}_1^{(-1_r)})_r), \quad \tilde{\epsilon}_\theta = \frac{1}{4r} (\tilde{u}_1 + \tilde{u}_1^{(-1_r)} + \tilde{u}_1^{(-1_z)} + \tilde{u}_1^{(-1,-1)}), \\ \tilde{\epsilon}_z &= \frac{1}{2} ((\tilde{u}_2)_z + (\tilde{u}_2^{(-1_r)})_z), \quad 2\tilde{\epsilon}_{rz} = \frac{1}{2} ((\tilde{u}_1)_z + (\tilde{u}_1^{(-1_r)})_z + (\tilde{u}_2)_r + (\tilde{u}_2^{(-1_z)})_r), \\ \tilde{E}_r &= -\frac{1}{2} (\tilde{\varphi}_r + (\tilde{\varphi}^{(-1_z)})_r), \quad \tilde{E}_z = -\frac{1}{2} (\tilde{\varphi}_z + (\tilde{\varphi}^{(-1_r)})_z). \end{aligned}$$

Similar to the one-dimensional case (see details in [22]), the construction of the difference scheme for the solution of problem (3.1)–(3.11) is based on the obtaining a difference analogue of the energy conservation law. The procedure for the construction may be split into two stages [25, 21]

- first, using the method of approximation of quadratic functional of energy, we perform discretisation in space,
- then, the differential-difference scheme obtained on the first stage is discretised in time.

Below, we consider these stages in some detail.

The integral of kinetic energy is approximated by the following quadrature formula

$$K^h = \frac{\rho}{2} \sum_{\omega_h} r \hbar_1 \hbar_2 \left[\left(\frac{d\tilde{u}_1}{dt} \right)^2 \left(\frac{d\tilde{u}_2}{dt} \right)^2 \right] \quad (5.1)$$

(where $\hbar_\alpha = h_\alpha/2$ when $r \in \bar{\omega}_{h_\alpha}/\omega_{h_\alpha}$ and $\hbar_\alpha = h_\alpha$ when $r \in \omega_{h_\alpha}$, $\alpha = 1, 2$), and the integrals of elastic deformation and electric field by

$$\begin{aligned} W^h + P^h = \frac{1}{2} \sum_{\omega_h^+} \bar{r} h_1 h_2 & [c_{33} \tilde{\epsilon}_r^2 + c_{11} (\tilde{\epsilon}_\theta^2 + \tilde{\epsilon}_z^2) + 2c_{13} (\tilde{\epsilon}_\theta + \tilde{\epsilon}_z) \tilde{\epsilon}_r \\ & + 2c_{12} \tilde{\epsilon}_z \tilde{\epsilon}_\theta + 2c_{44} \tilde{\epsilon}_{rz}^2 + \epsilon_{33} \tilde{E}_r^2 + \epsilon_{11} \tilde{E}_z^2]. \end{aligned} \quad (5.2)$$

We note that if the solution of the problem (3.1)–(3.11) is from $[W_2^4(Q_T)]^2 \times L^2(I, W_2^4(G))$, then the quadrature formulas (5.1) and (5.2) are formulas of the second order accuracy in $|h| = (h_1^2 + h_2^2)^{1/2}$.

Taking into account differential-difference analogues of state equations

$$\begin{cases} \tilde{\sigma}_r = c_{33} \tilde{\epsilon}_r + c_{13} (\tilde{\epsilon}_\theta + \tilde{\epsilon}_z) - \epsilon_{33} \tilde{E}_r, \quad \tilde{\sigma}_\theta = c_{13} \tilde{\epsilon}_r + c_{11} \tilde{\epsilon}_\theta + c_{12} \tilde{\epsilon}_z - \epsilon_{13} \tilde{E}_r, \\ \tilde{\sigma}_z = c_{13} \tilde{\epsilon}_r + c_{12} \tilde{\epsilon}_\theta + c_{11} \tilde{\epsilon}_z - \epsilon_{13} \tilde{E}_r, \quad \tilde{\sigma}_{rz} = c_{44} \tilde{\epsilon}_{rz} - \epsilon_{15} \tilde{E}_z, \\ \tilde{D}_r = \epsilon_{33} \tilde{\epsilon}_r + c_{13} (\tilde{\epsilon}_\theta + \tilde{\epsilon}_z) + \tilde{\epsilon}_{33} \tilde{E}_r, \quad \tilde{D}_z = 2\epsilon_{15} \tilde{\epsilon}_{rz} + \tilde{\epsilon}_{11} \tilde{E}_z \end{cases} \quad (5.3)$$

we have the differential-difference analogue of (4.4) in the form

$$\begin{aligned} \frac{d\mathcal{E}^h}{dt} = \rho \sum_{\omega_h} r \hbar_1 \hbar_2 & \left[\frac{\partial \tilde{u}_1}{\partial t} \frac{\partial^2 \tilde{u}_1}{\partial t^2} + \frac{\partial \tilde{u}_2}{\partial t} \frac{\partial^2 \tilde{u}_2}{\partial t^2} \right] + \sum_{\omega_h^+} \bar{r} h_1 h_2 \left[\frac{\partial \tilde{\epsilon}_r}{\partial t} \tilde{\sigma}_r + \frac{\partial \tilde{\epsilon}_\theta}{\partial t} \tilde{\sigma}_\theta + \right. \\ & \left. \frac{\partial \tilde{\epsilon}_z}{\partial t} \tilde{\sigma}_z + \frac{\partial \tilde{\epsilon}_{rz}}{\partial t} (2\tilde{\sigma}_{rz}) \right] + \sum_{\omega_h^+} \bar{r} h_1 h_2 \left[\tilde{E}_r \frac{\partial \tilde{D}_r}{\partial t} + \tilde{E}_z \frac{\partial \tilde{D}_z}{\partial t} \right], \end{aligned} \quad (5.4)$$

where

$$\begin{aligned} \mathcal{E}^h = \frac{\rho}{2} \sum_{\omega_h} r \hbar_1 \hbar_2 & \left[\left(\frac{d\tilde{u}_1}{dt} \right)^2 \left(\frac{d\tilde{u}_2}{dt} \right)^2 \right] + \frac{1}{2} \sum_{\omega_h^+} \bar{r} h_1 h_2 [c_{33} \tilde{\epsilon}_r^2 + c_{11} (\tilde{\epsilon}_\theta^2 + \tilde{\epsilon}_z^2) + \\ & 2c_{13} (\tilde{\epsilon}_\theta + \tilde{\epsilon}_z) \tilde{\epsilon}_r + 2c_{12} \tilde{\epsilon}_z \tilde{\epsilon}_\theta + 2c_{44} \tilde{\epsilon}_{rz}^2 + \epsilon_{33} \tilde{E}_r^2 + \epsilon_{11} \tilde{E}_z^2]. \end{aligned} \quad (5.5)$$

Applying grid formulas for summation by parts [28, 29] from (5.4) we derive the differential-difference analogue of the conservation law (4.4)

$$\begin{aligned} \frac{d\mathcal{E}^h}{dt} = \sum_{\omega_h} r \hbar_1 \hbar_2 & \left[f_1 \frac{\tilde{u}_1}{\partial t} + f_2 \frac{\tilde{u}_2}{\partial t} \right] + \sum_{\omega_{h_1}} r \hbar_1 \left(\tilde{\sigma}_{rz} \frac{\tilde{u}_1}{\partial t} + \tilde{\sigma}_z \frac{\tilde{u}_2}{\partial t} \right) \Big|_{j=0}^{j=M} + \\ & \sum_{\omega_{h_2}} r \hbar_2 \left(\tilde{\sigma}_r \frac{\tilde{u}_1}{\partial t} + \tilde{\sigma}_{rz} \frac{\tilde{u}_2}{\partial t} \right) \Big|_{i=0}^{i=N} + \sum_{\omega_h} r \hbar_1 \hbar_2 \frac{\partial f_3}{\partial t}. \end{aligned} \quad (5.6)$$

Now, using (5.6), we can easily derive differential-difference analogues of the equations for motion of the electroelastic medium, the Maxwell equation, and *natural* boundary conditions (3.9), (3.10).

The reader may consult [22], where such a procedure were explained for one-dimensional problems, for details.

The second stage of the construction of difference schemes requires the time discretisation. As a result of such discretisation, we obtain the following difference scheme:

$$\begin{cases} \rho y_{it} = \Lambda_1(y, g, \mu) + F_1, \\ \rho g_{it} = \Lambda_2(y, g, \mu) + F_2, \\ \Lambda_3(y, g, \mu) = F_3, \end{cases} \quad (5.7)$$

where functions y , g and μ are discrete-argument functions that give approximations to the functions $u_r(r, z, t)$, $u_z(r, z, t)$ and $\varphi(r, z, t)$ respectively. The difference operators Λ_i , $i = 1, 2, 3$ and right hand sides F_i , $i = 1, 2, 3$ in (5.7) are defined as follows

$$\Lambda_1(y, g, \mu) = \begin{cases} \frac{1}{r} \left(\frac{\bar{\sigma}_r + \bar{\sigma}_r^{(+1_z)}}{2} \right)_r + \frac{1}{r} \left(\bar{r} \bar{\sigma}_{rz} + \bar{r}^{(+1)} \bar{\sigma}_{rz}^{(+1_r)} \right)_z - \\ \frac{\bar{\sigma}_\theta + \bar{\sigma}_\theta^{(+1_r)} + \bar{\sigma}_\theta^{(+1_z)} + \bar{\sigma}_\theta^{(+1,+1)}}{4r}, & (r, z) \in \omega_h, \\ \frac{1}{r} \left(\bar{r} \bar{\sigma}_r^{(+1_z)} \right)_r + \frac{1}{r} \frac{2}{h_2} \left(\frac{\bar{r} \bar{\sigma}_{rz}^{(+1_z)} + \bar{r}^{(+1)} \bar{\sigma}_{rz}^{(+1,+1)}}{2} \right) - \\ \frac{\bar{\sigma}_\theta^{(+1_z)} + \bar{\sigma}_\theta^{(+1,+1)}}{2r}, & (r, z) \in \gamma_1, \\ \frac{1}{r} \left(\bar{r} \bar{\sigma}_r \right)_r - \frac{1}{r} \frac{2}{h_2} \left(\frac{\bar{r} \bar{\sigma}_{rz} + \bar{r}^{(+1)} \bar{\sigma}_{rz}^{(+1_r)}}{2} \right) - \frac{\bar{\sigma}_\theta + \bar{\sigma}_\theta^{(+1_r)}}{2r}, & (r, z) \in \gamma_2, \\ \frac{1}{r} \bar{r}^{(+1)} \left(\bar{\sigma}_{rz}^{(+1_r)} \right)_z + \frac{1}{r} \frac{2}{h_1} \bar{r}^{(+1)} \left(\frac{\bar{\sigma}_r^{(+1_r)} + \bar{\sigma}_r^{(+1,+1)}}{2} \right) - \\ \frac{\bar{\sigma}_\theta^{(+1_r)} + \bar{\sigma}_\theta^{(+1,+1)}}{2r}, & (r, z) \in \gamma_3, \\ \frac{1}{r} \bar{r} \left(\bar{\sigma}_{rz} \right)_z - \frac{1}{r} \frac{2}{h_1} \bar{r} \left(\frac{\bar{\sigma}_r + \bar{\sigma}_r^{(+1_z)}}{2} \right) - \frac{\bar{\sigma}_\theta + \bar{\sigma}_\theta^{(+1_z)}}{2r}, & (r, z) \in \gamma_4, \\ \frac{1}{r} \frac{2}{h_1} \bar{r}^{(+1)} \bar{\sigma}_r^{(+1,+1)} + \frac{1}{r} \frac{2}{h_2} \bar{r}^{(+1)} \bar{\sigma}_{rz}^{(+1,+1)} - \frac{\bar{\sigma}_\theta^{(+1,+1)}}{r}, & (r, z) \in \gamma_{13}, \\ \frac{1}{r} \frac{2}{h_1} \bar{r}^{(+1)} \bar{\sigma}_r^{(+1_r)} - \frac{1}{r} \frac{2}{h_2} \bar{r}^{(+1)} \bar{\sigma}_{rz}^{(+1_r)} - \frac{\bar{\sigma}_\theta^{(+1_r)}}{r}, & (r, z) \in \gamma_{23}, \\ -\frac{1}{r} \frac{2}{h_1} \bar{r} \bar{\sigma}_r^{(+1_z)} + \frac{1}{r} \frac{2}{h_2} \bar{r} \bar{\sigma}_{rz}^{(+1_r)} - \frac{\bar{\sigma}_\theta^{(+1_z)}}{r}, & (r, z) \in \gamma_{14}, \\ -\frac{1}{r} \frac{2}{h_1} \bar{r} \bar{\sigma}_r - \frac{1}{r} \frac{2}{h_2} \bar{r} \bar{\sigma}_{rz} - \frac{\bar{\sigma}_\theta}{r}, & (r, z) \in \gamma_{24}, \end{cases}$$

$$\Lambda_2(y, g, \mu) = \begin{cases} \frac{1}{r} \left(\bar{r} \frac{\bar{\sigma}_{rz} + \bar{\sigma}_{rz}^{(+1_z)}}{2} \right)_r + \frac{1}{r} \left(\bar{r} \bar{\sigma}_z + \bar{r}^{(+1)} \bar{\sigma}_z^{(+1_r)} \right)_z, & (r, z) \in \omega_h, \\ \frac{1}{r} \left(\bar{r} \bar{\sigma}_{rz}^{(+1_z)} \right)_r + \frac{1}{r} \frac{2}{h_2} \left(\frac{\bar{r} \bar{\sigma}_z^{(+1_z)} + \bar{r}^{(+1)} \bar{\sigma}_z^{(+1,+1)}}{2} \right), & (r, z) \in \gamma_1, \\ \frac{1}{r} \left(\bar{r} \bar{\sigma}_{rz} \right)_r - \frac{1}{r} \frac{2}{h_2} \left(\frac{\bar{r} \bar{\sigma}_z + \bar{r}^{(+1)} \bar{\sigma}_z^{(+1_r)}}{2} \right) - \frac{\bar{\sigma}_\theta^+ \bar{\sigma}_\theta^{(+1_r)}}{2r}, & (r, z) \in \gamma_2, \\ \frac{1}{r} \bar{r}^{(+1)} \left(\bar{\sigma}_z^{(+1_r)} \right)_z + \frac{1}{r} \frac{2}{h_1} \bar{r}^{(+1)} \left(\frac{\bar{\sigma}_{rz}^{(+1_z)} + \bar{\sigma}_{rz}^{(+1,+1)}}{2} \right), & (r, z) \in \gamma_3, \\ \frac{1}{r} \bar{r} \left(\bar{\sigma}_z \right)_z - \frac{1}{r} \frac{2}{h_1} \bar{r} \left(\frac{\bar{\sigma}_{rz} + \bar{\sigma}_{rz}^{(+1_z)}}{2} \right), & (r, z) \in \gamma_4, \\ \frac{1}{r} \frac{2}{h_2} \bar{r}^{(+1)} \bar{\sigma}_z^{(+1,+1)} + \frac{1}{r} \frac{2}{h_1} \bar{r}^{(+1)} \bar{\sigma}_{rz}^{(+1,+1)}, & (r, z) \in \gamma_{13}, \\ \frac{1}{r} \frac{2}{h_1} \bar{r}^{(+1)} \bar{\sigma}_{rz}^{(+1_z)} - \frac{1}{r} \frac{2}{h_2} \bar{r}^{(+1)} \bar{\sigma}_z^{(+1_r)}, & (r, z) \in \gamma_{23}, \\ -\frac{1}{r} \frac{2}{h_1} \bar{r} \bar{\sigma}_{rz}^{(+1_z)} + \frac{1}{r} \frac{2}{h_2} \bar{r} \bar{\sigma}_z^{(+1_z)}, & (r, z) \in \gamma_{14}, \\ -\frac{1}{r} \frac{2}{h_1} \bar{r} \bar{\sigma}_{rz} - \frac{1}{r} \frac{2}{h_2} \bar{r} \bar{\sigma}_z, & (r, z) \in \gamma_{24}, \end{cases}$$

$$\Lambda_3(y, g, \mu) = \begin{cases} \frac{1}{r} \left(\bar{r} \frac{\bar{D}_r + \bar{D}_r^{(+1_z)}}{2} \right)_r + \frac{1}{r} \left(\bar{r} \bar{D}_z + \bar{r}^{(+1)} \bar{D}_z^{(+1_r)} \right)_z, & (r, z) \in \omega_h, \\ \frac{1}{r} \left(\bar{r} \bar{D}_r^{(+1_z)} \right)_r + \frac{1}{r} \frac{2}{h_2} \left(\frac{\bar{r} \bar{D}_z^{(+1_z)} + \bar{r}^{(+1)} \bar{D}_z^{(+1,+1)}}{2} \right), & (r, z) \in \gamma_1, \\ \frac{1}{r} \left(\bar{r} \bar{D}_r \right)_r - \frac{1}{r} \frac{2}{h_2} \left(\frac{\bar{r} \bar{D}_z + \bar{r}^{(+1)} \bar{D}_z^{(+1_r)}}{2} \right), & (r, z) \in \gamma_2, \\ \mu, & (r, z) \in \bar{\omega}_h / (\omega_h \cup \gamma_1 \cup \gamma_2), \end{cases}$$

$$F_3 = \begin{cases} f_3, & (r, z) \in \omega_h \cup \gamma_1 \cup \gamma_2, \\ 0 & (r, z) \in \bar{\omega}_h / (\omega_h \cup \gamma_1 \cup \gamma_2), \end{cases}$$

$$F_1 = f_1 + \begin{cases} 0, & (r, z) \in \omega_h, \\ -\frac{2}{h_2} p_{rt}^{(0)}, & (r, z) \in \gamma_1, \\ \frac{2}{h_2} p_{rt}^{(1)}, & (r, z) \in \gamma_2, \\ -\frac{2}{h_1} p_r^{(0)}, & (r, z) \in \gamma_3, \\ \frac{2}{h_1} p_{rt}^{(1)}, & (r, z) \in \gamma_4, \\ -\frac{2}{h_1} p_r^{(0)} - \frac{2}{h_2} p_{rt}^{(0)}, & (r, z) \in \gamma_{13}, \\ -\frac{2}{h_1} p_r^{(0)} + \frac{2}{h_2} p_{rt}^{(1)}, & (r, z) \in \gamma_{23}, \\ \frac{2}{h_1} p_r^{(1)} - \frac{2}{h_2} p_{rt}^{(0)}, & (r, z) \in \gamma_{14}, \\ \frac{2}{h_1} p_r^{(1)} + \frac{2}{h_2} p_{rt}^{(1)}, & (r, z) \in \gamma_{24}, \end{cases} \quad F_2 = f_2 + \begin{cases} 0, & (r, z) \in \omega_h, \\ -\frac{2}{h_2} p_z^{(0)}, & (r, z) \in \gamma_1, \\ \frac{2}{h_2} p_z^{(1)}, & (r, z) \in \gamma_2, \\ -\frac{2}{h_1} p_{zt}^{(0)}, & (r, z) \in \gamma_3, \\ \frac{2}{h_1} p_{zt}^{(1)}, & (r, z) \in \gamma_4, \\ -\frac{2}{h_1} p_{zt}^{(0)} - \frac{2}{h_2} p_z^{(0)}, & (r, z) \in \gamma_{13}, \\ -\frac{2}{h_1} p_{zt}^{(0)} + \frac{2}{h_2} p_z^{(1)}, & (r, z) \in \gamma_{23}, \\ \frac{2}{h_1} p_{zt}^{(1)} - \frac{2}{h_2} p_z^{(0)}, & (r, z) \in \gamma_{14}, \\ \frac{2}{h_1} p_{zt}^{(1)} + \frac{2}{h_2} p_z^{(1)}, & (r, z) \in \gamma_{24}. \end{cases}$$

The fully-discrete approximation of the state equations follow immediately from (5.3)

$$\begin{cases} \bar{\sigma}_r = c_{33}\bar{\epsilon}_r + c_{13}(\bar{\epsilon}_\theta + \bar{\epsilon}_z) - e_{33}\bar{E}_r, & \bar{\sigma}_\theta = c_{13}\bar{\epsilon}_r + c_{11}\bar{\epsilon}_\theta + c_{12}\bar{\epsilon}_z - e_{13}\bar{E}_r, \\ \bar{\sigma}_z = c_{13}\bar{\epsilon}_r + c_{12}\bar{\epsilon}_\theta + c_{11}\bar{\epsilon}_z - e_{13}\bar{E}_r, & \bar{\sigma}_{rz} = c_{44}\bar{\epsilon}_{rz} - e_{15}\bar{E}_z, \\ \bar{D}_r = \bar{E}_r + e_{33}\bar{\epsilon}_r + e_{13}(\bar{\epsilon}_\theta + \bar{\epsilon}_z), & \bar{D}_z = \epsilon_{11}\bar{E}_z + 2e_{15}\bar{\epsilon}_{rz}, \end{cases} \quad (5.8)$$

where

$$\begin{aligned} \bar{E}_r &= \frac{1}{2} (\mu_r + \mu_r^{(-1_z)}), \quad \bar{E}_z = \frac{1}{2} (\mu_z + \mu_z^{(-1_r)}), \\ \bar{\epsilon}_r &= \frac{1}{2} (y_r + y_r^{(-1_z)}), \quad \bar{\epsilon}_\theta = \frac{1}{4\bar{r}} (y + y^{(-1_r)} + y^{(-1_z)} + y^{(-1,-1)}), \\ \bar{\epsilon}_z &= \frac{1}{2} (g_z + g_z^{(-1_r)}), \quad 2\bar{\epsilon}_{rz} = \frac{1}{2} (y_z + y_z^{(-1_r)} + g_r + g_r^{(-1_z)}). \end{aligned}$$

The first pair of initial conditions is approximated exactly

$$y(r, z, 0) = u_r^{(0)}(r, z), \quad g(r, z, 0) = u_z^{(0)}(r, z). \quad (5.9)$$

The second pair of initial conditions is approximated by the central difference derivative with subsequent elimination of the (-1)st fictitious time-layer by using the first two equation of the system (5.7) for $t = 0$. The result is

$$\rho y_t = \rho u_r^{(1)} + \frac{\tau}{2} (F_1 + \Lambda_1(y, g, \mu)), \quad \rho g_t = \rho u_z^{(1)} + \frac{\tau}{2} (F_2 + \Lambda_2(y, g, \mu)). \quad (5.10)$$

6 Bound on the Energy Functional in Nonstationary Coupled Models.

When expressed as a finite set of equations, mathematical conservation laws may be only an approximate reflection of the real-world situation. In order to guarantee an adequateness of such a reflection, it is important to obtain an upper bound estimate for the energy functional. On the one hand, such estimates may be used as a measure of model applicability to specific practical problems. On the other hand, such estimates play a fundamental role in investigating system stability. In this section we derive such an estimate from the energy balance equation (4.4).

Integrating (4.4) with respect to t from zero to t_1 ($0 < t_1 \leq T$) and taking into account conditions (3.9), (3.10) we have

$$\begin{aligned} \mathcal{E}(t_1) = & \mathcal{E}(0) + \int_0^{t_1} \int \int_{\Omega} r \left[\frac{\partial D_r}{\partial t} E_r + \frac{\partial D_z}{\partial t} E_z \right] d\Omega dt + \int_0^{t_1} \int_{R_0}^{R_1} r \left[\frac{\partial u_r}{\partial t} \Big|_{Z_1} p_{rt}^{(1)} - \right. \\ & \left. \frac{\partial u_r}{\partial t} \Big|_{Z_0} p_{rt}^{(0)} \right] dr dt + \int_0^{t_1} \int_{R_0}^{R_1} r \left[\frac{\partial u_z}{\partial t} \Big|_{Z_1} p_z^{(1)} - \frac{\partial u_z}{\partial t} \Big|_{Z_0} p_z^{(0)} \right] dr dt + \\ & \int_0^{t_1} \int_{Z_0}^{Z_1} r \left[\frac{\partial u_r}{\partial t} \Big|_{R_1} p_r^{(1)} - \frac{\partial u_r}{\partial t} \Big|_{R_0} p_r^{(0)} \right] dz dt + \int_0^{t_1} \int_{Z_0}^{Z_1} r \left[\frac{\partial u_z}{\partial t} \Big|_{R_1} p_{zt}^{(1)} - \right. \\ & \left. \frac{\partial u_z}{\partial t} \Big|_{R_0} p_{zt}^{(0)} \right] dz dt + \int_0^{t_1} \int \int_{\Omega} \left[f_1 \frac{\partial u_r}{\partial t} + f_2 \frac{\partial u_z}{\partial t} \right] d\Omega dt. \end{aligned} \quad (6.1)$$

Now we have to estimate additives in the right hand side of (6.1). As an example, we consider

$$\begin{aligned} I_1 = & \int_0^{t_1} \int_{R_0}^{R_1} r \left[\frac{\partial u}{\partial t} \Big|_{Z_1} p_{rt}^{(1)} - \frac{\partial u}{\partial t} \Big|_{Z_0} p_{rt}^{(0)} \right] dr dt = \int_{R_0}^{R_1} r \left\{ [p_{rt}^{(1)}(r, t_1) u_r(r, Z_1, t_1) - \right. \\ & \left. p_{rt}^{(1)}(r, 0) u_r(r, Z_1, 0)] - [p_{rt}^{(0)}(r, t_1) u_r(r, Z_0, t_1) - p_{rt}^{(0)}(r, 0) u_r(r, Z_0, 0)] \right\} dr + \\ & \int_0^{t_1} \int_{R_0}^{R_1} r \left\{ u_r(r, Z_0, t) \frac{\partial p_{rt}^{(0)}}{\partial t} - u_r(r, Z_1, t) \frac{\partial p_{rt}^{(1)}}{\partial t} \right\} dr dt = I_1^{(1)} + I_1^{(2)}, \end{aligned} \quad (6.2)$$

where

$$\begin{aligned} I_1^{(1)} = & \int_{R_0}^{R_1} r \left\{ [p_{rt}^{(1)}(r, t_1) u_r(r, Z_1, t_1) - p_{rt}^{(1)}(r, 0) u_r(r, Z_1, 0)] - \right. \\ & \left. [p_{rt}^{(0)}(r, t_1) u_r(r, Z_0, t_1) - p_{rt}^{(0)}(r, 0) u_r(r, Z_0, 0)] \right\} dr \\ I_1^{(2)} = & \int_0^{t_1} \int_{R_0}^{R_1} r \left\{ u_r(r, Z_0, t) \frac{\partial p_{rt}^{(0)}}{\partial t} - u_r(r, Z_1, t) \frac{\partial p_{rt}^{(1)}}{\partial t} \right\} dr dt. \end{aligned}$$

Using the Cauchy-Schwartz inequality and the ϵ -inequality [28, 29, 22] we get

$$\begin{aligned} I_1^{(1)} \leq & M_5^{(1)} \int_{R_0}^{R_1} [|u_r(r, Z_0, t_1)|^2 + |u_r(r, Z_1, t_1)|^2] dr + \\ & M_5^{(2)} \int_{R_0}^{R_1} [|u_r(r, Z_0, 0)|^2 + |u_r(r, Z_1, 0)|^2] dr + \\ & M_1^{(3)} \int_{R_0}^{R_1} [|p_{rt}^{(0)}(r, t_1)|^2 + |p_{rt}^{(1)}(r, t_1)|^2 + |p_{rt}^{(0)}(r, 0)|^2 + |p_{rt}^{(1)}(r, 0)|^2] dr. \end{aligned} \quad (6.3)$$

We apply the functional analysis technique based on embedding theorems [28, 32] in order to obtain estimates of the first two terms in the RHS of (6.3):

$$\begin{aligned} \int_{R_0}^{R_1} [|u_r(r, Z_0, t_1)|^2 + |u_r(r, Z_1, t_1)|^2] dr &\leq M_6^{(1)} \int \int_{\Omega} \left(\frac{\partial u_r}{\partial z} \right)^2 \Big|_{t=t_1} d\Omega, \\ \int_{R_0}^{R_1} [|u_r(r, Z_0, 0)|^2 + |u_r(r, Z_1, 0)|^2] dr &\leq M_6^{(2)} \int \int_{\Omega} \left(\frac{\partial u_r}{\partial z} \right)^2 \Big|_{t=0} d\Omega. \end{aligned}$$

Let $M_5^{(i)} M_6^{(i)} = M_1^{(i)}$, $i = 1, 2$. Then from (6.3) we get

$$\begin{aligned} I_1^{(1)} &\leq M_1^{(1)} \int \int_{\Omega} \left(\frac{\partial u_r}{\partial z} \right)^2 \Big|_{t=t_1} d\Omega + M_1^{(2)} \int \int_{\Omega} \left(\frac{\partial u_r}{\partial z} \right)^2 \Big|_{t=0} d\Omega + \\ &M_1^{(3)} \int R_0^{R_1} [|p_{rt}^{(0)}(r, t_1)|^2 + |p_{rt}^{(1)}(r, t_1)|^2 + |p_{rt}^{(0)}(r, 0)|^2 + |p_{rt}^{(1)}(r, 0)|^2] dr. \end{aligned} \quad (6.4)$$

In a similar way, we transform the other three additives of the right hand side of (6.1).

Estimates of integrals of $I_1^{(2)}$ -type can also be performed by applying the ϵ -inequality and embedding theorems. For example, it is not difficult to obtain that

$$I_1^{(2)} \leq M_1^{(4)} \int_0^{t_1} \left(\frac{\partial u_r}{\partial z} \right)^2 dt + M_1^{(5)} \int_0^{t_1} \int_{R_0}^{R_1} \left[\left(\frac{\partial p_{rt}^{(0)}}{\partial t} \right)^2 + \left(\frac{\partial p_{rt}^{(1)}}{\partial t} \right)^2 \right] dr dt. \quad (6.5)$$

We take into account (6.4), (6.5) in estimating (6.1) and apply the following lemma on integral inequality:

Lemma 6.1 [11]. *Let function $I(t) \geq 0$ ($0 \leq t \leq T$) is continuous, differentiable and such that it satisfies the following inequality*

$$I(t_2) \leq I(t_1) + M \int_{t_1}^{t_2} I(t) dt + N \int_{t_1}^{t_2} \sqrt{I(t)} dt \text{ for any } 0 \leq t_1 \leq t_2 \leq T \text{ ($M > 0$, $N \geq 0$)}.$$

Then

$$\sqrt{I(t)} \leq \sqrt{I(0)} \exp \left(\frac{M}{2} t \right) + \frac{N}{M} \left(\exp \left(\frac{M}{2} t \right) - 1 \right).$$

As a result, from (6.1) we derive the following estimate

$$\begin{aligned} \mathcal{E}(t_1) &\leq M_1 \mathcal{E}(0) + M_2 \left\{ \int_{R_0}^{R_1} \left[\sum_{i,j=0}^1 (|p_{rt}^{(i)}(r, t_j)|^2 + |p_{zt}^{(i)}(r, t_j)|^2) \right] dr + \right. \\ &\int_{Z_0}^{Z_1} \left[\sum_{i,j=0}^1 (|p_r^{(i)}(r, t_j)|^2 + |p_{zt}^{(i)}(r, t_j)|^2) \right] dz + \int_0^{t_1} \int_{R_0}^{R_1} \sum_{i=0}^1 \left[\left(\frac{\partial p_{rt}^{(i)}}{\partial t} \right)^2 + \left(\frac{\partial p_z^{(i)}}{\partial t} \right)^2 \right] dr dt + \\ &\left. \int_0^{t_1} \int_{Z_0}^{Z_1} \sum_{i=0}^1 \left[\left(\frac{\partial p_r^{(i)}}{\partial t} \right)^2 + \left(\frac{\partial p_{zt}^{(i)}}{\partial t} \right)^2 \right] dz dt + \int_0^{t_1} \int \int_{\Omega} r(f_1^2 + f_2^2) \right\} d\Omega dt, \end{aligned} \quad (6.6)$$

where the time variable is taken at $t = t_1$ if $j = 1$ and at $t = t_0 = 0$ if $j = 0$. Note that in inequality (6.6) we have taken into account that

$$\frac{\partial D_r}{\partial t} = \frac{\partial D_z}{\partial t} = 0$$

(see [22] for details).

Now we have to estimate the total energy of the electro-mechanical system at the initial moment of time, using initial conditions (3.7), (3.8). The main difficulty (as it was in the one-dimensional case), consists of obtaining an estimate of electric-field-energy integral.

First, let us estimate the quantity

$$\int \int_{\Omega} r [E_r^2 + E_z^2] \Big|_{t=0} d\Omega.$$

It is straightforward to get the following identity

$$\int \int_{\Omega} r [D_r E_r + D_z E_z] d\Omega = \int \int_{\Omega} r \lambda (E_r + E_z) d\Omega, \quad (6.7)$$

where the quantity λ is defined by the relationships

$$\frac{\partial \lambda}{\partial r} + \frac{\partial \lambda}{\partial z} = f_3, \quad \lambda(R_0, z, t) = \lambda(r, Z_0, t) = 0. \quad (6.8)$$

Taking into account state equations (3.4) from (6.7) we obtain

$$\begin{aligned} \min\{c_1, c_2\} \int \int_{\Omega} r (E_r^2 + E_z^2) \Big|_{t=0} d\Omega &\leq c_1 \int \int_{\Omega} r E_r^2 \Big|_{t=0} d\Omega + \\ c_2 \int \int_{\Omega} r E_z^2 \Big|_{t=0} d\Omega &\leq M_3 \int \int_{\Omega} r [\lambda^2 + \epsilon_r^2 + \epsilon_\theta^2 + \epsilon_z^2 + \epsilon_{rz}^2] \Big|_{t=0} d\Omega, \end{aligned}$$

where c_1, c_2 are constants that depend on properties of the electro-elastic medium. Therefore

$$\begin{aligned} \mathcal{E}(0) &\leq \frac{\rho}{2} \int \int_{\Omega} r [(u_r^{(1)})^2 + (u_z^{(1)})^2] d\Omega + \frac{1}{2} \int \int_{\Omega} r [c_{33}\epsilon_r^2 + c_{11}(\epsilon_\theta^2 + \epsilon_z^2) + \\ &2c_{13}(\epsilon_\theta\epsilon_r + \epsilon_z\epsilon_r) + 2c_{12}\epsilon_z\epsilon_\theta + 2c_{44}\epsilon_{rz}^2] d\Omega + \frac{1}{2} \max\{\epsilon_{11}, \epsilon_{33}\} \times \\ &\frac{M_3}{\min\{c_1, c_2\}} \int \int_{\Omega} r [\lambda^2 + \epsilon_r^2 + \epsilon_\theta^2 + \epsilon_z^2 + \epsilon_{rz}^2] d\Omega, \end{aligned} \quad (6.9)$$

where all additives in the right hand side of (6.9) are computed for $t = 0$.

Finally, if we take into account the condition of non-negativity of the potential energy of deformation,

$$\delta_1 \sum_{i=1}^4 \xi_i^2 \leq c_{33}\xi_1^2 + c_{11}(\xi_2^2 + \xi_3^2) + 2c_{13}(\xi_2\xi_1 + \xi_3\xi_1) + 2c_{12}\xi_3\xi_2 + 2c_{44}\xi_4^2, \quad \delta_1 > 0, \quad (6.10)$$

then from (6.6) we obtain an upper bound on the energy functional for the problem (3.1)–(3.11). Hence we have proved the following theorem.

Theorem 6.1 If the condition (6.10) holds, then the solution of the problem (3.1)–(3.11) satisfies the following energy bound

$$\begin{aligned}
\mathcal{E}(t_1) \leq M \left\{ & \rho \int \int_{\Omega} r [(u_r^{(1)})^2 + (u_z^{(1)})^2] d\Omega + \int \int_{\Omega} r [c_{33}\epsilon_r^2 + c_{11}(\epsilon_\theta^2 + \epsilon_z^2) + \right. \\
& 2c_{13}(\epsilon_\theta + \epsilon_z)\epsilon_r + 2c_{12}\epsilon_z\epsilon_\theta + 2c_{44}\epsilon_{rz}^2] \Big|_{t=0} d\Omega + \int_{R_0}^{R_1} \left[\sum_{i,j=0}^1 (|p_{rt}^{(i)}(r, t_j)|^2 + \right. \\
& |p_z^{(i)}(r, t_j)|^2) \Big] dr + \int_{Z_0}^{Z_1} \left[\sum_{i,j=0}^1 (|p_r^{(i)}(z, t_j)|^2 + |p_{zt}^{(i)}(z, t_j)|^2) \right] dz + \\
& \int_0^{t_1} \int_{R_0}^{R_1} \sum_{i=0}^1 \left[\left(\frac{\partial p_{rt}^{(i)}}{\partial t} \right)^2 + \left(\frac{\partial p_z^{(i)}}{\partial t} \right)^2 \right] dr dt + \int_0^{t_1} \int_{Z_0}^{Z_1} \sum_{i=0}^1 \left[\left(\frac{\partial p_r^{(i)}}{\partial t} \right)^2 + \left(\frac{\partial p_{zt}^{(i)}}{\partial t} \right)^2 \right] dz dt + \\
& \left. \int \int_{\Omega} r \lambda^2 \Big|_{t=0} d\Omega + \int_0^{t_1} \int \int_{\Omega} r (f_1^2 + f_2^2) d\Omega dt \right\}, \tag{6.11}
\end{aligned}$$

where $\mathcal{E}(t)$ is the total energy of the electro-mechanical system at time t , and λ is defined by the relationships (6.8).

Theorem 6.1 provides the basis for the investigation of stability of the electromechanical system. We will address the stability issues elsewhere.

7 Numerical Experiments

The important direction of piezoceramic application is connected with improved resolution and miniaturisation of technical devices. In such areas as biomedical imaging, nondestructive evaluation and hydroacoustics, devices produced from hollow piezoceramic cylinders or sphere may prove to be very useful [9, 22].

We consider the results of modelling oscillations of hollow PZT-piezoceramic cylinders in the case when the potential difference $2V=1$ is maintained on the “end-wall” of cylinders.

- We note that, when the ratio of the cylinder thickness (l) to its height (H) is small ($l/H < R_1/H = a$), the results of computation with model (3.1)–(3.11) are in good agreement with the results for “infinitely”-long cylinders. Such results were obtained earlier with the one-dimensional model (see [22, 24]).
- One of the characteristics of practical interest is the dynamic of radial displacements on the external surface of cylinders. We conclude that, for thin cylinders poled radially, the magnitude of such displacements grows considerably faster (with the decrease of thickness) than for cylinders poled circularly. When thickness becomes larger, the magnitude of displacements of cylinders with circular preliminary polarisation typically exceeds the magnitude of displacements for cylinders with radial preliminary polarisation. As an example we present the results of computations for a thick cylinder with $l=0.9$ (Figure 1). Other examples may be found in [21].
- For finite-length cylinders, the use of two-dimensional models is essential. The error in computation of displacements on the external surface of cylinders with $l=0.13$ and $H=1$ may exceed 27% for cylinders with radial preliminary polarisation and 16% for cylinders with circular preliminary polarisation.

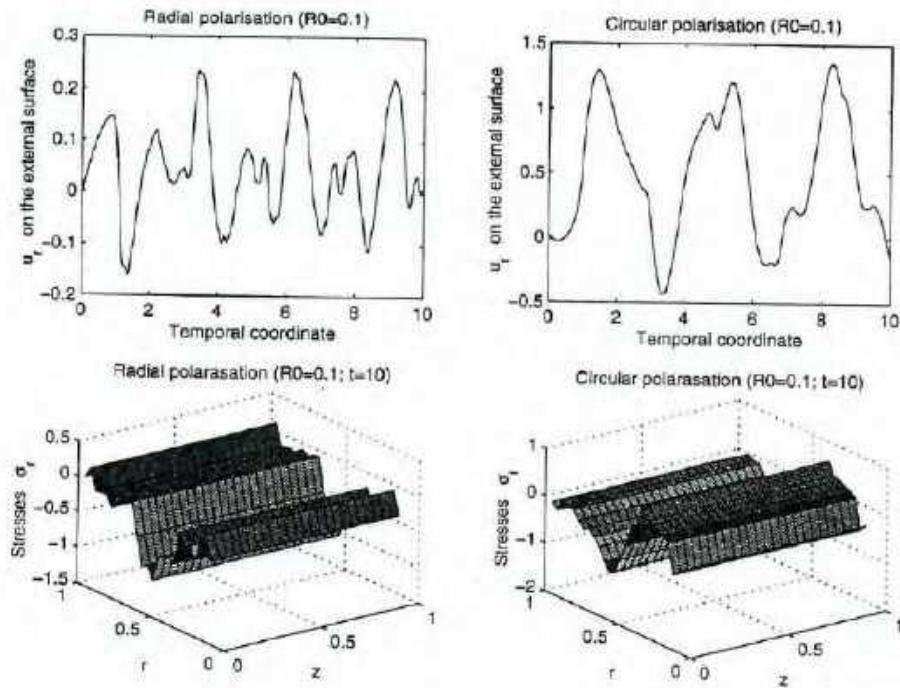


Figure 1: Displacements and radial stresses in an “infinitely-long” thick cylinder ($l=0.9$, $a=0.001$).

- Figures 2 and 3 present the results of computations for cylinders with radial preliminary polarisation ($l=0.9$). We compare these results for the same cylinder poled circularly (see Figures 4 and 5). We observe, that for thick cylinders, the circular preliminary polarisation produces larger amplitudes of oscillations on the external surfaces of cylinders. Axial stresses in cylinders poled circularly may essentially exceed axial stresses in cylinders with radial preliminary polarisation.
- For thin cylinders, even the qualitative picture becomes completely different (see Figures 6, 7 for results on cylinders poled radially and Figures 8, 9 for results on cylinders with circular preliminary polarisation). In the thin-cylinder case, both characteristics, displacements and stresses, are essentially larger for cylinders with radial preliminary polarisation. The results on quantitative estimates of such characteristics are an important prerequisite for the successful design of acoustic radiators, transducers, sensors and many other technical devices constructed from piezoceramic.

8 Concluding remarks.

Compared to pure elastic situations, the coupling elastic and electric fields, as well as anisotropy of physical properties of piezoelectric materials (for example, elastic, piezoelectric and dielectric moduli of piezoceramic), essentially complicates the analysis of wave phenomena that take place in piezoelectric bodies. Even the one-dimensional problem in the non-stationary case, presents a mathematically challenging and physically important problem. In order to obtain a plausible picture of the coupling phenomenon, methods for the solution of dynamical problems of electroelasticity may not always be based on thickness averaging of mechanical components of electro-elastic fields and the subsequent use of the Kirchoff-Lave-type hypothesis. Such averaging techniques, successful in the theory of plates and shells applied to elastic bodies, may not provide an appropriate tool in

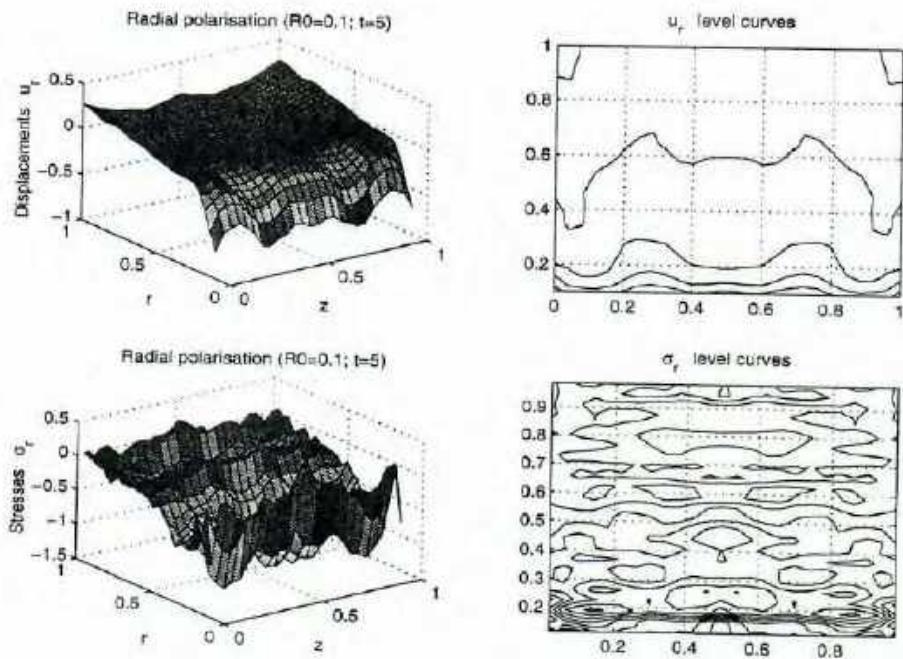


Figure 2: Displacements and radial stresses in a finite-length thick cylinder with radial preliminary polarisation ($l=0.9$, $a=1$).

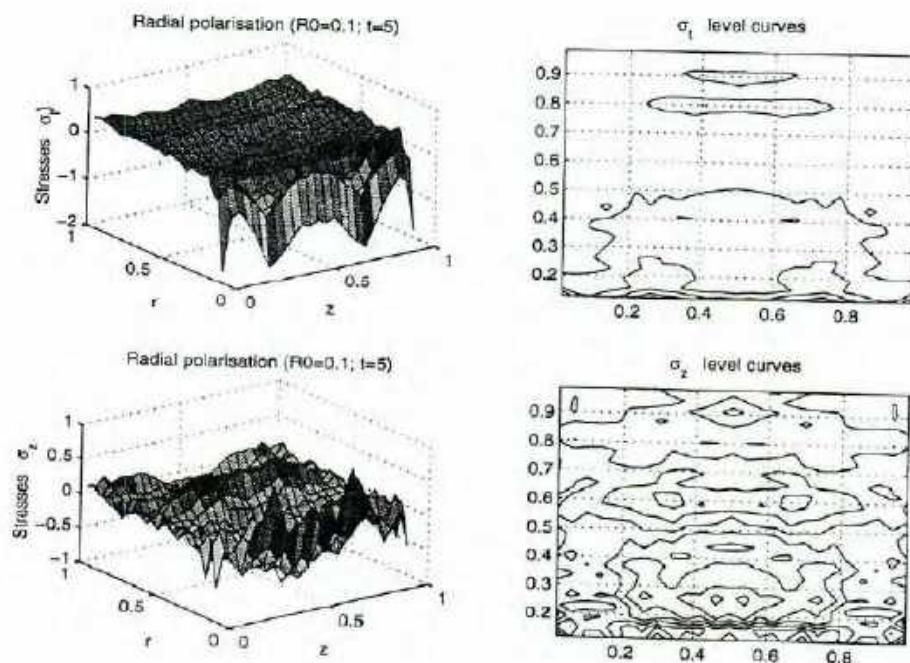


Figure 3: Stresses in a finite-length thick cylinder with radial preliminary polarisation ($l=0.9$, $a=1$).

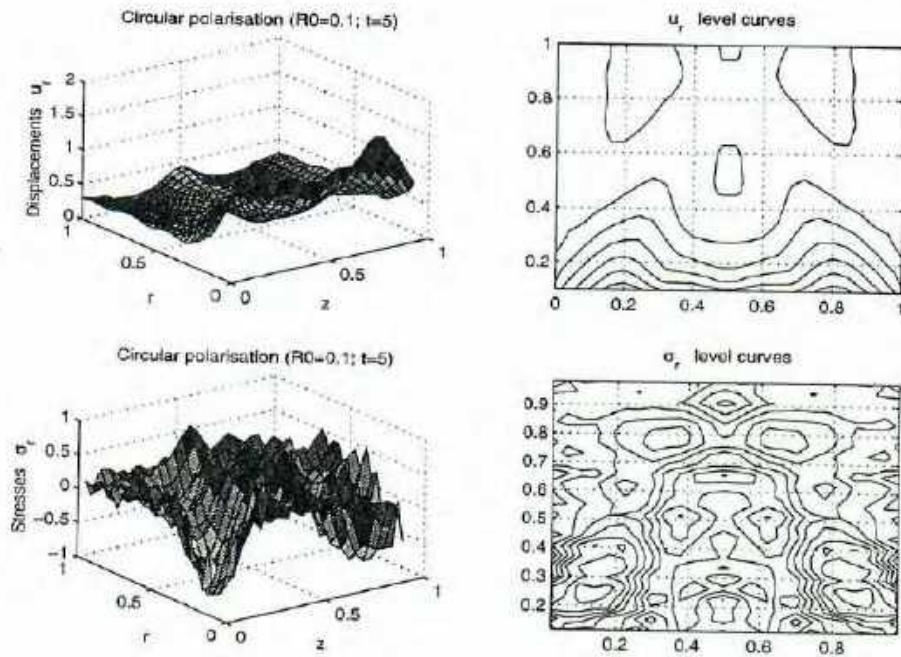


Figure 4: Displacements and radial stresses in a finite-length thick cylinder with circular preliminary polarisation ($l=0.9$, $a=1$).

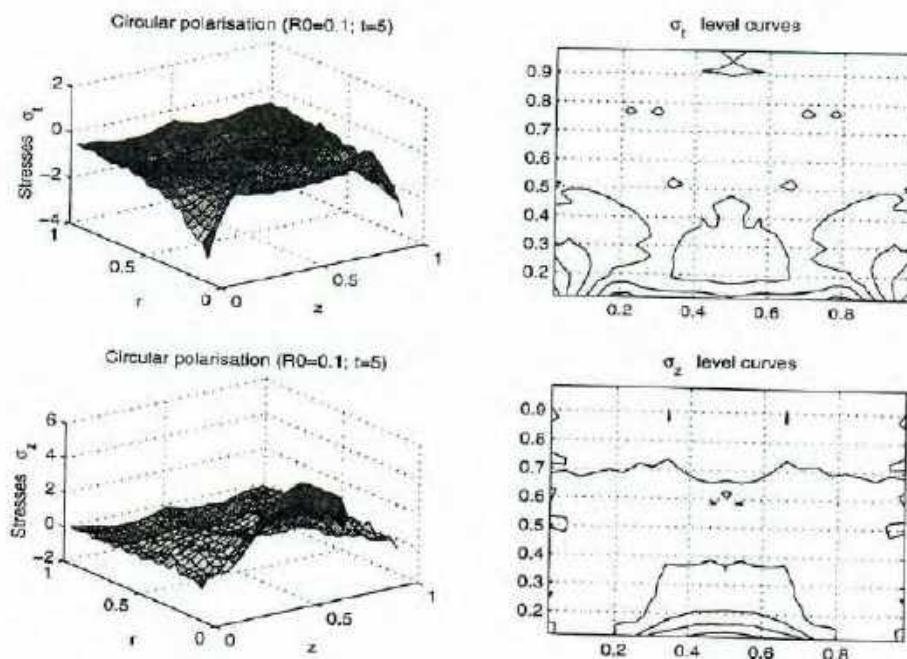


Figure 5: Stresses in a finite-length thick cylinder with circular preliminary polarisation ($l=0.9$, $a=1$).

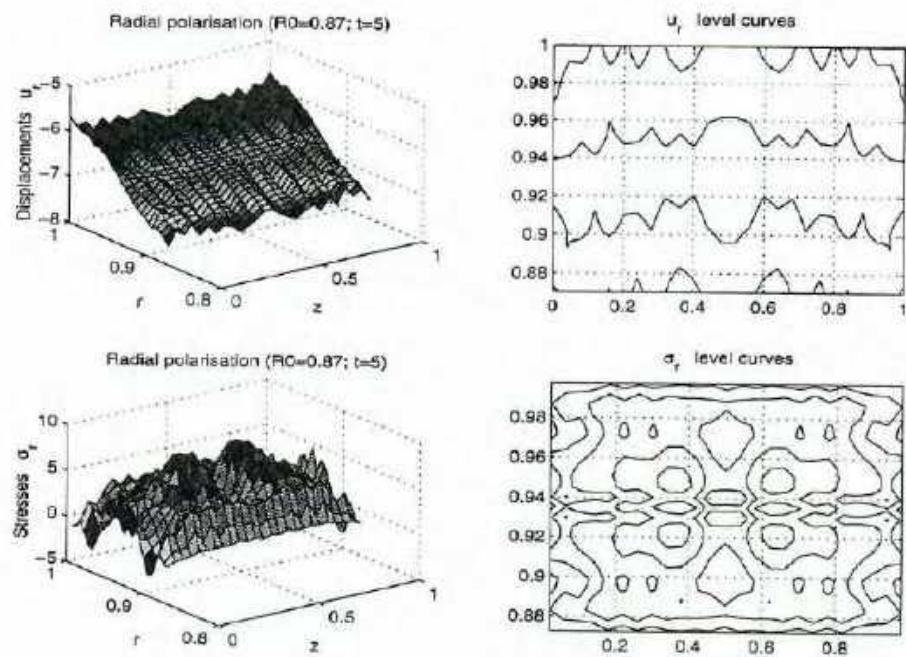


Figure 6: Displacements and radial stresses in a finite-length thin cylinder with radial preliminary polarisation ($l=0.13$, $a=1$).

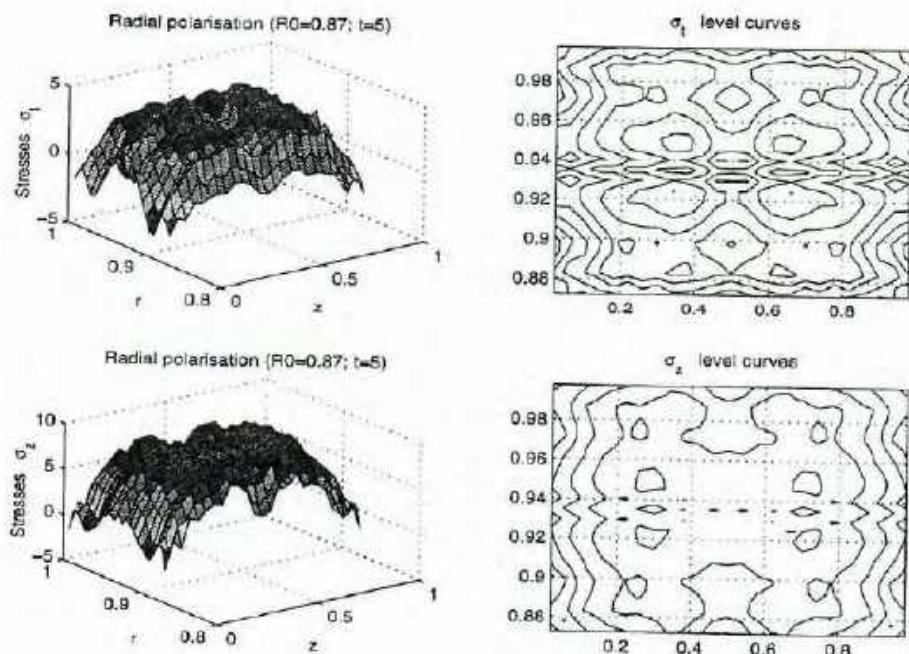


Figure 7: Stresses in a finite-length thin cylinder with radial preliminary polarisation ($l=0.13$, $a=1$).

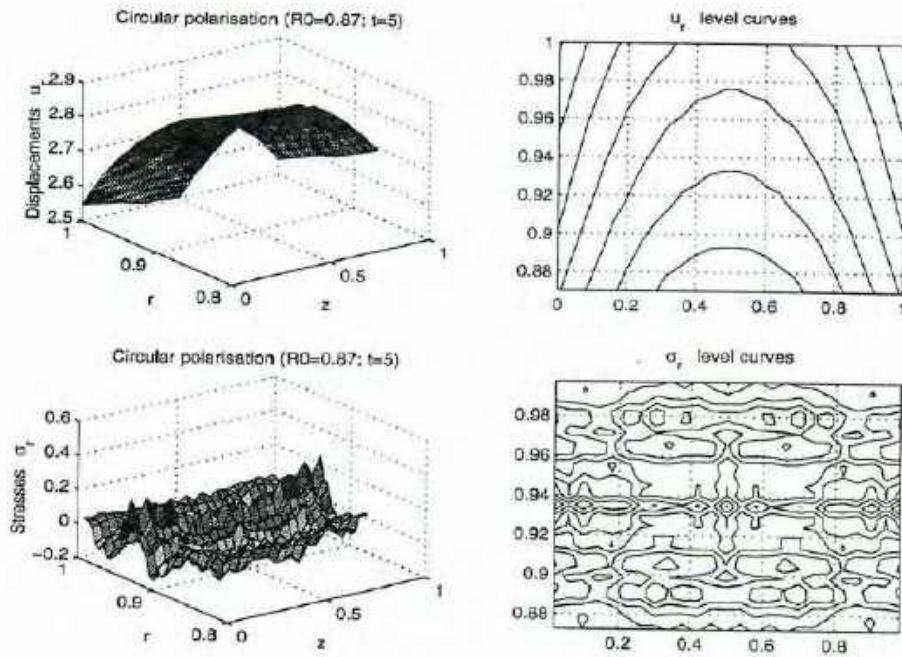


Figure 8: Displacements and radial stresses in a finite-length thin cylinder with circular preliminary polarisation ($l=0.13$, $a=1$).

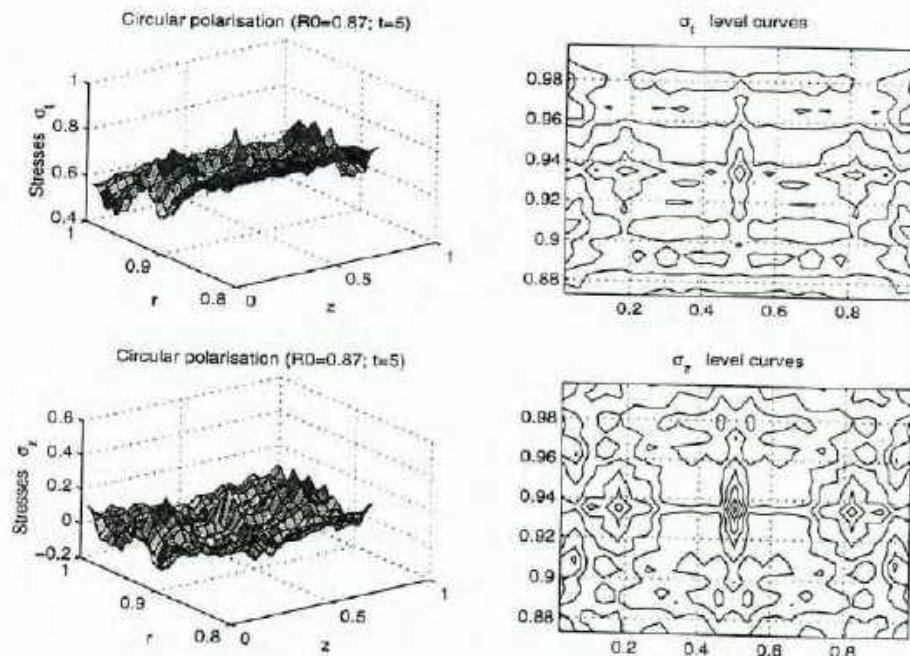


Figure 9: Stresses in a finite-length thin cylinder with circular preliminary polarisation ($l=0.13$, $a=1$).

coupled field theory. It is often more reasonable to directly apply an effective numerical technique to the original non-stationary coupled problem.

Piezoelectricity gives just one example where the coupling phenomenon between two different physical fields has to be dealt with in the early stage of mathematical modelling for the successful design of various technical devices. The directions of coupled field theory are virtually unexhaustable, with thermoelasticity, pyroelectricity, electrooptics, magnetoelasticity, magnetooptics giving just few additional examples. In the final analysis, it is clear that there is no single field in nature that acts on its own. Of course, in some cases it may be unreasonable and/or impractical to include the information about other fields. Then we may limit ourselves to a specific (and, perhaps, very good) approximation. Nevertheless, in understanding many physical, chemical and biological phenomena, it is important to include additional information about different interrelated components of these phenomena into our approximations. The improvement of such approximations is intrinsically connected with the physical parameterisation of mathematical models. In order to gain a more penetrating insight into the coupled field theory concepts, including those of piezoelectricity, we have to look deep down inside the structure of the material [6, 7]. Therefore, only a blend of the methods and techniques of mathematics, physics and material sciences can provide the necessary tools for the investigation of physical phenomena in finite structures.

Acknowledgements

The work was partially supported by grant USQ-PTRP 179389. I am deeply grateful to the School of Mathematics at the University of South Australia, where a part of this work was made, for their hospitality. I also would like to thank Anne Fuller for her helpful assistance at the final stage of preparation of this paper.

References

- [1] Antonova, E.E., Silvester, P.P., Finite Elements for Piezoelectric Vibrations with Open Electric Boundaries, *IEEE Trans. Ultrason., Ferroelect., Freq. Contr.*, Vol.44, No. 3, 1997, 548-556.
- [2] Ballato, A., Piezoelectricity: Old Effect, New Thrusts, *IEEE Trans. on Ultrasonics, Ferroelectrics*, **42**, No. 5, 916, 1995.
- [3] Berlincourt, D.A., Curran, D.R., and Jaffe, H. *Piezoelectric and Piezomagnetic Materials and Their Function in Transducers*, in "Physical Acoustics", Vol. 1A, Ed. W.P. Mason, New York and London: Academic Press, 1964, 204-236.
- [4] Brown, L.F., Ferroelectric Polymers: Current and Future Ultrasonic Applications, *Proceedings of the IEEE Ultrasonic Symposium*, 1992, 539-550.
- [5] Buchanan, G.R., Peddieson, Jr., J., Vibration of Infinite Piezoelectric Cylinders with Anisotropic Properties Using Cylindrical Finite Elements, *IEEE Trans. Ultrason., Ferroelect., Freq. Contr.*, Vol. 38, No.3, 1991, 291-301.
- [6] *Ceramic Materials for Electronics: Processing, Properties, and Applications*, Ed. R.C. Buchanan, Marcel Dekker, 1991.

- [7] *Mathematics of Microstructure Evolution*, Eds. Long-Qing Chen et al, SIAM & MMMS, Proceedings in Applied Mathematics 90, 1996.
- [8] Dieulesaint, E., Royer, D., *Elastic Waves in Solids: Applications to Signal Processing*, Chichester; N.Y.: J.Wiley, 1980.
- [9] Fielding, J.T. et al, Characterization of PZT Hollow-Sphere Transducers, *Proceedings of the IX IEEE International Symposium on Applications of Ferroelectrics*, 1994, 202–205.
- [10] Fukada, E., Poiseuille Medal Award Lecture: Piezoelectricity of biopolymers, *Biorheology*, 32, 593, 1995.
- [11] Godunov, S.K., *Equations of Mathematical Physics*, Moscow, Nauka, 1980.
- [12] Gururaja, T.R., Piezoelectric Transducers for Medical Ultrasonic Imaging, *American Ceramic Society Bulletin*, Vol. 73, No.5, 1994, 50–55.
- [13] Ikeda, T., *Fundamentals of Piezoelectricity*, Oxford: Oxford University Press, 1990.
- [14] Kagawa, Y., Tsuchiya, T., Kawashima, T., Finite Element Simulation of Piezoelectric Vibrator Gyroscopes, *IEEE Trans. Ultrason., Ferroelect., Freq. Contr.*, Vol. 43, No. 4, 1996, 509–120.
- [15] Kawawa, Y., Tsuchiya, T., Furukawa, G., Finite Element Simulation of Dynamic Responses of Piezoelectric Actuators, *J. of Sound and Vibration*, Vol. 191, No. 4, 1996, 519–528.
- [16] Lee, J.S., Boundary Element Method for Electroelastic Interaction in Piezoceramics, *Engineering Analysis with Boundary Elements*, Vol. 15, No. 4, 1995, 321–328.
- [17] Le Letty, R., Claeysen, F., Bossut, R., Combined Finite Element-Normal Mode Expansion Methods in Electroelasticity and Their Application to Piezoactive Motors, *Int. J. Numer. Methods Eng.*, Vol. 40, No. 18, 1997, 3385–3395.
- [18] Lerch, R., Simulation of Piezoelectric Devices by Two- and Three-Dimensional Finite Elements, *IEEE Trans. Ultrason., Ferroelect., Freq. Contr.*, Vol. 37, No. 3, 1990, 233–243.
- [19] Lu, P., Mahrenholtz, O., A Variational Boundary Element Formulation for Piezoelectricity, *Mechanics Research Communications*, Vol. 21, No.6, 1994, 605–615.
- [20] Melnik, R.V.N., Existence and Uniqueness Theorems of the Generalised Solutions for a Class of Non-stationary Problems of Coupled Electroelasticity, *S. Mathematics (Iz. VUZ)*, Vol. 35, No. 4, 1991, 24–32 (by Allerton Press).
- [21] Melnik, R.V.N., Moskalkov, M.N., Difference Schemes for and Analysis of Approximate Solutions of Two-Dimensional Nonstationary Problems in Coupled Electroelasticity, *Differential Equations*, Vol. 27, No. 7, 1991, 1220–1230 (by Plenum Publishing Corporation/Consultants Bureau, N.Y., 1992, 860–867).
- [22] Melnik, R.V.N., The stability condition and energy estimate for non-stationary problems of coupled electroelasticity, *Mathematics and Mechanics of Solids*, Vol. 2, No. 2, 1997 153–180.
- [23] Melnik, R.V.N., Intelligent structures and coupling in mathematical models: examples from dynamic electroelasticity, *Proceedings of the IEEE ICPADM'97*, Seoul, Korea, 1997.
- [24] Melnik, R.V.N., Convergence of the operator-difference scheme to generalized solutions of a coupled field theory problem, to appear in *Journal of Difference Equations and Applications*, 1998.

- [25] Moskalkov, M.N., Investigation of a difference scheme for the solution of the sound radiator problem for cylindric piezovibrators, *Differential Equations*, **22**, No.7, 1220-1226, 1986.
- [26] Nowacki, W., Electromagnetic Effects in Solids, Mir Publishers, Moscow, 1986.
- [27] *Piezoelectricity*, Eds. C.Z. Rosen, B.V. HIREMATH, R. NEWNHAM, N.Y. Americal Institute of Physics, 1992.
- [28] Samarskii, A.A., *Theorie der Differenzenverfahren*, (Series: Mathematik und ihre Anwendungen in Physik und Technik), Aufl. Leipzig: Academische Verlagsgesellschaft Geest & Portig, 1984.
- [29] Samarskii, A.A., Nikolaev, E.S., *Numerical Methods for Grid Equations*, Basel, Boston: Birkhauser Verlag, 1989.
- [30] Shashkov, M., Steinberg, S., *Conservative Finite Difference Methods on General Grids*, Boca Raton: CRC Press, 1995
- [31] Takahashi, S., *Ferroelectric Ceramics*, Eds. N. Setter, E.L. Colla, Basel: Birkhäuser Verlag, 1993, 349–362.
- [32] Trenogin, V.A., *Functional Analysis*, Moscow, Nauka, 1993.
- [33] Vatulyan, A.O., Kublikov, V.L., Boundary Element Method in Electroelasticity, *Boundary Elements Communications*, Vol. 6, No. 2, 1995, 59–61.

UNIVERSITY OF SOUTH AUSTRALIA



TECHNICAL REPORT SERIES

**Mathematical Models for Climate as a
Link between Coupled Physical Processes
and Computational Decoupling**

by

(R.) V Nick Melnik

Report No. 1997/1

CENTRE FOR INDUSTRIAL
AND APPLIED MATHEMATICS
SCHOOL OF MATHEMATICS

Faculty of Information Technology

The Levels, South Australia 5095, Telephone (08) 8302 3343 Facsimile (08) 8302 5785

TECHNICAL REPORT SERIES

**Mathematical Models for Climate as a
Link between Coupled Physical Processes
and Computational Decoupling**

by

(R.) V Nick Melnik

Report No. 1997/1

MATHEMATICAL MODELS FOR CLIMATE AS A LINK BETWEEN COUPLED PHYSICAL PROCESSES AND COMPUTATIONAL DECOUPLING

V. Nick Melnik

E-mail: matvnm@lv.levels.unisa.edu.au

Abstract

Mathematical models for climate studies are treated as a coupling link between physical and computational models. These models are characterized by the fact that small-scale phenomena influence the large-scale properties of the modelling system, yet the former cannot be extracted from the latter using available hardware and computational procedures. Climate systems belong to the class of systems whose dynamics are only observable in transient states. As a result, the sensitivity of models to coupling procedures requires an examination of the schemes responsible for transporting data between components. It is proposed to perform such an examination, based on the connection between error growth and the degree of coupling of model components, using adaptive error control.

Key words: coupling and decoupling procedures, hydrodynamic stability.

1 Introduction

Elements of the mathematical modelling of climate can be traced back to Aristotle's Meteorologica. The first contemporary achievement in this field is often attributed to U. Leverrier, who was the first to produce a weather map after a big storm in France in November 1854. A rigorous modern basis for short-range weather prediction was laid by Vilhelm and Jacob Bjerknes and other scientists of the Bergen school. For the

first time they rigorously treated the problem of weather forecasting as a mechanico-mathematical problem. Such a problem was described mathematically by an initial value problem for the hydrodynamic equations of a baroclinic fluid. One of the main concepts introduced in this work was *the concept of wave instability on frontal interfaces*, which is still of primary importance in theoretical meteorology.

An important step in mathematical climate studies was made by L. Richardson and Courant, Friedrichs, Lewy in the field of numerical analysis. The results of the latter group gave a guideline for the explanation of some failures in the former work, where meteorological “noises”¹ were included into the model.

During the early 1940s through to the 60s, many scientists contributed to filtering noise from the solution for the hydrodynamic system. In the field of theoretical meteorology a number of pioneering works were published, establishing basic principles for simplification of the hydrodynamic equations through the *quasi-geostrophic expansion*. Mathematical foundations of the theory were laid using a probabilistic approach [22, 39]. The main difficulty with the early works in this direction was the formulation of boundary conditions for the problem. Alternatives to the geostrophic approximations of the hydrodynamic equations were also developed. Among them was the *quasi-solenoidal approximation*.

Serious mathematical difficulties in the practical application of quasi-geostrophic and quasi-solenoidal approximations led in the late 1950s to a return to the initial hydrodynamic equations which were essentially used by L. Richardson. Such equations in theoretical meteorology are called *primitive*. Since that time, the main developments have been concentrated on numerical methods and the improvement of models by a better physical parameterization. Many important factors related to the physical parameterization were taken into consideration and implemented into models. This led to the creation of modern state-of-the-art models for climate that

- consist of relatively independent components that are responsible for interconnected parts of climate such as atmosphere, ocean, land surface, sea ice, etc, and
- require substantial computational power to obtain approximate solutions.

The quality of these approximate solutions depend significantly on the consistency between the mathematical model and the real climate. Since the improvement of mathematical models can be achieved by improved physical parameterization, *the concept of coupling between different components* becomes straightforward.

From the physical point of view, climate studies are essentially based on three fundamental theories. These are thermodynamics, the theory of radiation, and magnetohydrodynamics. Since a description of the climate system should include both the earth and the atmosphere, the overall system is often referred to as the atmosphere-active-layers (AAL) system. Both the earth and the atmosphere require a detailed physical parameterization, that leads to the difficulty of mathematically

¹such as acoustic waves

formalising *the interaction between underlying processes and phenomena*. The main difficulty with the description of the earth is to measure a purely gravitational force. In fact, we can only observe the combined effect of the two forces, gravitational and centrifugal, that is referred to as gravity. The measurement difficulties measurements stem from

- the variation of gravity at different latitudes;
- the variation of gravity vertically with respect to sea level;
- the variation of gravity erratically with respect to the earth's crust and other irregularities.

The main difficulty in the description of the atmosphere is to adequately represent transport effects in models. Amongst the most important constituents in the atmosphere are water (about 4% per volume), carbon dioxide (about 0.03% per volume), ozone (about $0.1 \times 10^{-5}\%$ per volume), oxygen (about 20%) and nitrogen (about 70%). However, if water and carbon dioxide are present throughout the atmosphere, then ozone becomes influential only at 20-30 km from the earth's surface, oxygen only from about 80 km and nitrogen even higher [37]. In addition, many gaseous constituents only have an indirect influence through the propagation of electromagnetic waves. There are also many important nongaseous constituents such as condensed H_2O , salt particles, dust and others that play important role in the description of clouds and precipitation.

We conclude that transport phenomena and gravity are key factors in an adequate description of the AAL system.

2 The structure of the paper and notation

This paper is organised as follows:

- Section 3 gives a brief outline of space-time scales that are important in a climate study. We recall the main hypothesis that is used in the mathematical modelling of synoptic processes.
- In Section 4 we consider the distinction between short and long range predictions on the basis of the concept of the relaxation time. The fundamental equations for the adiabatic approximation are also presented in this section.
- Section 5 is devoted to non-adiabatic models and their simplifications on the basis of quasi-geostrophic and quasi-solenoidal approximations.
- In Section 6 we consider advantages in the return to the primitive hydrodynamic equations, and difficulties in an adequate representation of the vertical structure of meteorological fields.
- Section 7 deals with non-adiabatic factors that lead to an approximation of the conservation law in mathematical models.

- Sections 8–10 are devoted to a dilemma between the concepts of coupling and independence as well as approaches for the numerical treatment of an interplay between these two concepts.
- In Sections 11 and 12 we formulate a finite set of differential equations that provides an approximation to the dynamics of climatic processes. We address two questions related to such an approximation, namely phase transitions and algorithmic stability.
- Section 13 deals with questions related to the validation of mathematical models for climate. We argue that for the validation of models, both a-priori assumptions and a-posteriori information are needed.
- In Section 14 we present some numerical results of the computation of meteorological fields on the basis of the NCAR CCM3 model.
- Section 15 concludes the paper. Directions for future development are also presented in this section.

The following notation is used throughout the paper:

- g is the acceleration due to gravity;
- ρ is the density;
- p is the pressure;
- T is the temperature;
- z is the altitude;
- m is the mass of the earth;
- c_0 is the isothermal sound of speed;
- h_a is the height of the atmosphere;
- p_0 is the average surface pressure;
- ρ_0 is the average surface density;
- T_0 is the average air temperature at sea level;
- k_b is the Boltzmann constant;
- k_v is the von Karman constant;
- $k_r = c_p/c_v$ is the ratio of specific heat capacity under constant pressure and constant volume;
- R is the gas constant, for example, $R = 287 \text{ J kg}^{-1} \text{ K}^{-1}$ for dry air and $R = 461 \text{ J kg}^{-1} \text{ K}^{-1}$ for water vapor;
- $\Omega = 7.292 \times 10^{-5} \text{ s}^{-1}$ is the angular velocity of the earth;
- φ is the latitude;
- $\chi = 2\Omega \sin \varphi$ is the Coriolis parameter;
- $\mathbf{F} = (F_x, F_y, F_z)$ is the field external to the earth, excluding pressure-gradient forces;
- c_p is the specific heat capacity at constant pressure, for example, $c_p = 1. \times 10^3 \text{ J kg}^{-1} \text{ K}^{-1}$ for dry air and $c_p = 1.81 \times 10^3 \text{ J kg}^{-1} \text{ K}^{-1}$ for water vapor.

Other notation is explained in the text as required.

3 Space-time scales and their interaction

In climate studies one of the most challenging problems is to adequately describe space-time scale interactions [23]. By interacting between themselves, different biogeochemical processes at different scales form a unified whole which we call climate. The problem of such interactions is typically simplified mathematically by regarding microturbulence as a dissipative factor which can be characterized by an *effective (or dynamic) viscosity coefficient*. Such a simplification allows us to effectively model *synoptic oscillations*, i.e. climate processes that are characterized by time scales from hours to several days. Diurnal oscillations also belong to this class. Amongst other types of oscillations, the following classes can be distinguished:

- Global oscillations, for example, planetary oscillations. They play an essential role in long-term weather predictions. Their time scales are characterized by the period from weeks to months. The Atmospheric Boundary Layer is a key factor in such processes.
- Seasonal oscillations that vary over a year.
- Interannual oscillations with time scales of several years. To this class belong, for example, glacial periods and ENSO-type phenomena.
- Micrometeorological oscillations with time scales of seconds to minutes. Small-scale turbulence, acoustic waves, and gravitational waves with small amplitudes provide examples of this type of oscillations.
- Mesometeorological oscillations such as thunderstorms, and gravitational waves with large amplitudes. They typically last from minutes to an hour.

We emphasize that mathematical models of climate systems are essentially “proxy” climate systems. Whatever model is chosen, small scale phenomena may substantially influence large-scale properties of the system, but computational procedures may not be available to extract the former from the latter.

4 Short and long range in the prediction of meteorological fields

The short-range prediction of meteorological fields is based on hydrodynamic theory in the case where the energy of sources and sinks is virtually ignored by using the *adiabatic approximation*. In quite a general setting, the resulting equations can be derived from the two conservation law equations, the conservation of the entropy and the conservation of the “vortex charge” [37],

$$\frac{d\Xi}{dt} = 0, \quad \frac{d\Psi}{dt} = 0, \quad (4.1)$$

where $\Psi = (\Psi_0 \cdot \nabla \Xi)/\rho$ is the potential Rossby vorticity. Due to (4.1), the entropy function Ξ and the absolute vorticity Ψ_0 are generators of differential adiabatic invariants, because any function of them is again an adiabatic invariant. As an integral invariant, all systems that can be described by the model (4.1) have the total energy of the system, \mathcal{E} , constant.

In the general case, the well-posedness of the model (4.1) is not an established mathematical fact. Any specific choice of two independent Lagrangian coordinates as well as the definition of two parts of the integral invariant², implies the necessity of addressing the problem of system stability. E.N. Lorenz was the first who proposed addressing this problem using the *macrostability parameter*, \mathcal{S} . Let

$$\Theta = T(p_0/p)^{(k_r - 1)/k_r} \quad (4.2)$$

be the potential temperature, where p_0 is the standard pressure. This quantity is often convenient as one of the Lagrangian coordinates. Then, the parameter of macrostability, \mathcal{S} , can be defined as the weighted average value of the vertical gradient of the potential temperature over the entire thickness of the atmosphere. If \mathcal{K} is the kinetic energy of the system, then the quantity $\mathcal{K} - \mathcal{S}$ is also adiabatic invariant. Hence, the quantity \mathcal{S} shows the amount of kinetic energy that is released/absorbed in the process of adiabatic transitions. This approach to stability requires an adequate specification of the vertical structure of the atmosphere.

An alternative approach is based on the concept of relaxation time, τ . From the mechanical point of view, the relaxation time, or "build-in" period, can be seen as the atmospheric efficiency coefficient, i.e. the rate at which potential energy, \mathcal{E}_p , is converted into kinetic energy:

$$\tau = \left(\frac{1}{\mathcal{E}_p} \frac{\partial \mathcal{E}_p}{\partial t} \right)^{-1}. \quad (4.3)$$

On the scale of synoptic processes, $\tau \approx 1$ week. If the time interval of interest $t - t_0$ (where t_0 is an initial moment of time) is less than τ the model (4.1) may provide a good approximation for short-range weather changes. However, for periods that satisfy the inequality

$$t - t_0 > \tau, \quad (4.4)$$

practically all regions of the atmosphere have sufficient time to interact with each other and the model (4.1) becomes inappropriate. Since the atmosphere is a rapidly changing component with low inertia of the whole AAL system, an essential part of investigation in the field of climate study is being concentrated on atmospheric modelling. This approach gives rise to the major difficulty in the modelling of long-term meteorological changes. We have to fix the initial state of the whole physical system, namely when $t = t_0$. Of course, this requires more careful examination

²for example, the kinetic energy and the labile energy that, in turn, consists of the sum of the potential energy and the internal energy of the system

of other components of the system, in particular the ocean, which is a component with a large thermal inertia. Hence, one of the most important initial conditions for the whole model is the temperature field. It is well-known, for example, that incompleteness of such data causes problems in modelling processes such as ENSO phenomenon, and other processes in the equatorial zones where the Coriolis parameter vanishes and the structure of the boundary layer of the atmosphere has to be modelled with an increased precision [16, 62, 6, 33, 60]. The implementation of *non-local* features of the system into the model becomes important for the validity of the model [13]. In the end, this requires an appropriate description of *turbulence in the boundary layer* as a major factor responsible for the *mixing* of heat, momentum, passive scalars, moisture etc. This emphasises the importance of taking into account both the interaction of time-scales and the interaction of spatial scales [48].

As we mentioned in the introduction, three main types of physical processes, namely

- thermodynamical,
- radiative, and
- magnetohydrodynamical,

influence the output of mathematical models of climate subjected to the physical parameterization [29, 45]. As a result, many efforts during recent times have been concentrated on improvements of existing physical parameterizations. In particular, much attention has been devoted to an adequate modelling of radiative processes [4] that require appropriate models for cloudiness and the transport of tracer species [47]. The analysis of sensitivity to transport phenomena has led to an increased interest in the semi-Lagrangian approach as an alternative to the well-established spectral approaches [46]. The semi-Lagrangian approach requires special numerical procedures for interpolation to compensate for the sparse character of data.

The necessity of an adequate representation of transport phenomena in mathematical models naturally leads to the development of the concept of coupling with respect to different components of climate [5]. In addition to *the large spatial scale* that cover thousands of kilometers other scales, such as

- mesoscale (from kilometers to hundreds of kilometers),
- small scale (from dozens of meters to kilometers), and
- microscales (from millimeters to dozens of meters).

become important for long-range prediction. Of course, in climate applications smaller scale structures are inevitably described statistically. It does not follow, however, that such structures are necessarily random in nature. The only a-priori conclusion we can draw is that in long-term processes *the atmosphere as a whole* does not act as a *closed system* [37]. It acts as a component of a bigger AAL system composed of the atmosphere and active layers that can be described by a coupling of many different physical, chemical and biological fields.

5 Physical hypotheses and mathematical approximations

In the long-term prediction, equations for the conservation of entropy and the potential vorticity cannot provide an adequate description of underlying processes that are essentially nonadiabatic. Since in this case we should be able to adequately describe sources and dissipation of energy, certain assumptions about the laws of dissipation and accumulation of energy should be made. In this case, model (4.1) should be replaced by evolutionary equations that more adequately represent transport phenomena. Let us denote by Λ the rate of energy increase per unit mass, and by \mathbf{F}_v the viscous force per unit mass. Then such equations can be written in the form

$$T \frac{d\Xi}{dt} = \Lambda, \quad \rho \frac{d\Psi}{dt} = \operatorname{div} [(\Lambda/T)\Psi_0 + \Xi(\operatorname{curl}\mathbf{F}_v)]. \quad (5.1)$$

This model allows us to take into account *non-adiabatic effects*. It introduces the two new variables, the momentum and heat fluxes, and requires additional hypotheses on sources and sinks in the AAL system. The central hypothesis is the hypothesis of *local thermodynamic equilibrium*. During recent years theoretical and experimental physicists proved it was necessary to go beyond the framework of this hypothesis [38, 35]. We note, that model (5.1) is a model of local type. In fact, to derive this model, the Obukhov hypothesis on the conservative properties of the potential vorticity is used [37], and the field

$$(\Lambda/T)\Psi_0 + \Xi(\operatorname{curl}\mathbf{F}_v) \quad (5.2)$$

is assumed to have a solenoidal structure. Nevertheless, in models of this type, difficulties connected with the inherently approximate modelling of physical processes such as the polarization of radiation, refraction, dispersion, and cloudiness, may still be partially overcome by introducing some feedback mechanisms, like regulators through the cloud cover, sea ice, snow cover etc.

The model (5.1) with a realistic physical parameterization is extremely difficult to deal with without some additional simplifications. Once again, we can use the local equilibrium hypothesis for mathematical simplifications of climate system models. This hypothesis leads to satisfactory results, at least in the case of small-amplitude waves. In this particular case it is sufficient to relate small oscillations of the atmosphere to the *equilibrium state*. When oscillations are small the *perturbation theory technique* is natural for the investigation of solutions of resulting models. In its essence, the successful application of this theory in many areas has its roots in the local equilibrium hypothesis.

Let us assume that in the equilibrium state pressure, density and temperature, p , ρ and T , depend only on the amplitude z . Then, as a primary task, we have to describe the *dynamics of these variables near the equilibrium*. In climate study, these variables are connected, as a rule, by the "timeless" *hydrostatic* (or quasi-

static) equation,

$$\frac{\partial p}{\partial z} = -\rho g, \quad (5.3)$$

and the Clapeyron equation

$$p = \rho RT, \quad (5.4)$$

where, as usual, g is the acceleration due to gravity and R is the specific gas constant.

For now, we shall assume that at the initial moment of time t_0 the atmospheric motion is

- quasi-static, i.e.

$$\partial p_e / \partial z = -g p_e; \quad (5.5)$$

- horizontal, i.e. for the velocity field $\mathbf{v} = (u, v, w)$ we assume

$$w = 0; \text{ and} \quad (5.6)$$

- geostrophic i.e. the velocity field is assumed to be non-divergent,

$$u_e = -\frac{1}{\chi\rho} \frac{\partial p_e}{\partial y}, \quad v_e = \frac{1}{\chi\rho} \frac{\partial p_e}{\partial x}, \quad (5.7)$$

where χ is the Coriolis parameter and the subscript index e stands for the values of thermodynamic parameters at moment t_0 . Of course, in this case one can introduce the stream function by the formula

$$\psi = p_e / (\chi\rho), \quad (5.8)$$

and the formulas (5.7) may be rewritten as

$$u_e = -\frac{\partial \psi}{\partial y}, \quad v_e = \frac{\partial \psi}{\partial x}. \quad (5.9)$$

One of the assumptions that is often made is that these three properties (referred to as *consistency conditions*) will be preserved in the future. This guarantees a stationary solution for the set of equations governing the atmosphere. Such motion is called motion of the first kind or *slow motion*. By the standard technique we can also account for the curvature of the earth by the transformation of the stationary solutions into slow gyroscopic Rossby waves.

If consistency conditions are violated in any region of space, X , *fast motion*, or motion of the second kind, has to be taken into consideration. In reality, we observe a continuous competition between a violation of the consistency conditions and adaptation of the meteorological fields \mathbf{v} , p , ρ , T . Since traditional methods assume that "meteorological noise" has little significance in the weather prediction, numerous attempts have been made to filter out motions of the second kind. However, if we accept the quasi-static approximation, all frequencies of the internal acoustic

waves go to infinity ("complete filtering"). Frequencies of the gravity waves become overestimated, though the error decreases for longer waves.

The weather is typically associated with synoptic processes. Hence, in order to describe such processes at a "minimal cost", we have to filter out from solutions of the Eulerian hydrodynamic equations (they contain both slow and fast motions) the motions of the second kind. If it is assumed that only the motion of the first kind is important for synoptic processes, then in the Eulerian equations we have two dimensional parameters, g and χ , that can be used for such filtering. Nonlinear systems in which fast oscillations occur along with slow ones are typical in applications³. Asymptotic methods and the theory of perturbations were developed in order to mathematically treat such systems. The idea of expansion with respect to a small parameter can also be used as a simplification of the hydrodynamic system to make it appropriate for the description of synoptic processes.

Let L and U be typical length and speed scales for synoptic processes. Then the role of small parameters may be played by the Rossby-Kibel number

$$R_k = U/(\chi L), \quad (5.10)$$

or the Mach number

$$M_a = U/c_0. \quad (5.11)$$

The isothermal sound speed, c_0 , is defined using the height of the atmosphere, h_a , as

$$c_0 = \sqrt{gh_a}, \text{ where } h_a = p_0/(\rho_0 g). \quad (5.12)$$

The Rossby-Kibel number may be interpreted as the ratio of the typical relative acceleration, U^2/L , to the typical Coriolis acceleration χU [37]. Using standard series expansion, the consistency conditions for the first kind motion may be defined from conditions for the vertical potential vorticity,

$$\Psi_z = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} = \frac{g}{f} \nabla^2 z + O\left(\frac{UR_k}{L}\right), \quad (5.13)$$

and for the horizontal divergence,

$$D_h = \frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} = O\left(\frac{UR_k}{L}\right). \quad (5.14)$$

This expansion made with respect to the parameter UR_k/L is referred to as the *quasi-geostrophic expansion*. The quantities p_0 and ρ_0 in (5.12) are defined at the surface of the earth, and have to be included in the boundary conditions of the problem. We also need the initial field of the pressure to compute the pressure at future moments of time. However, the advantage of the quasi-geostrophic approximation is that formally we are not required to know the initial distribution of

³Probably, the most widely cited example is the Van-der-Pol equation for the description of oscillations in an electric circuit containing a vacuum tube with feedback

the velocity field. It should be noted that, strictly speaking, the resulting system can only be represented in the form of differential equations for synoptic processes in a barotropic-atmosphere approximation. A weak formulation of the problem is required for the more general case.

The quasi-geostrophic approximation is one of the possible approximations to conservation laws for the horizontal motion. In spite of its partial advantages it has serious drawbacks. The major drawback is a violation of the assumption of smallness for the Rossby-Kibel number, R_k , in the vicinity of the equator, since the Coriolis parameter $\chi = 2\Omega \sin \varphi$ decreases near the equator. Moreover, there is evidence that this approximation may be inadequate even outside of the tropical zones. In such cases the consistency conditions (5.13), (5.14) for the synoptic fields must be changed. For example, in many cases it is reasonable to assume that the horizontal divergence D_h is small compared to the vertical vorticity Ψ_z , where

$$\Psi_z = O(U/L), \quad (5.15)$$

$$D_h = O\left(\frac{U}{L} \frac{M_a^2}{\alpha_0^2}\right). \quad (5.16)$$

In (5.16), α_0 is the parameter of (hydro)static stability. It is defined as

$$\alpha_0^2 = -(T/T_0)(p/c_p)\partial\Xi/\partial p, \quad (5.17)$$

and has an obvious connection with the Richardson number [37]

$$R_i = \frac{T_0^2}{M_a^2 T^2} \alpha_0^2, \quad (5.18)$$

(see also the definition in [11]). These modified consistency conditions, (5.15), (5.16), are derived with the assumption that in the horizontal velocity field for the slow (synoptic) motion, the potential component is small compared to the solenoidal component. Such an assumption leads to the *quasi-solenoidal approximation*. The mathematical model that corresponds to this approximation is typically formulated in terms of the stream function ψ , where

$$u = -\frac{\partial\psi}{\partial y}, \quad v = \frac{\partial\psi}{\partial x}. \quad (5.19)$$

The model requires both

- an initial field ψ for a given z , and
- values of ψ on the boundary.

It is well-known that the balance equation written for ψ is the *Monge-Ampere-type equation* that should be solved in a given bounded region. The well-posedness

of such a problem in the general setting can be guaranteed when it is elliptical. However, the ellipticity condition [37],

$$g\nabla^2 z + \chi^2/2 > 0, \quad (5.20)$$

can only be satisfied when the quantity R_k is small. As we mentioned above, this condition is often violated in applications. Such a violation is most noticeable in the tropical zones.

In addition to the mathematical difficulties in using the quasi-solenoidal approximation, and unsatisfactory practical results obtained in many cases with the quasi-geostrophic approximation, there remains the problem of filtering fast waves from the hydrodynamic equations open. However, progress in computational software and hardware has led to the possibility of solving the complete set of equations governing the atmosphere.

6 Primitive equations and the vertical structure of meteorological fields

Since in many cases neither quasi-geostrophic nor quasi-solenoidal approximations are appropriate in applications, computational complexity is caused by filtering procedures. The development of numerical procedures from the original hydrodynamic system in these cases is no more complex. The main problem in modelling with the primitive equations is the formulation of appropriate boundary conditions. If boundary conditions are not properly formulated, the stability of the solution cannot be guaranteed. In fact, we need

- the normal velocity on the entire boundary, and
- the potential vorticity on the part of the boundary where air motion is directed toward the interior.

Because of the approximate character of available data for these boundary conditions, the formulation of the mathematical model requires two types of equations, prognostic equations and tendency equations. With the quasi-static approximation, the derivation of the tendency equation is straightforward, provided the vertical structure of meteorological processes is known. In the particular case of barotropic atmosphere, the problem of the vertical structure of synoptic processes is reasonably simple. This explains the early development of the theory in the direction of simplifications described in Section 5.

However, in the general case of a baroclinic atmosphere, the problem of the vertical structure of meteorological fields remains one of the most serious problems for mathematical modelling using the primitive equations. At the initial stage of development of the model, mainly pressure was used as the vertical coordinate. The implementation of the earth's orography [41] led to the adoption of a σ -system

coordinate for the model. Among the first to use the geometrical altitude as a vertical coordinate was L. Richardson [50]. This idea was developed further by V. Starr (see references in [20]) who introduced quasi-Lagrangian coordinate systems. The present development of the vertical structure of meteorological fields in the NCAR CCM3 model [1] is based on the works of Kasahara and Washington [19], Kasahara [20], and Simmons and Stufing (see references in [21]). Among other approaches to vertical coordinates we include initial attempts at using the potential temperature (see (4.2)). This very fruitful idea has not received a proper development in the literature due to difficulties connected with lower boundary conditions.

From the mathematical point of view, the primitive equations are characterised by the “restoration” of the hyperbolic operator in the model. From the physical point of view we retain gravitational waves among the solutions. As a result, on the one hand this approach requires a large number of initial data. On the other hand, the computational complexity of resulting approximate solutions depends on values of small parameters such as the Rossby-Kibel number, R_k . We recall that this number characterizes the ratio between the inertial force of the system and the Coriolis force. Hence, if R_k is large⁴, the Coriolis effect related to small-scale effects may be neglected, as is usually done in our everyday life, in spite of the rotation of the earth. However, if R_k is small, the complexity of approximating algorithms increases.

For many meteorological and oceanographic phenomena it is important to take into account the dynamic nature of such “small” parameters induced by the interaction of different space-time scales. At present, such an interaction is modelled on the basis of the classical law of viscosity that gives a connection between the stress σ and the speed $|\mathbf{v}|$ through the effective viscosity coefficient μ ,

$$\sigma = \mu \frac{\partial |\mathbf{v}|}{\partial z}. \quad (6.1)$$

For practical applications, (6.1) should be supplemented by the law of energy dissipation. When microturbulence is treated as a dissipative factor, we can always find a reasonable analogy between the motion of molecules and the motion of macroscopic elements of turbulent fluids. This idea was first used by Prandtl [42], who attempted to treat the case of turbulent momentum exchange in this way. He arrived at the mixing-length hypothesis (see, for example, [12])

$$K_v = l^2 \left| \frac{\partial \mathbf{v}}{\partial z} \right|, \quad (6.2)$$

that gives a connection between the coefficient of vertical diffusion K_v and the velocity field \mathbf{v} through the mixing length l . The connection between (6.1) and (6.2) should be provided by a scaling law. A major portion of current investigations in

⁴for example, when Ω is small or when L is relatively small

climate study is based on the scaling law of logarithmic type

$$\bar{u} = \frac{u^*}{k_v} \ln \frac{z}{z_0}, \quad (6.3)$$

where \bar{u} denotes the mean velocity (in the x-direction), u^* is the friction velocity, z_0 is the roughness parameter, and k_v is the von Karman constant. Recently, new theoretical and experimental evidence was given to confirm that this law may be inappropriate as an adequate description of turbulent processes [2, 3].

7 Long-term meteorological processes and non-adiabatic factors

The main difference between short and long range meteorological processes is that in the long-range the atmosphere cannot be regarded as a *closed system*, since it is a part of a bigger AAL system. Among the most important active layers is the ocean. After approximately 1-2 weeks the upper layer of the ocean has a substantial influence on atmospheric processes. As a result, one of the most important *initial conditions* in the model is the temperature field, in particular the temperature of the active layer of the ocean. At present, the practical availability of large datasets for such conditions has led to different ideas aimed at lengthening the period of validity of short-range models for the atmosphere. Initially, we can assume a constant temperature for the ocean, then use a slab-ocean model. Further, we can increase the number of formally *independent mathematical models* that can interact between themselves through message passing in a computational algorithm.

The main difficulty in the construction of long range models stems from the inadequateness of the adiabatic approximations. Conservation laws become approximate in nature, and one should take into consideration sources and dissipations of energy. In Section 5 we defined an approximation by model (5.1) that is based on the Obukhov hypothesis. Ultimately, the validity of approximations of this type is based on appropriate scaling laws. From the physical point of view, the adequate construction of the model essentially depends on taking into account dissipative effects and sources, including

- heat sources such as solar and terrestrial radiation;
- cloud dispersion and absorbtion;
- local/nonlocal boundary layer diffusion etc.

The solution of mathematical models subjected to physical parameterizations that take into consideration such dissipative effects and sources can only be approached numerically. Moreover, the quality of the algorithm will decisively depend on the adequateness of the parameterization of bio-chemico-physical processes in the mathematical model. This is why, without additional simplifying assumptions with respect

to non-atmospheric components of climate like ocean or sea ice, the “exactness” of conservation laws cannot be justified for any mathematical model. In practice, we always have to overcome difficulties arising from the approximate character of conservation laws in “proxy” climate models. Nevertheless, since the model is solved numerically, we can always use the idea of conservation on a finite grid [53, 5]. Indeed, the representation of conservation laws in mathematical models is of an approximate nature. For example, by a requirement on the vertical finite differences of the model to conserve the global integral of total energy in the absence of sources and sinks [5], we still neglect lack of conservation. In general, the stability conditions of the “proxy system” are not only different from those for the stability of the system itself, but they also are sensitively dependent on the degree of coupling achieved in the “proxy system”. The approximate character of initial and boundary data in models with a hyperbolic-type operator does not allow consideration of the atmosphere in long-term processes as a *closed system*. Mathematical difficulties for such an approach are obvious. If the atmosphere is a dependent component of the AAL system, then the questions “*how many such components are sufficient to adequately describe climatic processes*” and “*what are these other components*” have to be answered.

A wider mathematical “freedom” is allowed by looking at the evolution of states of the atmosphere as a random process $\omega(t)$. In this case it is possible to approach the task of studying the *possibilities* of the statistical extrapolation of this process using Kolmogorov’s hypothesis. Namely, a random process $\omega(t)$ describing the evolution of the turbulent flow in an environment with vanishing viscosity asymptotically approaches a Markov process for large t . From such a consideration it follows that the distribution of probabilities $P^t(d\omega)$ for $t > t_0$ may, in principle, be uniquely determined by the state $\omega(t_0)$, and not be dependent on the remote history of the process when $t < t_0$. Although the assumption of the *negligible viscosity approximation* can be justified on a finite grid⁵, validation of the original mathematical model is intrinsically connected with the processing of incomplete information, which requires an adequate formulation of scaling laws. It appears that, in the general case, the Kolmogorov-Obukhov scaling law for local structures [22, 39], that is typically used, becomes inappropriate for this purpose. We discuss these issues in the next sections.

8 Information exchange between components of mathematical models for climate

There are many “proxy climate” models which allow the simulation of interactions between different components by message passage in corresponding computational models. One of the models of this type is the NCAR CSM, where the original

⁵using, for example, the four-thirds Richardson’s law

problem of climate study is reduced to that of a *controled exchange of information between the model components* under the assumption that conservative properties⁶ can be preserved when exchanged between model components [5]. In general, this assumption can be justified numerically, and it is reasonable to hope that by improving the physical (chemical, biological) parameterization we can improve the correctness and reliability of “proxy climate” models. Clearly, such improvements, as well as improvements in hardware and software, may continue indefinitely. Hence, it is necessary to develop a strategy which permits an analysis of the trade-off between a level of coupling implied by achieved parameterization and a possible error induced by the incompleteness of available information.

In climate study, an “exact” realization of conservation laws is closely connected with the consistency conditions for “slow” motion that implies hydrostatic, (5.5), horizontal, (5.6), and geostrophic, (5.7), approximations. In reality these conditions are continuously violated. As a result, approaches to *filter out the fast waves from the solution* of coupled system of PDEs are quite restricted in their applicability to climate study. Such approaches are typically based on the Kolmogorov-Obukhov scaling laws for local structures [22, 39]⁷, and more generally on the local equilibrium hypothesis [38]. It is often the case that statistical field theory can be used to give a practical explanation of these hypotheses. However, for many complex dynamic system such as climate, mutual re-adjustment and self-adaptation of fields of different space-time scales is in the nature of the underlying processes. In such cases, the classical Kolmogorov-Obukhov scaling law may not lead to an adequate approximation of these processes [2, 3]. As we mentioned, the key point behind this fact is that the approximate character of initial data in models with a hyperbolic-type operator (such as the primitive equations of the hydrodynamic theory) does not allow consideration of the atmosphere in long-term processes as a *closed system*. Moreover, since the stability of any closed “proxy system” does not imply stability of the system itself, we need a trade-off strategy between coupling and stability concerns. To define such a strategy we have to appeal to the idea that a division between long-range and short-range depends on the definition of the *relaxation time* (or “build-up” period), τ (see (4.3)). This “coefficient of atmospheric efficiency” depends on the degree of coupling of atmosphere to its active layers that is ultimately defined by the problem solver or modeler. Mathematically, a formal division between “long” and “short” ranges is defined by the sign of the inequality between $t - t_0$ and τ :

$$\tau = \text{sign}\{t - t_0 - \tau\}. \quad (8.1)$$

The possibility of the existence of the two simultaneous limits

$$\tau \rightarrow 0^+, \text{ and } t - t_0 \rightarrow \infty \quad (8.2)$$

is often taken for granted as an a-priori mathematical assumption in investigations of complex dynamic system. Mathematical models based on this assumption are

⁶e.g. momentum, heat, freshwater

⁷we give more details in Section 10

characterised by a strong singularity at $\tau = 0$ and the parabolic features of the underlying dynamics. From the physical point of view such models are close associates of Fourier's original ideas on a diffusion mechanism for heat conduction. It is well-known, however, that the Fourier prediction may underestimate the peak temperature during a rapid transient period. Since experimental work on the wave behavior of heat transport [40] and theoretical work in this field [58], interest in hyperbolic-type models for processes that include diffusion is being dramatically increased. In the general non-linear case, such models preclude the assumption that small-scale phenomena can be extracted from a large-scale flow. As a result, one should overcome the problem of the approximate character of conservation laws in mathematical models of the "proxy systems". For example, heat in the "proxy climate system" might be conserved only under the assumption that it is neither gained nor lost at the top of the atmosphere. In fact, *it is not conserved* under inadequate parameterization⁸. Furthermore, in general the model can conserve energy only if we neglect the lack of conservation due to *a-priori* regularity assumptions for our approximations. From the physical point of view, relaxation of these assumptions requires more careful examination of the non-local features of boundary layers [13]. We also note that in CCM-type models⁹ the vertical advection of temperature is not used to conserve mass/energy. However, it is well known that the interaction between the vertical semi-Lagrangian approximations and the convective parameterization may seriously affect system predictability. In the general case, *the a-priori assumption on the existence of conservation laws for "proxy" climate systems leads to a-priori regularity assumptions for "exact" solutions of mathematical models.*

Since the construction of mathematical models for the evolution of thermodynamic systems has to be undertaken under analysis of uncertainty and the processing of incomplete information, it is always important to investigate the stability of associated computational models. However, in some cases it is possible to approach the issues of the well-posedness of the mathematical model without the investigation of stability in a computational sense. Indeed, mathematical analysis of the model itself can often be reduced to a general stochastic control problem. In turn, this problem can be often associated with a PDE of the Hamilton-Jacobi-Bellman type [9]. The quality of mathematical models based on such PDEs is essentially determined by the quality of the *approximation of the system Hamiltonian*, and *approximate initial data* for the models. In this case, the adequateness of the model to the real-world situation will be completely defined by the smoothness assumption on the sought-for solution. In the case where it is assumed that initial data for the model can be given exactly, the semi-continuity assumption [54] is natural. However, with application to real dynamic systems, such models can be reasonably validated if we are able to analyse the distinction between

⁸for example, if the long wave radiation in the atmosphere component uses the average sea surface temperature

⁹we use a version of CCM in our computational experiment

- external error growth due to model deficiencies (such as physical parameterization), and
- the internal error growth due to mathematical assumptions (resulting from the unstable “self-growth” of the initial data errors).

The level of parameterization defines an *upper bound* for such an internal error. This provides a way to investigate the connection between coupling in mathematical models and the level of uncertainty in the model prediction (see [49, 8] and references therein). On the other hand, such a bound introduces hyperbolic features into the model [38].

9 Hybrid Eulerian-Lagrangian models and numerical schemes for transport effects

Together with a wide use of the hybrid vertical coordinate, the interest in the hybrid Eulerian-Lagrangian type of mathematical models for climate study increases. In turn, this leads to attempts to implement into modern climate system models *semi-Lagrangian advection approximations* instead of the standard Eulerian approximations. One of the advantages of the semi-Lagrangian version is that in many cases it allows us to relax the normal advective Courant-Fridrichs-Lowy (CFL) stability condition. In fact, it is well-known that for standard spectral models the typical resolution of the model may lead to instability¹⁰ if one applies a standard time-step. To obviate this problem, *limited filtering* is often used on the top model layer¹¹. The interest in the semi-Lagrangian type of models is stimulated by the claim (see [61] and references therein) that application of the semi-Lagrangian version may not only exclude the above-mentioned filtering, but also eliminate the normal advective CFL time-step restriction. However, it has to be recalled that the main problem with a semi-Lagrangian formulation consists of the fact that the result of interpolation with pointwise values is *not a-priori conservative*. As a result, *long-term simulation* can be seriously affected. We should admit that both semi-Lagrangian and EST versions have a serious deficiency. Small-scale features in the solutions may be underestimated more in semi-Lagrangian versions, whereas the EST approach has to deal with Gibbs phenomenon and spectral truncation. Hence, in practice, the EST methods can be successfully applied to adiabatic approximations, whereas the semi-Lagrangian approach is more natural for problems related to the advection of fields with *large horizontal gradients*¹². We should also take into consideration the

¹⁰which may be observed in the Southern Hemisphere

¹¹such models are often referred to as Eulerian with Spectral Transform (EST), though one realises that they are not Eulerian because the water vapor transport is usually treated in a semi-Lagrangian manner

¹²for example, when modelling water vapor transport

fact that advantages in the stability of semi-Lagrangian advection schemes may be lost when we use hybrid Eulerian-Lagrangian models.

The author believes that the most efficient schemes must not require an explicit “subgrid-scale turbulence” parameterization and spacial filtering. Instead of the classical semi-Lagrangian or EST approaches it is reasonable to use 1D-Flux-Corrected-Transport (FCT) schemes. For climate study, this idea was used in [28]. The same idea was used in a different area of application to avoid explicit turbulence parameterization in [34]. It is straightforward to apply this idea to 3D transport modelling by means of time splitting in a manner explained, for example, in [32]. For climate models, zonal transport at high latitude can be split in time to satisfy the local CFL restriction, whereas at low latitude we can use a larger step. Such schemes provide an increasing accuracy with increasing resolution even when *discontinuities or steep gradients* are encountered. This property is extremely important to overcome the singular nature of the spherical coordinate system near poles, especially for systems with incomplete information, including those for which additional data from observations may be added in stages.

10 Turbulence, vanishing viscosity, and scaling laws

The core of hydrodynamic theory is the system of the Navier-Stokes equations. The *ensembles of solutions* of these equations is usually associated with the behaviour of turbulence. Many theoretical works in this field are concentrated on the limiting case of vanishing viscosity in the Navier-Stokes system (the inviscid limit).

From the physical point of view, the system itself is a mathematical expression of conservation laws. Such laws are obtainable, at least in principle, from a hyperbolic system by adding a small viscosity coefficient. As follows from [25] (see also [26]), in the limit of vanishing viscosity one expects to be able to recover entropy solutions of the *original hyperbolic system*. The whole procedure is based on the assumption that a formal mathematical transformation from the hyperbolic system to the Navier-Stokes system preserves conservative properties. Naturally, if the original system is conservative, then after the vanishing-viscosity-limit transformation the system remains conservative. However, if we add to the original hyperbolic system a *dispersive term*, then we cannot expect that solutions of this modified system in the limit of vanishing dispersion are well-behaved. Existence or non-existence of solutions of such a modified system depend on the *regularity of solutions of the original hyperbolic system*. Therefore, though the Navier-Stokes system is parabolic in nature [38], in the general case it has both *dispersive and hyperbolic features*.

These features complicate the quantification of the behaviour of ensembles of the solutions of this system. Moreover, in the general case, the solutions of this system are non-stationary, and depend on the initial conditions for the model. From the theoretical point of view, not much can be said about the accuracy in the definition of initial conditions for complex dynamic systems. One of the standard

approaches to the problem of investigating such systems is to start from stationary solutions, and try to approximate the time-average non-stationary solutions by averages of stationary statistical solutions. Such attempts are based on the ergodic hypothesis. From the physical point of view, such stationary random solutions are a generalization of steady state in the N-body problem. This generalization requires a well-defined concept of equilibrium. In classical theory, the equilibrium is associated with a “microcanonical distribution” obtained either by the equi-partition over the set of approximately equal energy systems, or by the Gibbs procedure. If the time-invariance of associated probabilities can be preserved, near-equilibrium processes can be investigated through equilibrium properties of the system using the approach introduced in the practice of mathematical modelling by Langevin (see, for example [14]).

If turbulence is regarded as a perturbation of an equilibrium, then more precise definitions of “equilibrium” and “perturbation” are required. If the equilibrium is obtained from the idea of the equi-partition of the set of approximately equal energy systems and the initial energy density is finite, then the resulting ensemble does not possess the ergodic property [2]. This type of equilibrium was first constructed in [15, 27] using the Fourier expansion. In the limit of vanishing truncation of the Fourier series, the underlying process can be described by a set of functions that are *almost nowhere differentiable*. From the physical point of view the process is characterised by *the infinite energy density*. If the topological space in which the set of these functions can be embedded is a *complete linear metric space*, then under quite general assumptions [10] there exists no measures which are at least quasi-invariant relative to all translations, except for those identical to zero measure. Taking this into account, it is natural to regard real turbulent processes as those that are far from the Hopf-Lee equilibrium [2]. Mathematical investigations of such processes are intrinsically connected with the Fourier transform, and the theory of probability due to classic results of harmonic analysis¹³. At the core of such investigations from the physical point of view is the assumption that the provision of energy at large scales is the dominant effect which *can be modelled* through the limit of vanishing viscosity at small scales. There are two main types of mathematical models that are intrinsic to this approach:

- stationary models with exactly given boundary conditions;
- non-stationary models with exactly given initial conditions.

Both types of models have a common feature. Namely, non-linear terms in the models can be treated as a perturbation expansion ordered by a *small time-independent parameter*. In climate system study the Rossby-Kibel number often plays the role of such a parameter, as we explained in Section 5. This approach encounters serious difficulty when the viscosity decreases and equilibrium is understood in the Hopf-Lee sense. In this case the singularity of turbulence increases simultaneously with the decrease in viscosity. Mathematically, this difficulty can be

¹³the Bochner theorem and its generalization to Banach spaces [44, 51]

formally overcome by using the Kolmogorov-Obukhov scaling law [22, 39], which implies a certain character for energy distribution between scales. That is, an inertial cascade of energy from the “whirling” scales to the dissipation scales can be represented by the Kolmogorov spectrum [2]

$$E(k) = C\epsilon^{2/3}k^{-5/3}, \quad (10.1)$$

where k is the wave number, ϵ is the rate of energy transfer across the spectrum, and C is an absolute constant. Recent results on deviations from this law can be found, for example, in [2]. Such deviations may be expected when *dispersion* is intrinsic to the model. This is always the case for mathematical models of real phenomena or processes where physical (chemical or biological) parameterization cannot be performed with infinite precision. In the general case, it is not enough to consider the limit of vanishing viscosity in order to adequately describe turbulence. We also need information of the character of dispersion [26]. Such a *correlation between viscosity and dispersion* has stimulated the search for different principles on which the statistical theory should be built [2].

The other interpretation of turbulence which has been recently proposed is based on the assumption of a small perturbation of a suitable Gibbsian equilibrium. However, if we accept the Gibbs hypothesis (see [35] and references therein) the nature of the convergence of the probabilities in the limit of vanishing viscosity remains open. The answer to this question is kept in the approximation of the Hamiltonian. In fact, the Gibbs probability distribution is defined as the probability of a collection of states by the Lebesgue integral of

$$C_a = 1/n_0 \exp(-\beta H), \quad \beta = 1/k_b T, \quad (10.2)$$

with respect to the Liouville measure. In (10.2) the notation is standard, that is k_b is the Boltzmann constant, T is the temperature of the system (macroscopic temperature), n_0 is a normalizing factor and H is the system Hamiltonian. The definition of the system Hamiltonian is a hierarchically approximating procedure. This implies a procedure for obtaining conditions that single out the canonical ensemble measure from the class of all probability measures on the phase space of the system. Such conditions determine stability conditions of the model [36].

If the scaling law is agreed upon then the nature of turbulence can be studied through different *approximations of the system Hamiltonian*. The connection between the approximate character of the Hamiltonian and the scaling law can quantify turbulence numerically whenever the *law of dispersion* is established. As a result, the quality of mathematical models for turbulence is essentially determined by an adequate physical parameterization of the model that is linked to the hierarchical approximation of the system Hamiltonian.

11 Differential mathematical models for the climate study

Many rigorous mathematical results in the investigation of climate models were obtained for the barotropic atmosphere. In this case, there is in the governing system of equations at least one equation that relates two thermodynamic variables on the basis of an individual particle from time to time. An alternative formulation can be given by using a *piezotropic equation*, which allows us to relate two thermodynamic variables from one spacial point to another at a given moment of time. Since statistical spacial data for climate study is typically sparse, the latter case leads to the weak rather than differential formulation of the problem. In general, non-linear relationships between thermodynamic parameters which define the model suggest solution by numerical methods. Then, the quality of the underlying algorithm is completely determined by the correspondence of the model to the real climate.

We refer to the paper of J. Smagorinsky [52] on the basic experiment performed in collaboration with J.G. Charney, N.A. Phillips, and J. von Neumann. It is not reasonable to review here all subsequent steps in the development of the model that we use in this paper. The description is well-documented and the appropriate references related to the NCAR CCM3 model can be found in [21]. Instead, we give below the set of differential equations governing the atmosphere that is fundamental in climate study. It consists of five equations that relate the three components of wind $\mathbf{v} = (u, v, \omega)$, pressure p , density ρ and temperature T , namely

- the equations of motion

$$\left\{ \begin{array}{l} \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} + \omega \frac{\partial u}{\partial z} = 2\Omega(v \sin \varphi - \omega \cos \varphi) - \frac{1}{\rho} \frac{\partial p}{\partial x} + \frac{F_x}{m}, \\ \frac{\partial v}{\partial t} + u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} + \omega \frac{\partial v}{\partial z} = -2\Omega u \sin \varphi - \frac{1}{\rho} \frac{\partial p}{\partial y} + \frac{F_y}{m}, \\ \frac{\partial \omega}{\partial t} + u \frac{\partial \omega}{\partial x} + v \frac{\partial \omega}{\partial y} + \omega \frac{\partial \omega}{\partial z} = 2\Omega u \cos \varphi - \frac{1}{\rho} \frac{\partial p}{\partial z} - g + \frac{F_z}{m}, \end{array} \right. \quad (11.1)$$

- the continuity equation

$$\frac{\partial \rho}{\partial t} + u \frac{\partial \rho}{\partial x} + v \frac{\partial \rho}{\partial y} + \omega \frac{\partial \rho}{\partial z} = -\rho \left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial \omega}{\partial z} \right) \quad (11.2)$$

- the thermodynamic equation

$$d\varepsilon = c_p dT - \frac{1}{\rho} dp, \quad (11.3)$$

- and the state equation

$$\rho = \rho(p, T). \quad (11.4)$$

It is customary to use the Clapeyron equation $\rho = p/(RT)$ as the state equation. In this case the question of an adequate model of the latent heat of the phase transition is of primary importance because the gas constant R is different for different phases. In (11.3) $d\varepsilon$ is the heat added per unit mass, and the field of external forces in (11.1) is written excluding pressure-gradient forces (the pressure gradient is directed from high towards low pressure). In essence, equations (11.1)-(11.4) are *prognostic*. They provide a theoretical possibility for the mathematical forecast in climate study if we insert observed climatic conditions at a given time. A built-in physical parameterization splits the system (11.1)–(11.4) into prognostic equations and *tendency equations*.

Let us briefly recall typical assumptions made for simplifications of the system (11.1)–(11.4). The adiabatic approximation which we discussed in Section 4 can be obtained by setting $d\varepsilon = 0$, which provides a sufficient approximation for short range prediction. Moreover, if the atmosphere is assumed to be *barotropic*, the density at each point can be determined solely by the pressure at that point. In this case we have

$$\rho = \rho(p), \quad T = T(p), \quad (11.5)$$

(the second relationship follows from the Clapeyron equation).

Furthermore, if the motion is assumed to be horizontal, we have

$$\left\{ \begin{array}{l} \chi v = \frac{1}{\rho} \frac{\partial p}{\partial x}, \\ \chi u = -\frac{1}{\rho} \frac{\partial p}{\partial y}, \\ g = -\frac{1}{\rho} \frac{\partial p}{\partial z} + 2\Omega u \cos \varphi. \end{array} \right. \quad (11.6)$$

If we neglect the Coriolis term on the right-hand side of the last equation in (11.6), the equation turns into the *(hydro)static equation*. The first two equations of the system are *geostrophic wind equations*¹⁴ which may provide a good approximation in middle and high latitudes. This approximation simplifies numerical procedures because, for the barotropic atmosphere, geostrophic wind does not increase with height [11].

From the system (11.6) for the hydrostatic approximation pressure can be excluded. This results in the *thermal wind equations*

$$\left\{ \begin{array}{l} \frac{\partial v}{\partial z} = \frac{g}{\chi T} \frac{\partial T}{\partial x} + \frac{v}{T} \frac{\partial T}{\partial z}, \\ \frac{\partial u}{\partial z} = -\frac{g}{\chi T} \frac{\partial T}{\partial y} + \frac{u}{T} \frac{\partial T}{\partial z}. \end{array} \right. \quad (11.7)$$

Again, numerically it leads to essential simplifications, since in the barotropic atmosphere the vertical temperature-gradient term is equal and opposite to the horizontal

¹⁴or gradient balance equations

temperature gradient term. Of course, this gradient is not negligible in the general case.

In the baroclinic atmosphere a strong dependency between the horizontal temperature gradient and the vertical wind shear requires an appropriate choice of *the vertical coordinate*. The interdependency of different components of climate becomes important. However, the majority of implementations of the hydrological cycle into model (11.1)–(11.4), as well as the development of global general circulation models¹⁵, are essentially based on *the hydrostatic equilibrium assumption* [30, 31, 12]. The necessity for the development of a non-hydrostatic type of model has been realised during recent years [55].

Having included more than one climate components into a unified mathematical model, it becomes increasingly important to adequately formulate the hypothesis on *subgrid scale vertical/horizontal mixing*. Currently this hypothesis is formulated on the basis of the von-Karman-Prandtl logarithmic law (6.3) in the region of wall-bounded turbulent shear flow [59, 42, 43]. Since in general this law may lead to an inappropriate scaling [2, 3], more general laws for the interaction of space-time scales should be applied for the climate study.

12 Phase transitions and the algorithmic stability

When temperature changes, water vapor departs from the ideal conditions, making an adequate model of water vapor transport one of the most important and difficult problem in climate study. The thermodynamics of water vapor and moist air is closely connected with the problem of *latent heat and phase transitions*. Even if the temperature of a substance remains constant, whenever this substance changes phase (evaporates, melts, condenses, freezes etc) a quantity of heat, called *the latent heat of the phase change*, must be supplied to or taken away from the substance. The quantification of the latent heat is based on the concept of entropy, and is often performed by using the Clausius-Clapeyron equation that relates the saturation of vapor pressure to the latent heat of a phase transition. Conceptually, this equation together with the state equation (11.4) is time-independent. This leads to an approximation of the mathematical model (11.1)-(11.4) whenever a physical parameterization is applied.

In most latitudes at most times of the year the atmospheric pressure and temperature vary continuously with time. As a result, *the geostrophic balance is never reached and maintained no matter how small the time-interval is assumed*. Rather, we observe a *continuous re-adjustment of the fields* with changing pressure and temperature fields. This requires the formulation of tendency equations that in turn require some a-priori knowledge of the vertical structure of meteorological fields. On the other hand, the knowledge of the vertical structure of the meteorological pro-

¹⁵that include in addition to the atmosphere other climate components such as ocean, land surface, sea ice

cesses is a major output in integrating the prognostic equations. Hence, although prognostic equations can provide *a-posteriori information*, they must always be supplemented by the tendency equations (which are based on *a-priori information*) in order to form a *closed system of mathematical equations*. The tendency equations are typically based on additional physical hypotheses (like hydrostaticity), and are approximate in their nature. The original system (11.1)-(11.4) is always replaced by its approximation, not only because of inevitable approximations of the functions F_x, F_y, F_z and initial and boundary conditions, but also because of the approximate nature of the equations (11.3) and (11.4) for any specific model. Naturally, this leads to attempts to improve such approximations by "building-in" to the model other components of climate such as the ocean, land-surface, sea ice. In these cases the vertical structure of meteorological processes cannot be defined in the simple manner of the barotropic approximation with the altitude typically defined by

$$z(x, y, t, p) = z_0(x, y, t) \psi_0(p). \quad (12.1)$$

On the other hand, if in the general baroclinic case the altitude is approximated in a multilevel manner as

$$z(x, y, t, p) = \sum_{i=0}^N z_i(x, y, t) \psi_i(p), \quad (12.2)$$

then the functions $\psi_i(p), i = 0, 1, \dots, N$ should be chosen on the basis of *a-priori information*. For any finite number N , each of the functions $z_i(x, y, t)$ becomes a parameter of a given mathematical model that can, in principle, be expressed in terms of the values of $z(x, y, t, p)$ at the given level of pressure p_i . Hence the question arises as for the *optimal a-priori choice* of the functions $\psi_i(p)$. Such a choice is *multilevel* by its nature. It requires an interpolation between given levels on the basis of some qualitative *a-priori assumptions*. These assumptions have to ensure the *well-posedness of the model*.

We recall that in short-range climate study, neither external energy generation nor dissipation of energy due to internal processes are taken into account. Let us assume now that h is the smallest scale of motion described by a discretized system of prognostic equations. Firstly, we note that the consideration of a discretized system is natural, at least due to imperfections in the measurement of meteorological fields. In reality, even at very high resolution, the scale h still exceeds the scales of regions of energy dissipation. However, if we neglect the energy dissipation then the energy transferred between scale spectrum finally reaches the scale of the order h and accumulates there without a dissipation. As a result, the non-homogeneities of meteorological fields with scales of the order h may increase in time when $t - t_0$ increases, inducing *non-linear instability*. This leads to a continuous correction of the model by a more precise definition of the law of dissipation. From the mathematical point of view, whenever

$$h \rightarrow 0 \text{ and/or } t - t_0 \rightarrow \infty, \quad (12.3)$$

the dependency of h on τ and/or the dependency of $t - t_0$ on τ becomes important.

13 Computational decoupling

The system (11.1)-(11.4) is a strongly coupled system of mathematical equations. Its solution cannot be obtained by analytical approaches unless substantial simplifications are made. Such simplifications may dramatically influence the validity of the final result. On the other hand, any specific physical parameterization of the model also implies an inevitable mathematical approximation as we explained in Section 12. Due to such an approximation, conservative properties of the original system may only be preserved approximately. The accuracy of approximations of conservation laws is determined by *the physical parameterization and the degree of coupling in the original mathematical model*. Essentially, any physical parameterization that is “built-in” to a mathematical model *splits the model into components*. However, in principle, the connections between such components can be restored computationally. The quality of the restoration depends on the number of model components and the quality of the physical parameterization with respect to the real processes and phenomena.

In the Climate System Model developed by NCAR there are four main components, namely atmosphere, ocean, land, and sea-ice. The connection between these components are realized using the Flux Coupler code [5]. This code is constructed under the assumption that conservative properties can be preserved for momentum, heat, and freshwater under message passing. In turn, this assumption inevitably leads to an approximation of the original model. Even if we assume that initial data (initial and boundary conditions) are given with an appropriate precision, additional assumptions for the energy dissipation law at the top of the atmosphere should be made by *a-priori arguments*. This implies an approximate character not only for the mathematical expression of physical laws, but an approximate character for the physical parameterization of the model as well.

Reasonable *a-priori assumptions* may be derived on the basis of experimentation and observations. They can provide a tool for the analysis of *the adaptive re-adjustment of meteorological fields*. However, a formal expression of such adaptive procedures requires some *a-posteriori arguments*. Such *a-posteriori arguments* are usually based on *the concept of continuity* [17]. Having both *a-priori assumptions and a-posteriori arguments* we can, in principle, validate the model ensuring its stability. In the general case the validation of mathematical models for complex dynamic systems can only be conducted with incomplete information. As a result, in reality it is practically impossible to achieve 100% reliability of the model. However, it is possible to achieve a *balance between the reliability of the model subjected to the physical parameterization and the efficiency of a numerical algorithm for its solution*. The procedure for achieving such a balance requires *the adaptive error control* that is based on *a-posteriori information about the computed solution*. In turn, the processing of such information requires *a-priori information on the exact solution*. Under quite general assumptions problems of this type can be formalized mathematically in the form of a hyperbolic type partial differential equation with

respect to the control function [35]. From the practical point of view, the well-posedness of the original model depends on a *constant of hydrodynamic stability*, C^s , that quantifies stability properties of a dual problem with coefficients which depend on exact and computed solutions as well as on the period of time $T = t - t_0$ during which the model is integrated [18]. In this interpretation, the validation of the original mathematical model is eventually determined by the evaluation of the quantity C^s , which for climate system models is closely associated with the relaxation time τ defined by (4.3). We will address the issues of such an evaluation elsewhere. Here we note that the foundation of theory in this direction was laid by the works [17, 7, 18] (see also references therein).

14 Numerical results.

In order to conduct a computational experiment we have used the National Center for Atmospheric Research model [1]. It can be run in three main modes, namely interactive, batch and message passing. Only the first two were used in our experiment. The original file contains a C-shell script "setup" that can be used for the configuration of the model (type of dynamics, resolution etc). The "setup" creates a directory with the configuration specific name, for example, in one of our cases it was "cray.t42.spectral.som/". On a SUN station with the SunOS operating system at the CIAM, University of South Australia we compiled the model with the command: "make sunos". This creates the executable file "ccm3bin" in a subdirectory "run/". Boundary datasets were taken from the NCAR WWW domain in the IEEE binary format. In the interactive mode the "setup" generates two standard namelists for initial and restart runs that can be used to tune in the model. The namelist files can be written at the discretion of a user as described in [1], p.27-44. We also note that running the model on SUN SPARC stations an increase of the stack size is often necessary, subject to the resolution used.

When we have to perform the simulation for a longer period it is convenient to run the model in the batch mode. The Fujitsu VPP-300 supercomputer at the Australian Supercomputer Facility was used for such a simulation. The VPP system is a distributed memory machine. Availability on this machine of the vectorizing and parallelizing UXP/V Fortran-90 [56, 57] compiler and other software capabilities makes this computer effective for the high-speed computation required for a CCM3 run. In this computational experiment only one processor was used to run a vectorized version of the CCM3 code in standard and slab ocean versions. A few bugs reported recently through the CCM-Users E-Mail Group were fixed. For a parallelized code the message passing using the PVM facility should be implemented, which can be seen as a future development of this work.

Below we present typical outputs obtained as a result of climate simulation. In this paper we have not attempted to investigate the error of this simulation. As follow from the above discussion, the total error consists of three parts:

- the error of initial data at the start of computer simulation,
- error of the finite set of differential equations in the description of climate, and
- the error of the numerical algorithm that is used.

For such models as the NCAR CSM, the total error obtained from the contribution from all three sources is practically infeasible. Instead, we are currently developing a technique for the evaluation of such an error for a simplified model. The purpose of numerical results presented here is to give a comprehensive graphical interpretation of several physical fields that have been computed and can be used for future analysis.

In the simulation of climate we used the standard input datasets ([1], p.45-50). As the main output, the model generates so-called *history files* that are in a binary format. They provide the information on a set of temporal samples. Field values at any given moment of time correspond to different latitudes ([1], p.51). Different plotting programs can be used for the interpretation of the results of outputs. We used a modified version of the code developed at the Global Change Research Center, Portland State University by Gerhard W. Gross whose help is gratefully acknowledged. This code reads history files and plots them using the GNUPLOT Plotting Program. In Figure 1 - 5, typical initial distributions are presented for the following fields respectively

- surface geopotential in m^2/s units;
- surface pressure in Pa units;
- zonal wind component in m/s units;
- meridional wind component in m/s units;
- sea-surface temperature field in C units.

In Figure 6 and 7 zonal and meridional components of the wind are presented at vertical level 15. The temperature field is presented in Figure 8 for vertical level 15 in K units. Finally, the water vapour field at the vertical level 15 in Kg_{H_2O}/Kg_{air} units is presented in Figure 9. The complete Master Field List of the NCAR CCM3 model and available options for the output of the model can be found in [1], p. 45-68.

15 Conclusions and future directions

The coupled simulation for climate system models provides an efficient tool for climate study. Moreover, the concept of coupling in modelling complex dynamic systems reflects one of the most general ways of implementing new effects and new

CLIMATE MODELLING WITH CCM3: FIELD PHIS (lat,long) Vert level = 1 Time Step = 1

PHIS

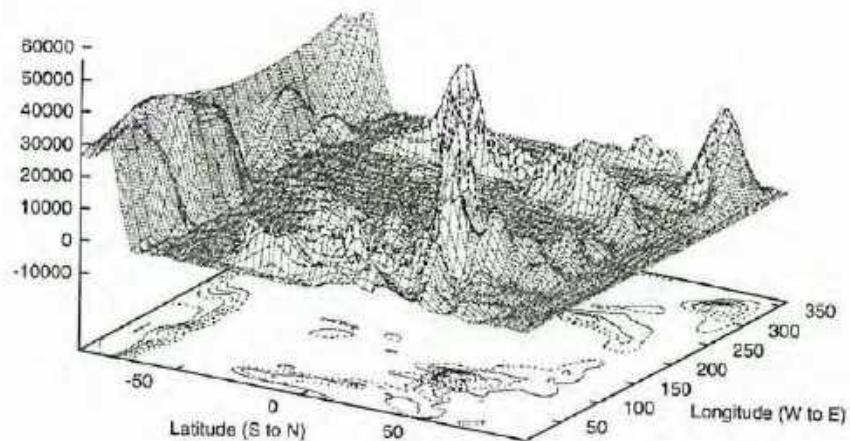


Figure 1: Surface geopotential (PHIS).

CLIMATE MODELLING WITH CCM3: FIELD PS (lat,long) Vert level = 1 Time Step = 1

PS

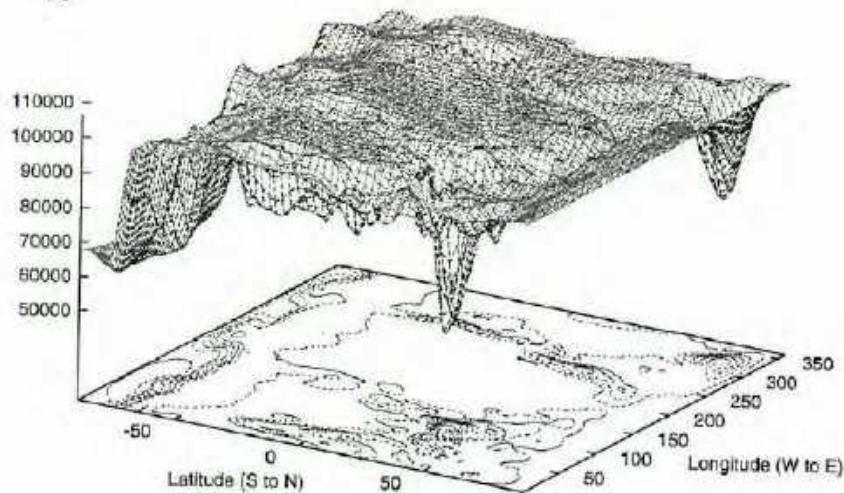


Figure 2: Surface pressure (PS).

CLIMATE MODELLING WITH CCM3: FIELD U (lat,long) Vert level = 1 Time Step = 1

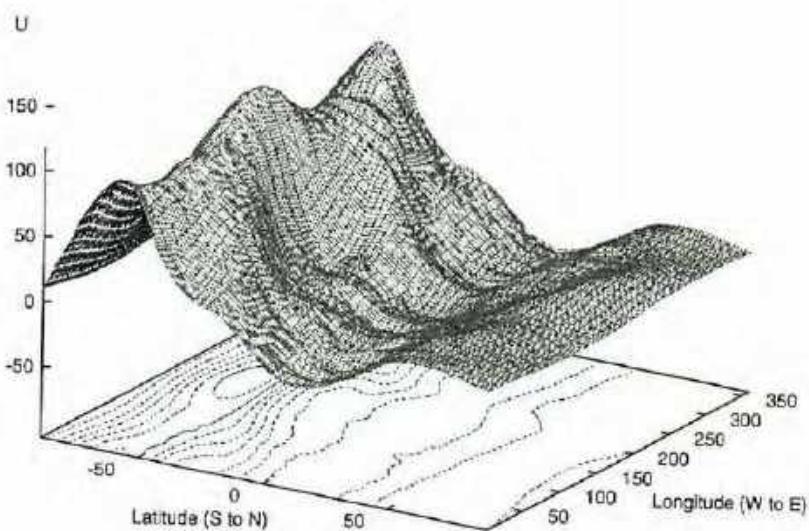


Figure 3: Zonal component of the wind (U, vertical level 1).

CLIMATE MODELLING WITH CCM3: FIELD V (lat,long) Vert level = 1 Time Step = 1

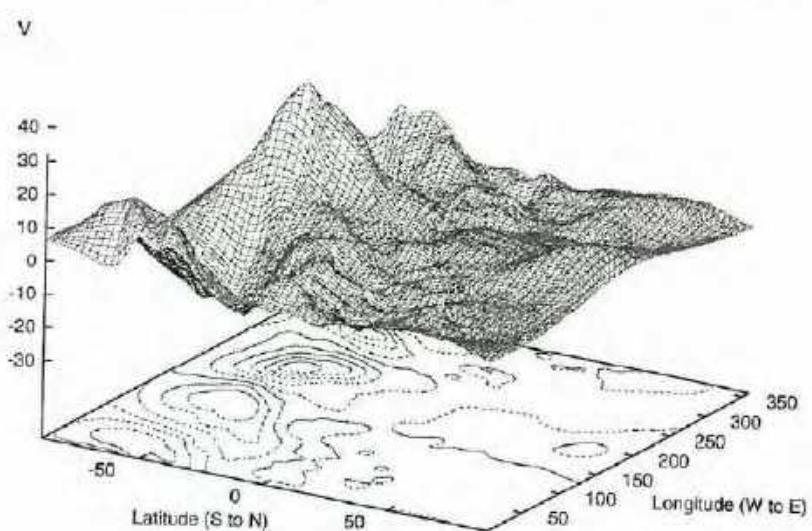


Figure 4: Meridional component of the wind (V, vertical level 1).

CLIMATE MODELLING WITH CCM3: FIELD SST (lat,long) Vert level = 1 Time Step = 1

SST C

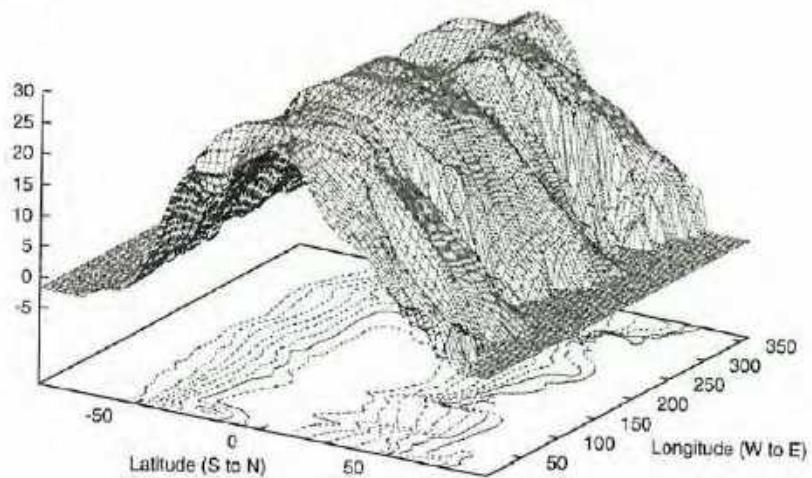


Figure 5: Sea-surface temperature distribution (SST).

CLIMATE MODELLING WITH CCM3: FIELD U (lat,long) Vert level = 15 Time Step = 1

U

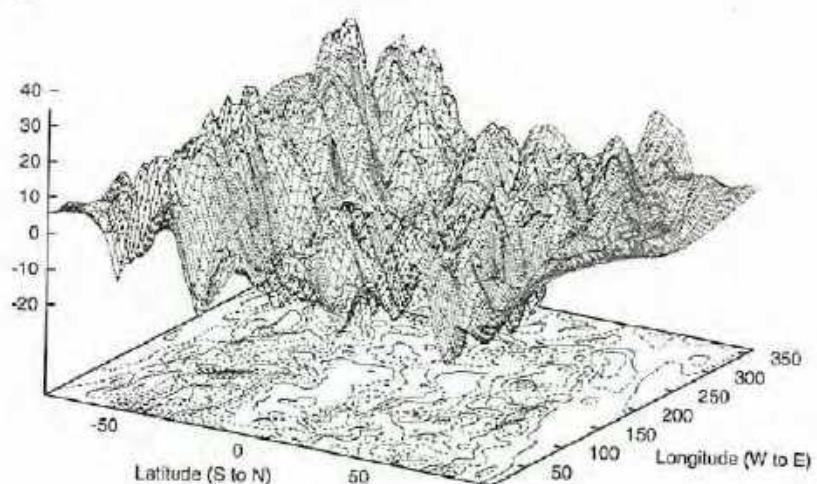


Figure 6: Zonal component of the wind (U, vertical level 15).

CLIMATE MODELLING WITH CCM3: FIELD V (lat,long) Vert level = 15 Time Step = 1

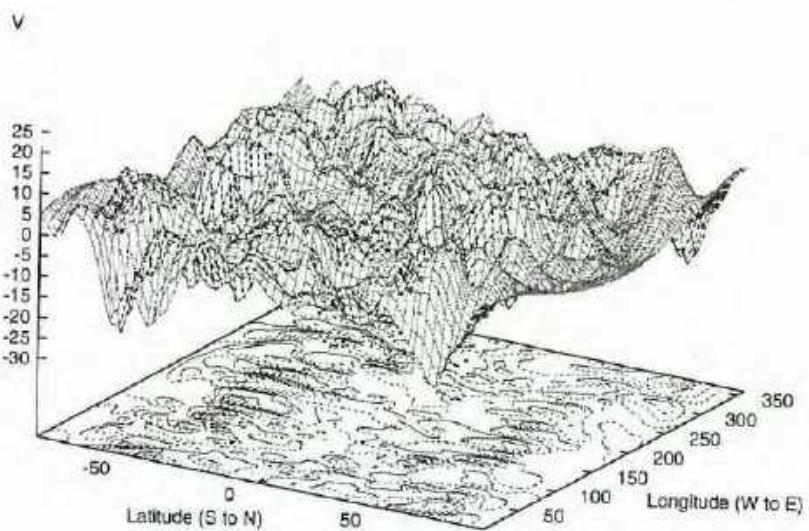


Figure 7: Meridional component of the wind (V, vertical level 15).

CLIMATE MODELLING WITH CCM3: FIELD T (lat,long) Vert level = 15 Time Step = 1

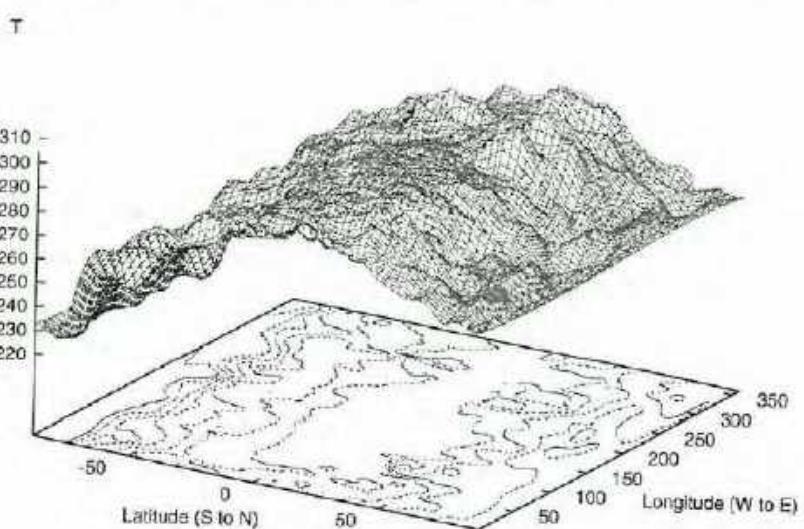


Figure 8: Temperature field at vertical level 15.

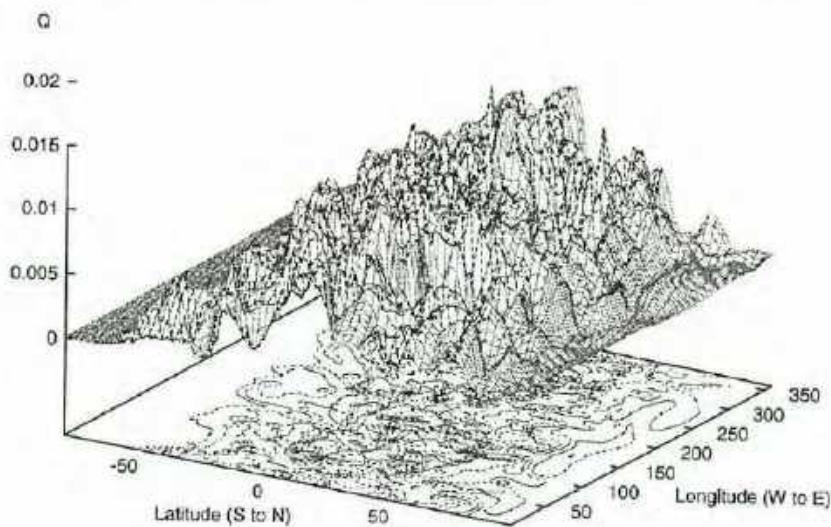


Figure 9: Water vapour at vertical level 15.

information into mathematical models. However, the refinements of the approach based on coupling procedures for such complex systems as climate may continue indefinitely. The two natural ways to meet the arising challenge were discussed in this paper. First, it is natural to start with a *finite set of independent models* and try to couple them by informational message passing using certain physical principles such as conservation laws. The NCAR Climate System Model is of this type. We presented several numerical examples obtained on the basis of this approach. For the NCAR CSM the set of independent components consists of four mathematical models for atmosphere, ocean, land surface, and sea ice. Although the conservative properties of the whole system cannot be guaranteed in general, certain key components, such as momentum, heat, freshwater, can be preserved *numerically*. The quality of such models depend on the dynamics of the error propagation that can be controlled by the Coupler Code. With additional information becoming available, this approach requires an increasingly complex code. In the end, we have to deal with the coupling phenomenon from the very beginning of the process of modelling. On the other hand, the sparse character of available informational datasets for such complex dynamic systems as climate makes *the concept of independency* for the model components natural, at least at the initial step of the process of modelling.

The dilemma between independence and coupling has led us to the necessity of considering another approach to the modelling of complex dynamic systems. We came to the conclusion that such an approach has to be based not only on a-priori information about the system (that is incomplete in its nature) but also on a-posteriori

information. This allows us to construct numerical algorithms that can be analyzed in the traditional manner of using a-priori information, but the procedure of construction has to be based on a-posteriori information. In this case *the choice of the norm for the error control is model-specific, being influenced by the physical parameterization of the mathematical model*. For any given physical parameterization, conservation laws may be implemented only approximately into mathematical models of complex dynamic systems. As a result, the standard energy norms may not provide an appropriate choice for the error control in mathematical models of such systems. This idea is the basis for the future development of the presented work. Under a fixed degree of coupling and a given physical parameterization we need a scale of a-priori and a-posteriori estimates in a spectrum of norms to ensure the stability of the model. Numerical algorithms based on such estimates allow an adaptive error control within the chosen spectrum. In turn, this allows the construction of adaptive computational codes that can be effectively used in the study of complex dynamic systems with many transient states.

ACKNOWLEDGEMENT.

The author wishes to acknowledge the support of the School of Mathematics at the University of South Australia. The author thanks Dr Murray Dow from the Australian Supercomputer Facility for permission to use a vectorized version of the CCM3 model to obtain a series of computational results on the VPP-300 supercomputer. The support of the leader of the Environmental Modelling Research Group at the Centre for Industrial and Applied Mathematics Professor J.Filar and the help of his collaborators, Mr P.Gaertner, Dr J.Day, and Dr S.Lucas is also gratefully acknowledged.

References

- [1] Acker, T.L., et al User's Guide to NCAR CCM3, *NCAR TN-421*, May, 1996.
- [2] Barenblatt, G.I., and Chorin, A.J. Scaling laws and vanishing viscosity limits in turbulence theory, *Department of Mathematics, University of California, Berkeley*, 1996.
- [3] Barenblatt, G.I., Chorin, A.J., and Prostokishin, V.M. Scaling laws for Fully Developed Turbulent Flow in Pipes: Discussion of Experimental Data, *Department of Mathematics, University of California, Berkeley*, 1996.
- [4] Briegleb, B.P. Delta-Eddington Approximation for Solar Radiation in the NCAR Community Climate Model, *Journal of Geophysical Research*, Vol. 97, No.D7, 1992, 7603-7612.
- [5] Bryan, F.O., et al The NCAR CSM Flux Coupler, *NCAR TN*, May, 1996.
- [6] Chang, P., and Philander, S.G. A coupled ocean-atmosphere instability of relevance to the seasonal cycle, *Journal of the Atmospheric Sciences*, 51, No.24, 1994, 3627-3648.
- [7] Eriksson, K., and Johnson, C. Adaptive finite element methods for parabolic problems V: long-time integration, *SIAM J. Numer. Anal.*, 32, No. 6, 1995, 1750-1763.

- [8] Farrell, B.F. Small Error Dynamics and the Predictability of Atmospheric Flows, *Journal of the Atmospheric Sciences*, 47, No. 20, 1990, 2409-2416.
- [9] Fleming, W.H., and Soner, H.M. Controlled Markov Processes and Viscosity Solutions, *Springer-Verlag*, 1993.
- [10] Gelfand, I.M., and Vilenkin, N. Ya. Generalized Functions. Applications to Harmonic Analysis, *Academic Press*, Vol. 4, 1964, Chapter 4, Section 5.
- [11] Hess, S. L. Introduction to Theoretical Meteorology, *Holt, Rinehart & Winston*, 1959, New York.
- [12] Holloway, J.L. Jr., and Manabe, S. Simulation of Climate by a Global General Circulation Model, *Monthly Weather Review*, Vol. 99, No. 5, 1971, 335-370.
- [13] Holtslag, A.A.M., and Boville, B.A. Local Versus Nonlocal Boundary-Layer Diffusion in a Global Climate Model, *Journal of Climate*, Vol. 6, 1993, 1825-1842.
- [14] Honerkamp, J. Stochastic Dynamical Systems. Concepts, Numerical Methods, Data Analysis. *New York: VCH*, 1994.
- [15] Hopf, E. Statistical hydrodynamics and functional calculus, *J. Rat. Mech. Anal.*, 1, 1952, 87-142.
- [16] Jin, F.F., and Neelin, J.D. Modes of interannual tropical ocean-atmosphere interaction - a unified view, *Journal of Atmospheric Sciences*, 50, No.21, 1993, 3477-3503.
- [17] Johnson, J.R. A new paradigm for adaptive finite element method, in *The Mathematics of Finite Elements and Applications*, Ed. by J.R. Whiteman, 1994, 105-120.
- [18] Johnson, C., Rannacher, R., and Boman, M. Numerics and hydrodynamic stability: toward error control in computational fluid dynamics, *SIAM J. Numer. Anal.*, 32, No.6, 1995, 158-1079.
- [19] Kasahara, A., and Washington, W. NCAR Global General Circulation Model of the Atmosphere, *Monthly Weather Review*, Vol. 95, No. 7, 1967, 389-402.
- [20] Kasahara, A. Various Vertical Coordinate Systems Used for Numerical Weather Prediction, *Monthly Weather Review*, Vol. 102, 1974, 509-522.
- [21] Kiehl, J.T., et al Description of the NCAR Community Climate Model (CCM3), *NCAR TN -420*, 1996.
- [22] Kolmogorov, A.N. Local structure of turbulence in an incompressible fluid at a very high Reynolds number, *Dokl. Akad. Nauk. SSSR*, 30, 1941, 299-302.
- [23] Kreiss, H.O. Problems with different time scales, in *Acta Numerica*, 1992, 101-139.
- [24] Kurihava, Y. Numerical Integration of the Primitive Equations on a Spherical Grid, *Monthly Weather Review*, Vol. 93, No. 7, 1965, 399-415.
- [25] Lax, P.D. Hyperbolic systems of conservation laws and the mathematical theory of shock waves, *SIAM Publications*, Philadelphia, 1972.
- [26] Lax, P.D. The zero dispersion limit, a deterministic analog of turbulence, *Comm. Pure Appl. Math.*, 44, 1991, 1047-1056.
- [27] Lee, T.D. On some statistical properties of hydrodynamic and hydromagnetic fields, *Quarterly Appl. Math.*, 1952, 69-72.
- [28] Lin, S.-J., et al A class of the van Leer-type transport schemes and its application to the moisture transport in a General Circulation Model, *Monthly Weather Review*, 122, 1994, 1574-1592.

- [29] Ma, C.-C., et al Sensitivity of a Coupled Ocean-Atmosphere Model to Physical Parameterizations, *Journal of Climate*, Vol. 7, 1994, 1883-1896.
- [30] Manabe, S., Smagorinsky, J., and Strickler, R. Simulated Climatology of a General Circulation Model with a Hydrological Cycle, *Monthly Weather Review*, Vol. 95, No. 7, 1967, 389-402.
- [31] Manabe, S. Climate and the Ocean Circulation, *Monthly Weather Review*, Vol. 97, No. 11, 1969, 739-774.
- [32] Marchuk, G.I. Splitting and Alternating Direction Methods, in *Handbook of Numerical Analysis*, Vol. 1, Ed. by P.G.Ciarlet and J.L.Lions, 1990, Elsevier Science Publishers, 199-462.
- [33] McCreary, J.P., Jr., and Anderson, D.L.T. An Overview of Coupled Ocean-Atmosphere Models of El-Nino and the Southern Oscillation, *Journal of Geophysical Research*, 96, No.28, 1991, 3125-3150.
- [34] Melnik, V.N. Semi-implicit finite-difference schemes with flow correction for quasihydrodynamic models, *Engineering Simulation*, 1995, 856-865.
- [35] Melnik, V.N. Nonconservation law equation in mathematical modelling: aspects of approximation, *Proc. of the International Conf. AEMC'96*, Sydney, 1996, 423-430.
- [36] Melnik, V.N. Optimal probabilistic trajectories of deterministic finite-state machines, *School of Mathematics, University of South Australia*, TR 15, 1996, 1-25.
- [37] Monin, A. S. Weather Forecasting as a Problem in Physics, *The MIT Press*, 1972.
- [38] Muller, I., and Ruggeri, T. Extended Thermodynamics, *Springer-Verlag*, 1993.
- [39] Obukhov, A.M. Spectral energy distribution in turbulent flow, *Dokl. Akad. Nauk SSSR*, 32, 1941, 22-24.
- [40] Peshkov, V.A. Second sound in helium II, *Journal of Physics*, 3, 1944, 381.
- [41] Phillips, N.A. A Coordinate System Having Some Special Advantages for Numerical Forecasting, *Journal of Meteorology*, Vol. 14, 1957, 184-185.
- [42] Prandtl, L. Bericht über Untersuchungen zur ausgebildeten Turbulenz, *Zeitschr. angew. Math. Mech.*, 5, 1925, 136-139.
- [43] Prandtl, L. Zur turbulenten Stroemung in Rohren und laengs Platten, *Ergeb. Aerodyn. Versuch.*, Series 4, 1932, Goettingen.
- [44] Prokhorov, Yu.V. Convergence of random processes and limit theorems of probability theory, *Theory Probab. and Appl.*, 1, 1956, 157-214.
- [45] Randall, D.A., et al Analysis of snow feedback in 14 general circulation models, *Journal of Geophysical Research*, 99, No. D10, 1994, 20757-20771.
- [46] Rasch, P.J., and Williamson, D.L. The Sensitivity of a General Circulation Model Climate to the Moisture Transport Formulation, *Journal of Geophysical Research*, 96, No. D7, 1991, 13123-13137.
- [47] Rasch, P.J., et al A three-dimensional transport model for the middle atmosphere, *Journal of Geophysical Research*, 99, No. D1, 1994, 999-1017.
- [48] Read, P.L. Applications of chaos to meteorology and climate, in *The Nature of Chaos*, Ed. by T. Mullin, Clarendon Press, Oxford, 1994, 222-260.
- [49] Reynolds, C.A., Webster, P.J., and Kalnay, E. Random error Growth in NMC's Global Forecasts, *Monthly Weather Review*, 122, 1994, 1281-1305.

- [50] Richardson, L.F. *Weather Prediction by numerical Process*, Cambridge Univ. Press, London, 1922.
- [51] Sazonov, V. A note on characteristic functionals, *Theory Probab. and Appl.*, 3, 1958, 188-192.
- [52] Smagorinsky, J. General Circulation Experiments with the Primitive Equations, *Monthly Weather Review*, Vol. 91, No. 3, 1963, 99-164.
- [53] Shashkov, M. Conservative Finite-Difference Methods on General Grid, *CRC Press*, 1996.
- [54] Struwe, M. *Variational Methods. Applications to Nonlinear PDEs and Hamiltonian Systems*, Springer-Verlag, 1990.
- [55] Tripoli, G.J. A nonhydrostatic mesoscale model designed to simulate scale interaction, *Monthly Weather Review*, 120, 1992, 1342-1359.
- [56] UXP/V Fortran-90/VP. User's Guide (V10), *J2U5-0050-01EN*, Fujitsu, 1995.
- [57] UXP/V Fortran-90 Messages (V10), *J2U5-0060-01EN*, Fujitsu, 1995.
- [58] Vernotte, P. La véritable équation de la chaleur, *Comptes Rendus Hebd. Séances Acad. Sci.*, 247(23), 1958, 2103-2105.
- [59] Von Karman, T., Mechanische Ähnlichkeit und Turbulenz, *Nach. Ges. Wiss. Goettingen Math.-Phys. Klasse*, 1932, 58-76.
- [60] Wakata, Y., and Sarachik, E.S. Nonlinear Effects in Coupled Atmosphere-Ocean Basin Modes, *Journal of Atmospheric Sciences*, 51, No.6, 1994, 909-920.
- [61] Williamson, D.L., and Olson, J.G. Climate simulation with a semi-Lagrangian version of the NCAR Community Climate Model, *Monthly Weather Review*, 122, 1994, 1595-1610.
- [62] Wu, G., and Lau, N.C. A GCM simulation of the relationship between tropical-storm formulation and ENSO, *Monthly Weather Review*, 120, 1992, 958-977.

MATHEMATICAL MODELS FOR CLIMATE AS A LINK BETWEEN COUPLED PHYSICAL PROCESSES AND COMPUTATIONAL DECOUPLING

Roderick V. Nicholas Melnik

Department of Mathematics and Computing
University of Southern Queensland
Toowoomba, QLD 4350, AUSTRALIA
E-mail: melnik@usq.edu.au

Abstract

Mathematical models for climate studies are treated as a coupling link between physical and computational models. These models are characterized by the fact that small-scale phenomena influence the large-scale properties of the modelling system, yet the former cannot be extracted from the latter using available hardware and computational procedures. Climate systems belong to the class of systems whose dynamics are only observable in transient states. As a result, the sensitivity of models to coupling procedures requires an examination of the schemes responsible for transporting data between components. It is proposed to perform such an examination, based on the connection between error growth and the degree of coupling of model components, using adaptive error control.

Key words: mathematical climate system models, coupling and decoupling procedures, hydrodynamic stability.

1 Introduction

Elements of the mathematical modelling of climate can be traced back to Aristotle's Meteorologica. The first contemporary achievement in this field is often attributed to U. Leverrier, who was the first to produce a weather map after a big storm in France in November 1854. A rigorous modern basis for short-range weather prediction was laid by Vilhelm and Jacob Bjerknes and other scientists of the Bergen school. For the first time they rigorously treated the problem of weather forecasting as a mechanico-mathematical problem. Such a problem was described mathematically by an initial value problem for the hydrodynamic equations of a baroclinic fluid. One of the main concepts introduced in this work was *the concept of wave instability on frontal interfaces*, which is still of primary importance in theoretical meteorology.

An important step in mathematical climate studies was made by L. Richardson and Courant, Friedrichs, Lewy in the field of numerical analysis. The results of the latter group gave a guideline for the explanation of some failures in the former work, where meteorological "noises"¹ were included into the model.

During the early 1940s through to the 60s, many scientists contributed to filtering noise from the solution for the hydrodynamic system. In the field of theoretical meteorology a number of pioneering works were published, establishing basic principles for simplification of the hydrodynamic equations through the *quasi-geostrophic expansion*. Mathematical foundations of the theory were laid using a probabilistic approach [22, 40]. The main difficulty with the early works in this direction was the formulation of boundary conditions for the problem. Alternatives to the geostrophic approximations of the hydrodynamic equations were also developed. Among them was the *quasi-solenoidal approximation*.

Serious mathematical difficulties in the practical application of quasi-geostrophic and quasi-solenoidal approximations led in the late 1950s to a return to the initial hydrodynamic equations which were essentially used by L. Richardson. Such equations in theoretical meteorology are called *primitive*. Since that time, the main developments have been concentrated on numerical methods and the improvement of models by a better physical parameterization. Many important factors related to the physical parameterization were taken into consideration and implemented into models. This led to the creation of modern state-of-the-art models for climate that

- consist of relatively independent components that are responsible for interconnected parts of climate such as atmosphere, ocean, land surface, sea ice, etc, and
- require substantial computational power to obtain approximate solutions.

The quality of these approximate solutions depend significantly on the consistency between the mathematical model and the real climate. Since the improvement of

¹such as acoustic waves

mathematical models can be achieved by improved physical parameterization, *the concept of coupling between different components* becomes straightforward.

From the physical point of view, climate studies are essentially based on three fundamental theories. These are thermodynamics, the theory of radiation, and magnetohydrodynamics. Since a description of the climate system should include both the earth and the atmosphere, the overall system is often referred to as the atmosphere-active-layers (AAL) system. Both the earth and the atmosphere require a detailed physical parameterization, that leads to the difficulty of mathematically formalising *the interaction between underlying processes and phenomena*. The main difficulty with the description of the earth is to measure a purely gravitational force. In fact, we can only observe the combined effect of the two forces, gravitational and centrifugal, that is referred to as gravity. The measurement difficulties measurements stem from

- the variation of gravity at different latitudes;
- the variation of gravity vertically with respect to sea level;
- the variation of gravity erratically with respect to the earth's crust and other irregularities.

The main difficulty in the description of the atmosphere is to adequately represent transport effects in models. Amongst the most important constituents in the atmosphere are water (about 4% per volume), carbon dioxide (about 0.03% per volume), ozone (about $0.1 \times 10^{-5}\%$ per volume), oxygen (about 20%) and nitrogen (about 70%). However, if water and carbon dioxide are present throughout the atmosphere, then ozone becomes influential only at 20-30 km from the earth's surface, oxygen only from about 80 km and nitrogen even higher [38]. In addition, many gaseous constituents only have an indirect influence through the propagation of electromagnetic waves. There are also many important nongaseous constituents such as condensed H_2O , salt particles, dust and others that play important role in the description of clouds and precipitation.

We conclude that transport phenomena and gravity are key factors in an adequate description of the AAL system.

2 The structure of the paper and notation

This paper is organised as follows:

- Section 3 gives a brief outline of space-time scales that are important in a climate study. We recall the main hypothesis that is used in the mathematical modelling of synoptic processes.
- In Section 4 we consider the distinction between short and long range predictions on the basis of the concept of the relaxation time. The fundamental equations for the adiabatic approximation are also presented in this section.

- Section 5 is devoted to non-adiabatic models and their simplifications on the basis of quasi-geostrophic and quasi-solenoidal approximations.
- In Section 6 we consider advantages in the return to the primitive hydrodynamic equations, and difficulties in an adequate representation of the vertical structure of meteorological fields.
- Section 7 deals with non-adiabatic factors that lead to an approximation of the conservation law in mathematical models.
- Sections 8–10 are devoted to a dilemma between the concepts of coupling and independence as well as approaches for the numerical treatment of an interplay between these two concepts.
- In Sections 11 and 12 we formulate a finite set of differential equations that provides an approximation to the dynamics of climatic processes. We address two questions related to such an approximation, namely phase transitions and algorithmic stability.
- Section 13 deals with questions related to the validation of mathematical models for climate. We argue that for the validation of models, both a-priori assumptions and a-posteriori information are needed.
- In Section 14 we present some numerical results of the computation of meteorological fields on the basis of the NCAR CCM3 model.
- Section 15 concludes the paper. Directions for future development are also presented in this section.

The following notation is used throughout the paper:

- g is the acceleration due to gravity;
- ρ is the density;
- p is the pressure;
- T is the temperature;
- z is the altitude;
- m is the mass of the earth;
- c_0 is the isothermal sound of speed;
- h_a is the height of the atmosphere;
- p_0 is the average surface pressure;
- ρ_0 is the average surface density;
- T_0 is the average air temperature at sea level;
- k_b is the Boltzmann constant;
- k_v is the von Karman constant;
- $k_r = c_p/c_v$ is the ratio of specific heat capacity under constant pressure and constant volume;
- R is the gas constant, for example, $R = 287 \text{ J kg}^{-1} \text{ K}^{-1}$ for dry air and $R = 461 \text{ J kg}^{-1} \text{ K}^{-1}$ for water vapor;
- $\Omega = 7.292 \times 10^{-5} \text{ s}^{-1}$ is the angular velocity of the earth;
- φ is the latitude;
- $\chi = 2\Omega \sin \varphi$ is the Coriolis parameter;

- $\mathbf{F} = (F_x, F_y, F_z)$ is the field external to the earth, excluding pressure-gradient forces;
 - c_p is the specific heat capacity at constant pressure, for example, $c_p = 1. \times 10^3 \text{ J kg}^{-1} \text{ K}^{-1}$ for dry air and $c_p = 1.81 \times 10^3 \text{ J kg}^{-1} \text{ K}^{-1}$ for water vapor.
- Other notation is explained in the text as required.

3 Space-time scales and their interaction

In climate studies one of the most challenging problems is to adequately describe space-time scale interactions [23]. By interacting between themselves, different bio-chemico-physical processes at different scales form a unified whole which we call climate. The problem of such interactions is typically simplified mathematically by regarding microturbulence as a dissipative factor which can be characterized by an *effective (or dynamic) viscosity coefficient*. Such a simplification allows us to effectively model *synoptic oscillations*, i.e. climate processes that are characterized by time scales from hours to several days. Diurnal oscillations also belong to this class. Amongst other types of oscillations, the following classes can be distinguished:

- Global oscillations, for example, planetary oscillations. They play an essential role in long-term weather predictions. Their time scales are characterized by the period from weeks to months. The Atmospheric Boundary Layer is a key factor in such processes.
- Seasonal oscillations that vary over a year.
- Interannual oscillations with time scales of several years. To this class belong, for example, glacial periods and ENSO-type phenomena.
- Micrometeorological oscillations with time scales of seconds to minutes. Small-scale turbulence, acoustic waves, and gravitational waves with small amplitudes provide examples of this type of oscillations.
- Mesometeorological oscillations such as thunderstorms, and gravitational waves with large amplitudes. They typically last from minutes to an hour.

We emphasize that mathematical models of climate systems are essentially “proxy” climate systems. Whatever model is chosen, small scale phenomena may substantially influence large-scale properties of the system, but computational procedures may not be available to extract the former from the latter.

4 Short and long range in the prediction of meteorological fields

The short-range prediction of meteorological fields is based on hydrodynamic theory in the case where the energy of sources and sinks is virtually ignored by using *the*

adiabatic approximation. In quite a general setting, the resulting equations can be derived from the two conservation law equations, the conservation of the entropy and the conservation of the “vortex charge” [38],

$$\frac{d\Xi}{dt} = 0, \quad \frac{d\Psi}{dt} = 0, \quad (4.1)$$

where $\Psi = (\Psi_0 \cdot \nabla \Xi)/\rho$ is the potential Rossby vorticity. Due to (4.1), the entropy function Ξ and the absolute vorticity Ψ_0 are generators of differential adiabatic invariants, because any function of them is again an adiabatic invariant. As an integral invariant, all systems that can be described by the model (4.1) have the total energy of the system, \mathcal{E} , constant.

In the general case, the well-posedness of the model (4.1) is not an established mathematical fact. Any specific choice of two independent Lagrangian coordinates as well as the definition of two parts of the integral invariant², implies the necessity of addressing the problem of system stability. E.N. Lorenz was the first who proposed addressing this problem using the *macrostability parameter*, \mathcal{S} . Let

$$\Theta = T(p_0/p)^{(k_r-1)/k_r} \quad (4.2)$$

be the potential temperature, where p_0 is the standard pressure. This quantity is often convenient as one of the Lagrangian coordinates. Then, the parameter of macrostability, \mathcal{S} , can be defined as the weighted average value of the vertical gradient of the potential temperature over the entire thickness of the atmosphere. If \mathcal{K} is the kinetic energy of the system, then the quantity $\mathcal{K} - \mathcal{S}$ is also adiabatic invariant. Hence, the quantity \mathcal{S} shows the amount of kinetic energy that is released/absorbed in the process of adiabatic transitions. This approach to stability requires an adequate specification of the vertical structure of the atmosphere.

An alternative approach is based on the concept of relaxation time, τ . From the mechanical point of view, the relaxation time, or “build-in” period, can be seen as the atmospheric efficiency coefficient, i.e. the rate at which potential energy, \mathcal{E}_p , is converted into kinetic energy:

$$\tau = \left(\frac{1}{\mathcal{E}_p} \frac{\partial \mathcal{E}_p}{\partial t} \right)^{-1}. \quad (4.3)$$

On the scale of synoptic processes, $\tau \approx 1$ week. If the time interval of interest $t - t_0$ (where t_0 is an initial moment of time) is less than τ the model (4.1) may provide a good approximation for short-range weather changes. However, for periods that satisfy the inequality

$$t - t_0 > \tau, \quad (4.4)$$

practically all regions of the atmosphere have sufficient time to interact with each other and the model (4.1) becomes inappropriate. Since the atmosphere is a rapidly

²for example, the kinetic energy and the labile energy that, in turn, consists of the sum of the potential energy and the internal energy of the system

changing component with low inertia of the whole AAL system, an essential part of investigation in the field of climate study is being concentrated on atmospheric modelling. This approach gives rise to the major difficulty in the modelling of long-term meteorological changes. We have *to fix the initial state of the whole physical system*, namely when $t = t_0$. Of course, this requires more careful examination of other components of the system, in particular the ocean, which is a component with a large thermal inertia. Hence, one of the most important initial conditions for the whole model is the temperature field. It is well-known, for example, that incompleteness of such data causes problems in modelling processes such as ENSO phenomenon, and other processes in the equatorial zones where the Coriolis parameter vanishes and the structure of the boundary layer of the atmosphere has to be modelled with an increased precision [16, 63, 6, 33, 61]. The implementation of *non-local* features of the system into the model becomes important for the validity of the model [13]. In the end, this requires an appropriate description of *turbulence in the boundary layer* as a major factor responsible for the *mixing* of heat, momentum, passive scalars, moisture etc. This emphasises the importance of taking into account both the interaction of time-scales and the interaction of spatial scales [49].

As we mentioned in the introduction, three main types of physical processes, namely

- thermodynamical,
- radiative, and
- magnetohydrodynamical,

influence the output of mathematical models of climate subjected to the physical parameterization [29, 46]. As a result, many efforts during recent times have been concentrated on improvements of existing physical parameterizations. In particular, much attention has been devoted to an adequate modelling of radiative processes [4] that require appropriate models for cloudiness and the transport of tracer species [48]. The analysis of sensitivity to transport phenomena has led to an increased interest in the semi-Lagrangian approach as an alternative to the well-established spectral approaches [47]. The semi-Lagrangian approach requires special numerical procedures for interpolation to compensate for the sparse character of data.

The necessity of an adequate representation of transport phenomena in mathematical models naturally leads to the development of the concept of coupling with respect to different components of climate [5]. In addition to *the large spatial scale* that cover thousands of kilometers other scales, such as

- mesoscale (from kilometers to hundreds of kilometers),
- small scale (from dozens of meters to kilometers), and
- microscales (from millimeters to dozens of meters).

become important for long-range prediction. Of course, in climate applications smaller scale structures are inevitably described statistically. It does not follow, however, that such structures are necessarily random in nature. The only a-priori conclusion we can draw is that in long-term processes *the atmosphere as a whole* does not act as a *closed system* [38]. It acts as a component of a bigger AAL system composed of the atmosphere and active layers that can be described by a coupling of many different physical, chemical and biological fields.

5 Physical hypotheses and mathematical approximations

In the long-term prediction, equations for the conservation of entropy and the potential vorticity cannot provide an adequate description of underlying processes that are essentially nonadiabatic. Since in this case we should be able to adequately describe sources and dissipation of energy, certain assumptions about the laws of dissipation and accumulation of energy should be made. In this case, model (4.1) should be replaced by evolutionary equations that more adequately represent transport phenomena. Let us denote by Λ the rate of energy increase per unit mass, and by \mathbf{F}_v the viscous force per unit mass. Then such equations can be written in the form

$$T \frac{d\Xi}{dt} = \Lambda, \quad \rho \frac{d\Psi}{dt} = \operatorname{div} [(\Lambda/T)\Psi_0 + \Xi(\operatorname{curl}\mathbf{F}_v)]. \quad (5.1)$$

This model allows us to take into account *non-adiabatic effects*. It introduces the two new variables, the momentum and heat fluxes, and requires additional hypotheses on sources and sinks in the AAL system. The central hypothesis is the hypothesis of *local thermodynamic equilibrium*. During recent years theoretical and experimental physicists proved it was necessary to go beyond the framework of this hypothesis [39, 35]. We note, that model (5.1) is a model of local type. In fact, to derive this model, the Obukhov hypothesis on the conservative properties of the potential vorticity is used [38], and the field

$$(\Lambda/T)\Psi_0 + \Xi(\operatorname{curl}\mathbf{F}_v) \quad (5.2)$$

is assumed to have a solenoidal structure. Nevertheless, in models of this type, difficulties connected with the inherently approximate modelling of physical processes such as the polarization of radiation, refraction, dispersion, and cloudiness, may still be partially overcome by introducing some feedback mechanisms, like regulators through the cloud cover, sea ice, snow cover etc.

The model (5.1) with a realistic physical parameterization is extremely difficult to deal with without some additional simplifications. Once again, we can use the local equilibrium hypothesis for mathematical simplifications of climate system models. This hypothesis leads to satisfactory results, at least in the case of small-amplitude waves. In this particular case it is sufficient to relate small oscillations

of the atmosphere to *the equilibrium state*. When oscillations are small *the perturbation theory technique* is natural for the investigation of solutions of resulting models. In its essence, the successful application of this theory in many areas has its roots in the local equilibrium hypothesis.

Let us assume that in the equilibrium state pressure, density and temperature, p , ρ and T , depend only on the amplitude z . Then, as a primary task, we have to describe the *dynamics of these variables near the equilibrium*. In climate study, these variables are connected, as a rule, by the “timeless” *hydrostatic* (or *quasi-static*) equation,

$$\frac{\partial p}{\partial z} = -\rho g, \quad (5.3)$$

and the Clapeyron equation

$$p = \rho RT, \quad (5.4)$$

where, as usual, g is the acceleration due to gravity and R is the specific gas constant.

For now, we shall assume that at the initial moment of time t_0 the atmospheric motion is

- quasi-static, i.e.

$$\partial p_e / \partial z = -g \rho_e; \quad (5.5)$$

- horizontal, i.e. for the velocity field $\mathbf{v} = (u, v, w)$ we assume

$$w = 0; \text{ and} \quad (5.6)$$

- geostrophic i.e. the velocity field is assumed to be non-divergent,

$$u_e = -\frac{1}{\chi\rho} \frac{\partial p_e}{\partial y}, \quad v_e = \frac{1}{\chi\rho} \frac{\partial p_e}{\partial x}, \quad (5.7)$$

where χ is the Coriolis parameter and the subscript index e stands for the values of thermodynamic parameters at moment t_0 . Of course, in this case one can introduce the stream function by the formula

$$\psi = p_e / (\chi\rho), \quad (5.8)$$

and the formulas (5.7) may be rewritten as

$$u_e = -\frac{\partial \psi}{\partial y}, \quad v_e = \frac{\partial \psi}{\partial x}. \quad (5.9)$$

One of the assumptions that is often made is that these three properties (referred to as *consistency conditions*) will be preserved in the future. This guarantees a stationary solution for the set of equations governing the atmosphere. Such motion is called motion of the first kind or *slow motion*. By the standard technique we can

also account for the curvature of the earth by the transformation of the stationary solutions into slow gyroscopic Rossby waves.

If consistency conditions are violated in any region of space, X , *fast motion*, or motion of the second kind, has to be taken into consideration. In reality, we observe a continuous competition between a violation of the consistency conditions and adaptation of the meteorological fields \mathbf{v} , p , ρ , T . Since traditional methods assume that “meteorological noise” has little significance in the weather prediction, numerous attempts have been made to filter out motions of the second kind. However, if we accept the quasi-static approximation, all frequencies of the internal acoustic waves go to infinity (“complete filtering”). Frequencies of the gravity waves become overestimated, though the error decreases for longer waves.

The weather is typically associated with synoptic processes. Hence, in order to describe such processes at a “minimal cost”, we have to filter out from solutions of the Eulerian hydrodynamic equations (they contain both slow and fast motions) the motions of the second kind. If it is assumed that only the motion of the first kind is important for synoptic processes, then in the Eulerian equations we have two dimensional parameters, g and χ , that can be used for such filtering. Nonlinear systems in which fast oscillations occur along with slow ones are typical in applications³. Asymptotic methods and the theory of perturbations were developed in order to mathematically treat such systems. The idea of expansion with respect to a small parameter can also be used as a simplification of the hydrodynamic system to make it appropriate for the description of synoptic processes.

Let L and U be typical length and speed scales for synoptic processes. Then the role of small parameters may be played by the Rossby-Kibel number

$$R_k = U/(\chi L), \quad (5.10)$$

or the Mach number

$$M_a = U/c_0. \quad (5.11)$$

The isothermal sound speed, c_0 , is defined using the height of the atmosphere, h_a , as

$$c_0 = \sqrt{gh_a}, \quad \text{where } h_a = p_0/(\rho_0 g). \quad (5.12)$$

The Rossby-Kibel number may be interpreted as the ratio of the typical relative acceleration, U^2/L , to the typical Coriolis acceleration χU [38]. Using standard series expansion, the consistency conditions for the first kind motion may be defined from conditions for the vertical potential vorticity,

$$\Psi_z = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} = \frac{g}{f} \nabla^2 z + O\left(\frac{UR_k}{L}\right), \quad (5.13)$$

³Probably, the most widely cited example is the Van-der-Pol equation for the description of oscillations in an electric circuit containing a vacuum tube with feedback

and for the horizontal divergence,

$$D_h = \frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} = O\left(\frac{UR_k}{L}\right). \quad (5.14)$$

This expansion made with respect to the parameter UR_k/L is referred to as the *quasi-geostrophic expansion*. The quantities p_0 and ρ_0 in (5.12) are defined at the surface of the earth, and have to be included in the boundary conditions of the problem. We also need the initial field of the pressure to compute the pressure at future moments of time. However, the advantage of the quasi-geostrophic approximation is that formally we are not required to know the initial distribution of the velocity field. It should be noted that, strictly speaking, the resulting system can only be represented in the form of differential equations for synoptic processes in a barotropic-atmosphere approximation. A weak formulation of the problem is required for the more general case.

The quasi-geostrophic approximation is one of the possible approximations to conservation laws for the horizontal motion. In spite of its partial advantages it has serious drawbacks. The major drawback is a violation of the assumption of smallness for the Rossby-Kibel number, R_k , in the vicinity of the equator, since the Coriolis parameter $\chi = 2\Omega \sin \varphi$ decreases near the equator. Moreover, there is evidence that this approximation may be inadequate even outside of the tropical zones. In such cases the consistency conditions (5.13), (5.14) for the synoptic fields must be changed. For example, in many cases it is reasonable to assume that the horizontal divergence D_h is small compared to the vertical vorticity Ψ_z , where

$$\Psi_z = O(U/L), \quad (5.15)$$

$$D_h = O\left(\frac{U}{L} \frac{M_a^2}{\alpha_0^2}\right). \quad (5.16)$$

In (5.16), α_0 is the parameter of (hydro)static stability. It is defined as

$$\alpha_0^2 = -(T/T_0)(p/c_p)\partial\Xi/\partial p, \quad (5.17)$$

and has an obvious connection with the Richardson number [38]

$$R_i = \frac{T_0^2}{M_a^2 T^2} \alpha_0^2, \quad (5.18)$$

(see also the definition in [11]). These modified consistency conditions, (5.15), (5.16), are derived with the assumption that in the horizontal velocity field for the slow (synoptic) motion, the potential component is small compared to the solenoidal component. Such an assumption leads to the *quasi-solenoidal approximation*. The mathematical model that corresponds to this approximation is typically formulated in terms of the stream function ψ , where

$$u = -\frac{\partial\psi}{\partial y}, \quad v = \frac{\partial\psi}{\partial x}. \quad (5.19)$$

The model requires both

- an initial field ψ for a given z , and
- values of ψ on the boundary.

It is well-known that the balance equation written for ψ is *the Monge-Ampere-type equation* that should be solved in a given bounded region. The well-posedness of such a problem in the general setting can be guaranteed when it is elliptical. However, the ellipticity condition [38],

$$g\nabla^2 z + \chi^2/2 > 0, \quad (5.20)$$

can only be satisfied when the quantity R_k is small. As we mentioned above, this condition is often violated in applications. Such a violation is most noticeable in the tropical zones.

In addition to the mathematical difficulties in using the quasi-solenoidal approximation, and unsatisfactory practical results obtained in many cases with the quasi-geostrophic approximation, there remains the problem of filtering fast waves from the hydrodynamic equations open. However, progress in computational software and hardware has led to the possibility of solving the complete set of equations governing the atmosphere.

6 Primitive equations and the vertical structure of meteorological fields

Since in many cases neither quasi-geostrophic nor quasi-solenoidal approximations are appropriate in applications, computational complexity is caused by filtering procedures. The development of numerical procedures from the original hydrodynamic system in these cases is no more complex. The main problem in modelling with the primitive equations is the formulation of appropriate boundary conditions. If boundary conditions are not properly formulated, the stability of the solution cannot be guaranteed. In fact, we need

- the normal velocity on the entire boundary, and
- the potential vorticity on the part of the boundary where air motion is directed toward the interior.

Because of the approximate character of available data for these boundary conditions, the formulation of the mathematical model requires two types of equations, prognostic equations and tendency equations. With the quasi-static approximation, the derivation of the tendency equation is straightforward, provided the vertical structure of meteorological processes is known. In the particular case of barotropic atmosphere, the problem of the vertical structure of synoptic processes is reasonably simple. This explains the early development of the theory in the direction of simplifications described in Section 5.

However, in the general case of a baroclinic atmosphere, the problem of the vertical structure of meteorological fields remains one of the most serious problems for mathematical modelling using the primitive equations. At the initial stage of development of the model, mainly pressure was used as the vertical coordinate. The implementation of the earth's orography [42] led to the adoption of a σ -system coordinate for the model. Among the first to use the geometrical altitude as a vertical coordinate was L. Richardson [51]. This idea was developed further by V. Starr (see references in [20]) who introduced quasi-Lagrangian coordinate systems. The present development of the vertical structure of meteorological fields in the NCAR CCM3 model [1] is based on the works of Kasahara and Washington [19], Kasahara [20], and Simmons and Stufing (see references in [21]). Among other approaches to vertical coordinates we include initial attempts at using the potential temperature (see (4.2)). This very fruitful idea has not received a proper development in the literature due to difficulties connected with lower boundary conditions.

From the mathematical point of view, the primitive equations are characterised by the “restoration” of the hyperbolic operator in the model. From the physical point of view we retain gravitational waves among the solutions. As a result, on the one hand this approach requires a large number of initial data. On the other hand, the computational complexity of resulting approximate solutions depends on values of small parameters such as the Rossby-Kibel number, R_k . We recall that this number characterizes the ratio between the inertial force of the system and the Coriolis force. Hence, if R_k is large⁴, the Coriolis effect related to small-scale effects may be neglected, as is usually done in our everyday life, in spite of the rotation of the earth. However, if R_k is small, the complexity of approximating algorithms increases.

For many meteorological and oceanographic phenomena it is important to take into account the dynamic nature of such “small” parameters induced by the interaction of different space-time scales. At present, such an interaction is modelled on the basis of the classical law of viscosity that gives a connection between the stress σ and the speed $|\mathbf{v}|$ through the effective viscosity coefficient μ ,

$$\sigma = \mu \frac{\partial |\mathbf{v}|}{\partial z}. \quad (6.1)$$

For practical applications, (6.1) should be supplemented by the law of energy dissipation. When microturbulence is treated as a dissipative factor, we can always find a reasonable analogy between the motion of molecules and the motion of macroscopic elements of turbulent fluids. This idea was first used by Prandtl [43], who attempted to treat the case of turbulent momentum exchange in this way. He arrived at the mixing-length hypothesis (see, for example, [12])

$$K_v = l^2 \left| \frac{\partial \mathbf{v}}{\partial z} \right|, \quad (6.2)$$

⁴for example, when Ω is small or when L is relatively small

that gives a connection between the coefficient of vertical diffusion K_v and the velocity field \mathbf{v} through the mixing length l . The connection between (6.1) and (6.2) should be provided by a scaling law. A major portion of current investigations in climate study is based on the scaling law of logarithmic type

$$\bar{u} = \frac{u^*}{k_v} \ln \frac{z}{z_0}, \quad (6.3)$$

where \bar{u} denotes the mean velocity (in the x-direction), u^* is the friction velocity, z_0 is the roughness parameter, and k_v is the *von Karman* constant. Recently, new theoretical and experimental evidence was given to confirm that this law may be inappropriate as an adequate description of turbulent processes [2, 3].

7 Long-term meteorological processes and non-adiabatic factors

The main difference between short and long range meteorological processes is that in the long-range the atmosphere cannot be regarded as a *closed system*, since it is a part of a bigger AAL system. Among the most important active layers is the ocean. After approximately 1-2 weeks the upper layer of the ocean has a substantial influence on atmospheric processes. As a result, one of the most important *initial conditions* in the model is the temperature field, in particular the temperature of the active layer of the ocean. At present, the practical availability of large datasets for such conditions has led to different ideas aimed at lengthening the period of validity of short-range models for the atmosphere. Initially, we can assume a constant temperature for the ocean, then use a slab-ocean model. Further, we can increase the number of formally *independent mathematical models* that can interact between themselves through message passing in a computational algorithm.

The main difficulty in the construction of long range models stems from the inadequateness of the adiabatic approximations. Conservation laws become approximate in nature, and one should take into consideration sources and dissipations of energy. In Section 5 we defined an approximation by model (5.1) that is based on the Obukhov hypothesis. Ultimately, the validity of approximations of this type is based on appropriate scaling laws. From the physical point of view, the adequate construction of the model essentially depends on taking into account dissipative effects and sources, including

- heat sources such as solar and terrestrial radiation;
- cloud dispersion and absorbtion;
- local/nonlocal boundary layer diffusion etc.

The solution of mathematical models subjected to physical parameterizations that take into consideration such dissipative effects and sources can only be approached

numerically. Moreover, the quality of the algorithm will decisively depend on the adequateness of the parameterization of bio-chemico-physical processes in the mathematical model. This is why, without additional simplifying assumptions with respect to non-atmospheric components of climate like ocean or sea ice, the “exactness” of conservation laws cannot be justified for any mathematical model. In practice, we always have to overcome difficulties arising from the approximate character of conservation laws in “proxy” climate models. Nevertheless, since the model is solved numerically, we can always use the idea of conservation on a finite grid [54, 5]. Indeed, the representation of conservation laws in mathematical models is of an approximate nature. For example, by a requirement on the vertical finite differences of the model to conserve the global integral of total energy in the absence of sources and sinks [5], we still neglect lack of conservation. In general, the stability conditions of the “proxy system” are not only different from those for the stability of the system itself, but they also are sensitively dependent on the degree of coupling achieved in the “proxy system”. The approximate character of initial and boundary data in models with a hyperbolic-type operator does not allow consideration of the atmosphere in long-term processes as a *closed system*. Mathematical difficulties for such an approach are obvious. If the atmosphere is a dependent component of the AAL system, then the questions “*how many such components are sufficient to adequately describe climatic processes*” and “*what are these other components*” have to be answered.

A wider mathematical “freedom” is allowed by looking at the evolution of states of the atmosphere as a random process $\omega(t)$. In this case it is possible to approach the task of studying the *possibilities* of the statistical extrapolation of this process using Kolmogorov’s hypothesis. Namely, a random process $\omega(t)$ describing the evolution of the turbulent flow in an environment with vanishing viscosity asymptotically approaches a Markov process for large t . From such a consideration it follows that the distribution of probabilities $P^t(d\omega)$ for $t > t_0$ may, in principle, be uniquely determined by the state $\omega(t_0)$, and not be dependent on the remote history of the process when $t < t_0$. Although the assumption of the *negligible viscosity approximation* can be justified on a finite grid⁵, validation of the original mathematical model is intrinsically connected with the processing of incomplete information, which requires an adequate formulation of scaling laws. It appears that, in the general case, the Kolmogorov-Obukhov scaling law for local structures [22, 40], that is typically used, becomes inappropriate for this purpose. We discuss these issues in the next sections.

⁵using, for example, the four-thirds Richardson’s law

8 Information exchange between components of mathematical models for climate

There are many “proxy climate” models which allow the simulation of interactions between different components by message passage in corresponding computational models. One of the models of this type is the NCAR CSM, where the original problem of climate study is reduced to that of a *controled exchange of information between the model components* under the assumption that conservative properties⁶ can be preserved when exchanged between model components [5]. In general, this assumption can be justified numerically, and it is reasonable to hope that by improving the physical (chemical, biological) parameterization we can improve the correctness and reliability of “proxy climate” models. Clearly, such improvements, as well as improvements in hardware and software, may continue indefinitely. Hence, it is necessary to develop a strategy which permits an analysis of the trade-off between a level of coupling implied by achieved parameterization and a possible error induced by the incompleteness of available information.

In climate study, an “exact” realization of conservation laws is closely connected with the consistency conditions for “slow” motion that implies hydrostatic, (5.5), horizontal, (5.6), and geostrophic, (5.7), approximations. In reality these conditions are continuously violated. As a result, approaches to *filter out the fast waves from the solution* of coupled system of PDEs are quite restricted in their applicability to climate study. Such approaches are typically based on the Kolmogorov-Obukhov scaling laws for local structures [22, 40]⁷, and more generally on the local equilibrium hypothesis [39]. It is often the case that statistical field theory can be used to give a practical explanation of these hypotheses. However, for many complex dynamic system such as climate, mutual re-adjustment and self-adaptation of fields of different space-time scales is in the nature of the underlying processes. In such cases, the classical Kolmogorov-Obukhov scaling law may not lead to an adequate approximation of these processes [2, 3]. As we mentioned, the key point behind this fact is that the approximate character of initial data in models with a hyperbolic-type operator (such as the primitive equations of the hydrodynamic theory) does not allow consideration of the atmosphere in long-term processes as a *closed system*. Moreover, since the stability of any closed “proxy system” does not imply stability of the system itself, we need a trade-off strategy between coupling and stability concerns. To define such a strategy we have to appeal to the idea that a division between long-range and short-range depends on the definition of the *relaxation time* (or “build-up” period), τ (see (4.3)). This “coefficient of atmospheric efficiency” depends on the degree of coupling of atmosphere to its active layers that is ultimately defined by the problem solver or modeler. Mathematically, a formal division between “long” and “short” ranges is defined by the sign of the inequality between

⁶e.g. momentum, heat, freshwater

⁷we give more details in Section 10

$t - t_0$ and τ :

$$r = \text{sign}\{t - t_0, \tau\}. \quad (8.1)$$

The possibility of the existence of the two simultaneous limits

$$\tau \rightarrow 0^+, \text{ and } t - t_0 \rightarrow \infty \quad (8.2)$$

is often taken for granted as an a-priori mathematical assumption in investigations of complex dynamic system. Mathematical models based on this assumption are characterised by a strong singularity at $\tau = 0$ and the parabolic features of the underlying dynamics. From the physical point of view such models are close associates of Fourier's original ideas on a diffusion mechanism for heat conduction. It is well-known, however, that the Fourier prediction may underestimate the peak temperature during a rapid transient period. Since experimental work on the wave behavior of heat transport [41] and theoretical work in this field [59], interest in hyperbolic-type models for processes that include diffusion is being dramatically increased. In the general non-linear case, such models preclude the assumption that small-scale phenomena can be extracted from a large-scale flow. As a result, one should overcome the problem of the approximate character of conservation laws in mathematical models of the "proxy systems". For example, heat in the "proxy climate system" might be conserved only under the assumption that it is neither gained nor lost at the top of the atmosphere. In fact, *it is not conserved* under inadequate parameterization⁸. Furthermore, in general the model can conserve energy only if we neglect the lack of conservation due to *a-priori* regularity assumptions for our approximations. From the physical point of view, relaxation of these assumptions requires more careful examination of the non-local features of boundary layers [13]. We also note that in CCM-type models⁹ the vertical advection of temperature is not used to conserve mass/energy. However, it is well known that the interaction between the vertical semi-Lagrangian approximations and the convective parameterization may seriously affect system predictability. In the general case, *the a-priori assumption on the existence of conservation laws for "proxy" climate systems leads to a-priori regularity assumptions for "exact" solutions of mathematical models.*

Since the construction of mathematical models for the evolution of thermodynamic systems has to be undertaken under analysis of uncertainty and the processing of incomplete information, it is always important to investigate the stability of associated computational models. However, in some cases it is possible to approach the issues of the well-posedness of the mathematical model without the investigation of stability in a computational sense. Indeed, mathematical analysis of the model itself can often be reduced to a general stochastic control problem. In turn, this problem can be often associated with a PDE of the Hamilton-Jacobi-Bellman type [9]. The quality of mathematical models based on such PDEs is essentially determined by

⁸for example, if the long wave radiation in the atmosphere component uses the average sea surface temperature

⁹we use a version of CCM in our computational experiment

the quality of the *approximation of the system Hamiltonian*, and *approximate initial data* for the models. In this case, the adequateness of the model to the real-world situation will be completely defined by the smoothness assumption on the sought-for solution. In the case where it is assumed that initial data for the model can be given exactly, the semi-continuity assumption [55] is natural. However, with application to real dynamic systems, such models can be reasonably validated if we are able to analyse the distinction between

- external error growth due to model deficiencies (such as physical parameterization), and
- the internal error growth due to mathematical assumptions (resulting from the unstable “self-growth” of the initial data errors).

The level of parameterization defines an *upper bound* for such an internal error. This provides a way to investigate the connection between coupling in mathematical models and the level of uncertainty in the model prediction (see [50, 8] and references therein). On the other hand, such a bound introduces hyperbolic features into the model [39].

9 Hybrid Eulerian-Lagrangian models and numerical schemes for transport effects

Together with a wide use of the hybrid vertical coordinate, the interest in the hybrid Eulerian-Lagrangian type of mathematical models for climate study increases. In turn, this leads to attempts to implement into modern climate system models *semi-Lagrangian advection approximations* instead of the standard Eulerian approximations. One of the advantages of the semi-Lagrangian version is that in many cases it allows us to relax the normal advective Courant-Fridrichs-Lewy (CFL) stability condition. In fact, it is well-known that for standard spectral models the typical resolution of the model may lead to instability¹⁰ if one applies a standard time-step. To obviate this problem, *limited filtering* is often used on the top model layer¹¹. The interest in the semi-Lagrangian type of models is stimulated by the claim (see [62] and references therein) that application of the semi-Lagrangian version may not only exclude the above-mentioned filtering, but also eliminate the normal advective CFL time-step restriction. However, it has to be recalled that the main problem with a semi-Lagrangian formulation consists of the fact that the result of interpolation with pointwise values *is not a-priori conservative*. As a result, *long-term simulation* can be seriously affected. We should admit that both semi-Lagrangian and

¹⁰which may be observed in the Southern Hemisphere

¹¹such models are often referred to as Eulerian with Spectral Transform (EST), though one realises that they are not Eulerian because the water vapor transport is usually treated in a semi-Lagrangian manner

EST versions have a serious deficiency. Small-scale features in the solutions may be underestimated more in semi-Lagrangian versions, whereas the EST approach has to deal with Gibbs phenomenon and spectral truncation. Hence, in practice, the EST methods can be successfully applied to adiabatic approximations, whereas the semi-Lagrangian approach is more natural for problems related to the advection of fields with *large horizontal gradients*¹². We should also take into consideration the fact that advantages in the stability of semi-Lagrangian advection schemes may be lost when we use hybrid Eulerian-Lagrangian models.

The author believes that the most efficient schemes must not require an explicit “subgrid-scale turbulence” parameterization and spacial filtering. Instead of the classical semi-Lagrangian or EST approaches it is reasonable to use 1D-Flux-Corrected-Transport (FCT) schemes. For climate study, this idea was used in [28]. The same idea was used in a different area of application to avoid explicit turbulence parameterization in [34]. It is straightforward to apply this idea to 3D transport modelling by means of time splitting in a manner explained, for example, in [32]. For climate models, zonal transport at high latitude can be split in time to satisfy the local CFL restriction, whereas at low latitude we can use a larger step. Such schemes provide an increasing accuracy with increasing resolution even when *discontinuities or steep gradients* are encountered. This property is extremely important to overcome the singular nature of the spherical coordinate system near poles, especially for systems with incomplete information, including those for which additional data from observations may be added in stages.

10 Turbulence, vanishing viscosity, and scaling laws

The core of hydrodynamic theory is the system of the Navier-Stokes equations. The *ensembles of solutions* of these equations is usually associated with the behaviour of turbulence. Many theoretical works in this field are concentrated on the limiting case of vanishing viscosity in the Navier-Stokes system (the inviscid limit).

From the physical point of view, the system itself is a mathematical expression of conservation laws. Such laws are obtainable, at least in principle, from a hyperbolic system by adding a small viscosity coefficient. As follows from [25] (see also [26]), in the limit of vanishing viscosity one expects to be able to recover entropy solutions of *the original hyperbolic system*. The whole procedure is based on the assumption that a formal mathematical transformation from the hyperbolic system to the Navier-Stokes system preserves conservative properties. Naturally, if the original system is conservative, then after the vanishing-viscosity-limit transformation the system remains conservative. However, if we add to the original hyperbolic system a *dispersive term*, then we cannot expect that solutions of this modified system in the limit of vanishing dispersion are well-behaved. Existence or non-existence of solutions of such a modified system depend on *the regularity of solutions of the*

¹²for example, when modelling water vapor transport

original hyperbolic system. Therefore, though the Navier-Stokes system is parabolic in nature [39], in the general case it has both *dispersive and hyperbolic features*.

These features complicate the quantification of the behaviour of ensembles of the solutions of this system. Moreover, in the general case, the solutions of this system are non-stationary, and depend on the initial conditions for the model. From the theoretical point of view, not much can be said about the accuracy in the definition of initial conditions for complex dynamic systems. One of the standard approaches to the problem of investigating such systems is to start from stationary solutions, and try to approximate the time-average non-stationary solutions by averages of stationary statistical solutions. Such attempts are based on the ergodic hypothesis. From the physical point of view, such stationary random solutions are a generalization of steady state in the N-body problem. This generalization requires a well-defined concept of equilibrium. In classical theory, the equilibrium is associated with a “microcanonical distribution” obtained either by the equi-partition over the set of approximately equal energy systems, or by the Gibbs procedure. If the time-invariance of associated probabilities can be preserved, near-equilibrium processes can be investigated through equilibrium properties of the system using the approach introduced in the practice of mathematical modelling by Langevin (see, for example [14]).

If turbulence is regarded as a perturbation of an equilibrium, then more precise definitions of “equilibrium” and “perturbation” are required. If the equilibrium is obtained from the idea of the equi-partition of the set of approximately equal energy systems and the initial energy density is finite, then the resulting ensemble does not possess the ergodic property [2]. This type of equilibrium was first constructed in [15, 27] using the Fourier expansion. In the limit of vanishing truncation of the Fourier series, the underlying process can be described by a set of functions that are *almost nowhere differentiable*. From the physical point of view the process is characterised by *the infinite energy density*. If the topological space in which the set of these functions can be embedded is a *complete linear metric space*, then under quite general assumptions [10] there exists no measures which are at least quasi-invariant relative to all translations, except for those identical to zero measure. Taking this into account, it is natural to regard real turbulent processes as those that are far from the Hopf-Lee equilibrium [2]. Mathematical investigations of such processes are intrinsically connected with the Fourier transform, and the theory of probability due to classic results of harmonic analysis¹³. At the core of such investigations from the physical point of view is the assumption that the provision of energy at large scales is the dominant effect which *can be modelled* through the limit of vanishing viscosity at small scales. There are two main types of mathematical models that are intrinsic to this approach:

- stationary models with exactly given boundary conditions;
- non-stationary models with exactly given initial conditions.

¹³the Bochner theorem and its generalization to Banach spaces [45, 52]

Both types of models have a common feature. Namely, non-linear terms in the models can be treated as a perturbation expansion ordered by a *small time-independent parameter*. In climate system study the Rossby-Kibel number often plays the role of such a parameter, as we explained in Section 5. This approach encounters serious difficulty when the viscosity decreases and equilibrium is understood in the Hopf-Lee sense. In this case the singularity of turbulence increases simultaneously with the decrease in viscosity. Mathematically, this difficulty can be formally overcome by using the Kolmogorov-Obukhov scaling law [22, 40], which implies a certain character for energy distribution between scales. That is, an inertial cascade of energy from the “whirling” scales to the dissipation scales can be represented by the Kolmogorov spectrum [2]

$$E(k) = C\epsilon^{2/3}k^{-5/3}, \quad (10.1)$$

where k is the wave number, ϵ is the rate of energy transfer across the spectrum, and C is an absolute constant. Recent results on deviations from this law can be found, for example, in [2]. Such deviations may be expected when *dispersion* is intrinsic to the model. This is always the case for mathematical models of real phenomena or processes where physical (chemical or biological) parameterization cannot be performed with infinite precision. In the general case, it is not enough to consider the limit of vanishing viscosity in order to adequately describe turbulence. We also need information of the character of dispersion [26]. Such a *correlation between viscosity and dispersion* has stimulated the search for different principles on which the statistical theory should be built [2].

The other interpretation of turbulence which has been recently proposed is based on the assumption of a small perturbation of a suitable Gibbsian equilibrium. However, if we accept the Gibbs hypothesis (see [35] and references therein) the nature of the convergence of the probabilities in the limit of vanishing viscosity remains open. The answer to this question is kept in the approximation of the Hamiltonian. In fact, the Gibbs probability distribution is defined as the probability of a collection of states by the Lebesgue integral of

$$C_a = 1/n_0 \exp(-\beta H), \quad \beta = 1/k_b T, \quad (10.2)$$

with respect to the Liouville measure. In (10.2) the notation is standard, that is k_b is the Boltzmann constant, T is the temperature of the system (macroscopic temperature), n_0 is a normalizing factor and H is the system Hamiltonian. The definition of the system Hamiltonian is a hierarchically approximating procedure. This implies a procedure for obtaining conditions that single out the canonical ensemble measure from the class of all probability measures on the phase space of the system. Such conditions determine stability conditions of the model [36].

If the scaling law is agreed upon then the nature of turbulence can be studied through different *approximations of the system Hamiltonian*. The connection between the approximate character of the Hamiltonian and the scaling law can quantify

turbulence numerically whenever *the law of dispersion* is established. As a result, the quality of mathematical models for turbulence is essentially determined by an adequate physical parameterization of the model that is linked to the hierarchical approximation of the system Hamiltonian.

11 Differential mathematical models for the climate study

Many rigorous mathematical results in the investigation of climate models were obtained for the barotropic atmosphere. In this case, there is in the governing system of equations at least one equation that relates two thermodynamic variables on the basis of an individual particle from time to time. An alternative formulation can be given by using a *piezotropic equation*, which allows us to relate two thermodynamic variables from one spacial point to another at a given moment of time. Since statistical spacial data for climate study is typically sparse, the latter case leads to the weak rather than differential formulation of the problem. In general, non-linear relationships between thermodynamic parameters which define the model suggest solution by numerical methods. Then, the quality of the underlying algorithm is completely determined by the correspondence of the model to the real climate.

We refer to the paper of J. Smagorinsky [53] on the basic experiment performed in collaboration with J.G. Charney, N.A. Phillips, and J. von Neumann. It is not reasonable to review here all subsequent steps in the development of the model that we use in this paper. The description is well-documented and the appropriate references related to the NCAR CCM3 model can be found in [21]. Instead, we give below the set of differential equations governing the atmosphere that is fundamental in climate study. It consists of five equations that relate the three components of wind $\mathbf{v} = (u, v, \omega)$, pressure p , density ρ and temperature T , namely

- the equations of motion

$$\left\{ \begin{array}{l} \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} + \omega \frac{\partial u}{\partial z} = 2\Omega(v \sin \varphi - \omega \cos \varphi) - \frac{1}{\rho} \frac{\partial p}{\partial x} + \frac{F_x}{m}, \\ \frac{\partial v}{\partial t} + u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} + \omega \frac{\partial v}{\partial z} = -2\Omega u \sin \varphi - \frac{1}{\rho} \frac{\partial p}{\partial y} + \frac{F_y}{m}, \\ \frac{\partial \omega}{\partial t} + u \frac{\partial \omega}{\partial x} + v \frac{\partial \omega}{\partial y} + \omega \frac{\partial \omega}{\partial z} = 2\Omega u \cos \varphi - \frac{1}{\rho} \frac{\partial p}{\partial z} - g + \frac{F_z}{m}, \end{array} \right. \quad (11.1)$$

- the continuity equation

$$\frac{\partial \rho}{\partial t} + u \frac{\partial \rho}{\partial x} + v \frac{\partial \rho}{\partial y} + \omega \frac{\partial \rho}{\partial z} = -\rho \left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial \omega}{\partial z} \right) \quad (11.2)$$

- the thermodynamic equation

$$d\varepsilon = c_p dT - \frac{1}{\rho} dp, \quad (11.3)$$

• and the state equation

$$\rho = \rho(p, T). \quad (11.4)$$

It is customary to use the Clapeyron equation $\rho = p/(RT)$ as the state equation. In this case the question of an adequate model of the latent heat of the phase transition is of primary importance because the gas constant R is different for different phases. In (11.3) $d\varepsilon$ is the heat added per unit mass, and the field of external forces in (11.1) is written excluding pressure-gradient forces (the pressure gradient is directed from high towards low pressure). In essence, equations (11.1)–(11.4) are *prognostic*. They provide a theoretical possibility for the mathematical forecast in climate study if we insert observed climatic conditions at a given time. A built-in physical parameterization splits the system (11.1)–(11.4) into prognostic equations and *tendency equations*.

Let us briefly recall typical assumptions made for simplifications of the system (11.1)–(11.4). The adiabatic approximation which we discussed in Section 4 can be obtained by setting $d\varepsilon = 0$, which provides a sufficient approximation for short range prediction. Moreover, if the atmosphere is assumed to be *barotropic*, the density at each point can be determined solely by the pressure at that point. In this case we have

$$\rho = \rho(p), \quad T = T(p), \quad (11.5)$$

(the second relationship follows from the Clapeyron equation).

Furthermore, if the motion is assumed to be horizontal, we have

$$\left\{ \begin{array}{l} \chi v = \frac{1}{\rho} \frac{\partial p}{\partial x}, \\ \chi u = -\frac{1}{\rho} \frac{\partial p}{\partial y}, \\ g = -\frac{1}{\rho} \frac{\partial p}{\partial z} + 2\Omega u \cos \varphi. \end{array} \right. \quad (11.6)$$

If we neglect the Coriolis term on the right-hand side of the last equation in (11.6), the equation turns into the *(hydro)static equation*. The first two equations of the system are *geostrophic wind equations*¹⁴ which may provide a good approximation in middle and high latitudes. This approximation simplifies numerical procedures because, for the barotropic atmosphere, geostrophic wind does not increase with height [11].

¹⁴or gradient balance equations

From the system (11.6) for the hydrostatic approximation pressure can be excluded. This results in *the thermal wind equations*

$$\begin{cases} \frac{\partial v}{\partial z} = \frac{g}{\chi T} \frac{\partial T}{\partial x} + \frac{v}{T} \frac{\partial T}{\partial z}, \\ \frac{\partial u}{\partial z} = -\frac{g}{\chi T} \frac{\partial T}{\partial y} + \frac{u}{T} \frac{\partial T}{\partial z}. \end{cases} \quad (11.7)$$

Again, numerically it leads to essential simplifications, since in the barotropic atmosphere the vertical temperature-gradient term is equal and opposite to the horizontal temperature gradient term. Of course, this gradient is not negligible in the general case.

In the baroclinic atmosphere a strong dependency between the horizontal temperature gradient and the vertical wind shear requires an appropriate choice of *the vertical coordinate*. The interdependency of different components of climate becomes important. However, the majority of implementations of the hydrological cycle into model (11.1)–(11.4), as well as the development of global general circulation models¹⁵, are essentially based on *the hydrostatic equilibrium assumption* [30, 31, 12]. The necessity for the development of a non-hydrostatic type of model has been realised during recent years [56].

Having included more than one climate components into a unified mathematical model, it becomes increasingly important to adequately formulate the hypothesis on *subgrid scale vertical/horizontal mixing*. Currently this hypothesis is formulated on the basis of the von-Karman-Prandtl logarithmic law (6.3) in the region of wall-bounded turbulent shear flow [60, 43, 44]. Since in general this law may lead to an inappropriate scaling [2, 3], more general laws for the interaction of space-time scales should be applied for the climate study.

12 Phase transitions and the algorithmic stability

When temperature changes, water vapor departs from the ideal conditions, making an adequate model of water vapor transport one of the most important and difficult problem in climate study. The thermodynamics of water vapor and moist air is closely connected with the problem of *latent heat and phase transitions*. Even if the temperature of a substance remains constant, whenever this substance changes phase (evaporates, melts, condenses, freezes etc) a quantity of heat, called *the latent heat of the phase change*, must be supplied to or taken away from the substance. The quantification of the latent heat is based on the concept of entropy, and is often performed by using the Clausius-Clapeyron equation that relates the saturation of vapor pressure to the latent heat of a phase transition. Conceptually, this equation together with the state equation (11.4) is time-independent. This leads

¹⁵that include in addition to the atmosphere other climate components such as ocean, land surface, sea ice

to an approximation of the mathematical model (11.1)-(11.4) whenever a physical parameterization is applied.

In most latitudes at most times of the year the atmospheric pressure and temperature vary continuously with time. As a result, *the geostrophic balance is never reached and maintained no matter how small the time-interval is assumed*. Rather, we observe a *continuous re-adjustment of the fields* with changing pressure and temperature fields. This requires the formulation of tendency equations that in turn require some *a-priori* knowledge of the vertical structure of meteorological fields. On the other hand, the knowledge of the vertical structure of the meteorological processes is a major output in integrating the prognostic equations. Hence, although prognostic equations can provide *a-posteriori information*, they must always be supplemented by the tendency equations (which are based on *a-priori information*) in order to form a *closed system of mathematical equations*. The tendency equations are typically based on additional physical hypotheses (like hydrostaticity), and are approximate in their nature. The original system (11.1)-(11.4) is always replaced by its approximation, not only because of inevitable approximations of the functions F_x, F_y, F_z and initial and boundary conditions, but also because of the approximate nature of the equations (11.3) and (11.4) for any specific model. Naturally, this leads to attempts to improve such approximations by “building-in” to the model other components of climate such as the ocean, land-surface, sea ice. In these cases the vertical structure of meteorological processes cannot be defined in the simple manner of the barotropic approximation with the altitude typically defined by

$$z(x, y, t, p) = z_0(x, y, t)\psi_0(p). \quad (12.1)$$

On the other hand, if in the general baroclinic case the altitude is approximated in a multilevel manner as

$$z(x, y, t, p) = \sum_{i=0}^N z_i(x, y, t)\psi_i(p), \quad (12.2)$$

then the functions $\psi_i(p), i = 0, 1, \dots, N$ should be chosen on the basis of *a-priori information*. For any finite number N , each of the functions $z_i(x, y, t)$ becomes a parameter of a given mathematical model that can, in principle, be expressed in terms of the values of $z(x, y, t, p)$ at the given level of pressure p_i . Hence the question arises as for *the optimal a-priori choice* of the functions $\psi_i(p)$. Such a choice is *multilevel* by its nature. It requires an interpolation between given levels on the basis of some qualitative *a-priori assumptions*. These assumptions have to ensure *the well-posedness of the model* (see details in [37]).

We recall that in short-range climate study, neither external energy generation nor dissipation of energy due to internal processes are taken into account. Let us assume now that h is the smallest scale of motion described by a discretized system of prognostic equations. Firstly, we note that the consideration of a discretized system is natural, at least due to imperfections in the measurement of meteorological

fields. In reality, even at very high resolution, the scale h still exceeds the scales of regions of energy dissipation. However, if we neglect the energy dissipation then the energy transferred between scale spectrum finally reaches the scale of the order h and accumulates there without a dissipation. As a result, the non-homogeneities of meteorological fields with scales of the order h may increase in time when $t - t_0$ increases, inducing *non-linear instability*. This leads to a continuous correction of the model by a more precise definition of the law of dissipation. From the mathematical point of view, whenever

$$h \rightarrow 0 \text{ and/or } t - t_0 \rightarrow \infty, \quad (12.3)$$

the dependency of h on τ and/or the dependency of $t - t_0$ on τ becomes important.

13 Computational decoupling

The system (11.1)-(11.4) is a strongly coupled system of mathematical equations. Its solution cannot be obtained by analytical approaches unless substantial simplifications are made. Such simplifications may dramatically influence the validity of the final result. On the other hand, any specific physical parameterization of the model also implies an inevitable mathematical approximation as we explained in Section 12. Due to such an approximation, conservative properties of the original system may only be preserved approximately. The accuracy of approximations of conservation laws is determined by *the physical parameterization and the degree of coupling in the original mathematical model*. Essentially, any physical parameterization that is “built-in” to a mathematical model *splits the model into components*. However, in principle, the connections between such components can be restored computationally. The quality of the restoration depends on the number of model components and the quality of the physical parameterization with respect to the real processes and phenomena.

In the Climate System Model developed by NCAR there are four main components, namely atmosphere, ocean, land, and sea-ice. The connection between these components are realized using the Flux Coupler code [5]. This code is constructed under the assumption that conservative properties can be preserved for momentum, heat, and freshwater under message passing. In turn, this assumption inevitably leads to an approximation of the original model. Even if we assume that initial data (initial and boundary conditions) are given with an appropriate precision, additional assumptions for the energy dissipation law at the top of the atmosphere should be made by *a-priori arguments*. This implies an approximate character not only for the mathematical expression of physical laws, but an approximate character for the physical parameterization of the model as well.

Reasonable *a-priori assumptions* may be derived on the basis of experimentation and observations. They can provide a tool for the analysis of *the adaptive re-adjustment of meteorological fields*. However, a formal expression of such adaptive

procedures requires some *a-posteriori arguments*. Such *a-posteriori* arguments are usually based on *the concept of continuity* [17]. Having both *a-priori assumptions and a-posteriori arguments* we can, in principle, validate the model ensuring its stability. In the general case the validation of mathematical models for complex dynamic systems can only be conducted with incomplete information. As a result, in reality it is practically impossible to achieve 100% reliability of the model. However, it is possible to achieve a *balance between the reliability of the model subjected to the physical parameterization and the efficiency of a numerical algorithm for its solution*. The procedure for achieving such a balance requires *the adaptive error control* that is based on *a-posteriori information about the computed solution*. In turn, the processing of such information requires *a-priori information on the exact solution*. Under quite general assumptions problems of this type can be formalized mathematically in the form of a hyperbolic type partial differential equation with respect to the control function [35]. From the practical point of view, the well-posedness of the original model depends on a *constant of hydrodynamic stability*, C^s , that quantifies stability properties of a dual problem with coefficients which depend on exact and computed solutions as well as on the period of time $T = t - t_0$ during which the model is integrated [18]. In this interpretation, the validation of the original mathematical model is eventually determined by the evaluation of the quantity C^s , which for climate system models is closely associated with the relaxation time τ defined by (4.3). We will address the issues of such an evaluation elsewhere. Here we note that the foundation of theory in this direction was laid by the works [17, 7, 18] (see also references therein).

14 Numerical results.

In order to conduct a computational experiment we have used the National Center for Atmospheric Research model [1]. It can be run in three main modes, namely interactive, batch and message passing. Only the first two were used in our experiment. The original file contains a C-shell script “setup” that can be used for the configuration of the model (type of dynamics, resolution etc). The “setup” creates a directory with the configuration specific name, for example, in one of our cases it was “cray.t42.spectral.som/”. On a SUN station with the SunOS operating system at the CIAM, University of South Australia we compiled the model with the command: “make sunos”. This creates the executable file “ccm3bin” in a subdirectory “run/”. Boundary datasets were taken from the NCAR WWW domain in the IEEE binary format. In the interactive mode the “setup” generates two standard namelists for initial and restart runs that can be used to tune in the model. The namelist files can be written at the discretion of a user as described in [1], p.27-44. We also note that running the model on SUN SPARC stations an increase of the stack size is often necessary, subject to the resolution used.

When we have to perform the simulation for a longer period it is convenient

to run the model in the batch mode. The Fujitsu VPP-300 supercomputer at the Australian Supercomputer Facility was used for such a simulation. The VPP system is a distributed memory machine. Availability on this machine of the vectorizing and parallelizing UXP/V Fortran-90 [57, 58] compiler and other software capabilities makes this computer effective for the high-speed computation required for a CCM3 run. In this computational experiment only one processor was used to run a vectorized version of the CCM3 code in standard and slab ocean versions. A few bugs reported recently through the CCM-Users E-Mail Group were fixed. For a parallelized code the message passing using the PVM facility should be implemented, which can be seen as a future development of this work.

Below we present typical outputs obtained as a result of climate simulation. In this paper we have not attempted to investigate the error of this simulation. As follow from the above discussion, the total error consists of three parts:

- the error of initial data at the start of computer simulation,
- error of the finite set of differential equations in the description of climate, and
- the error of the numerical algorithm that is used.

For such models as the NCAR CSM, the total error obtained from the contribution from all three sources is practically infeasible. Instead, we are currently developing a technique for the evaluation of such an error for a simplified model. The purpose of numerical results presented here is to give a comprehensive graphical interpretation of several physical fields that have been computed and can be used for future analysis.

In the simulation of climate we used the standard input datasets ([1], p.45-50). As the main output, the model generates so-called *history files* that are in a binary format. They provides the information on a set of temporal samples. Field values at any given moment of time correspond to different latitudes ([1], p.51). Different plotting programs can be used for the interpretation of the results of outputs. We used a modified version of the code developed at the Global Change Research Center, Portland State University by Gerhard W. Gross whose help is gratefully acknowledged. This code reads history files and plots them using the GNUPLOT Plotting Program. In Figure 1 - 5, typical initial distributions are presented for the following fields respectively

- surface geopotential in m^2/s units;
- surface pressure in Pa units;
- zonal wind component in m/s units;
- meridional wind component in m/s units;
- sea-surface temperature field in C units.

CLIMATE MODELLING WITH CCM3: FIELD PHIS (lat,long) Vert level = 1 Time Step = 1

PHIS

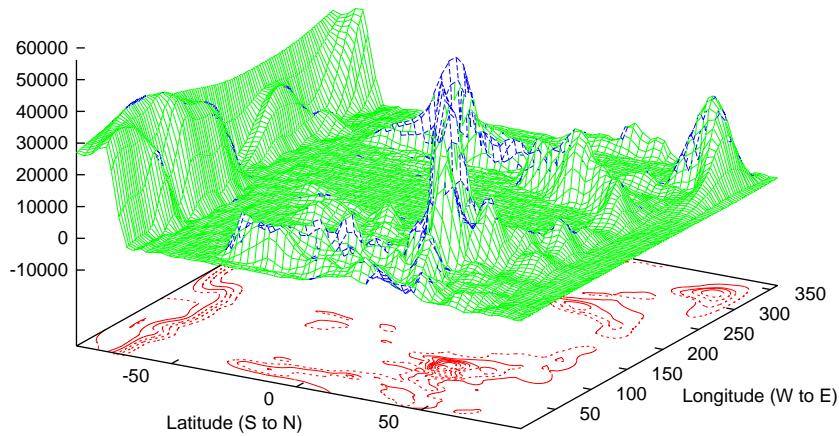


Figure 1: Surface geopotential (PHIS).

CLIMATE MODELLING WITH CCM3: FIELD PS (lat,long) Vert level = 1 Time Step = 1

PS

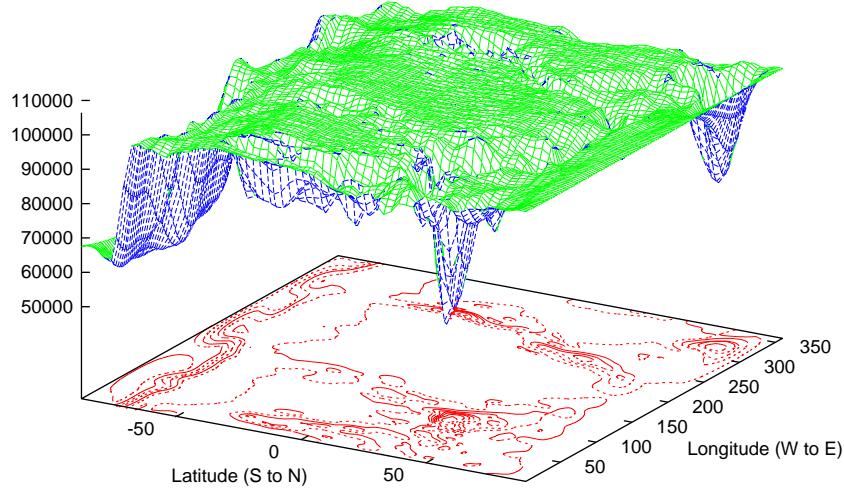


Figure 2: Surface pressure (PS).

CLIMATE MODELLING WITH CCM3: FIELD U (lat,long) Vert level = 1 Time Step = 1

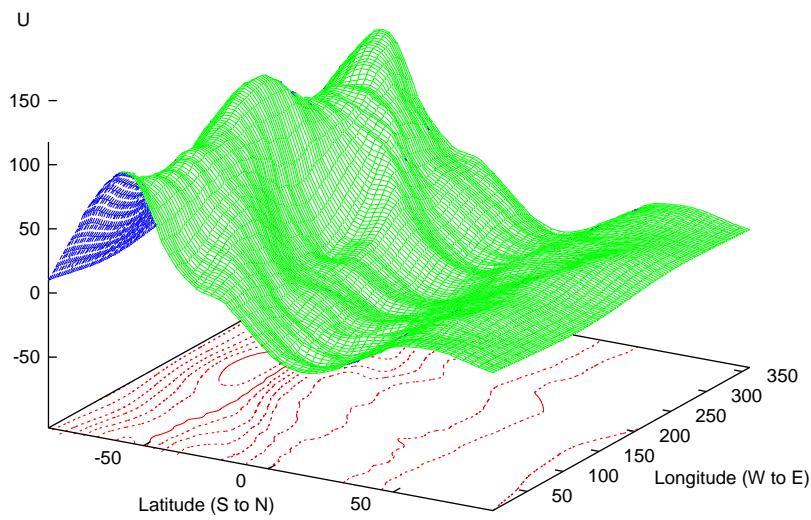


Figure 3: Zonal component of the wind (U, vertical level 1).

CLIMATE MODELLING WITH CCM3: FIELD V (lat,long) Vert level = 1 Time Step = 1

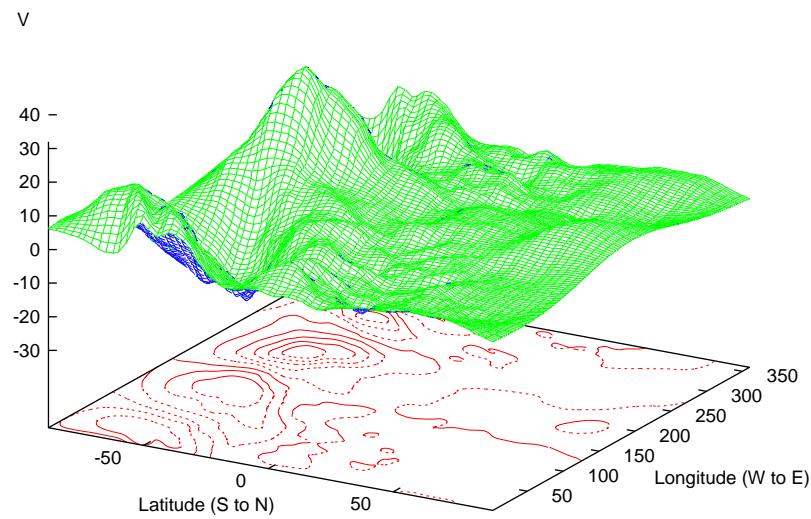


Figure 4: Meridional component of the wind (V, vertical level 1).

CLIMATE MODELLING WITH CCM3: FIELD SST (lat,long) Vert level = 1 Time Step = 1

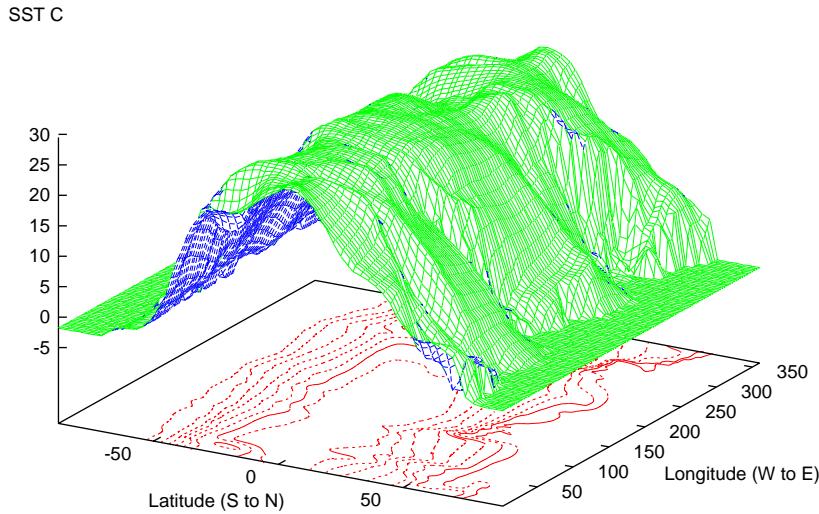


Figure 5: Sea-surface temperature distribution (SST).

In Figure 6 and 7 zonal and meridional components of the wind are presented at vertical level 15. The temperature field is presented in Figure 8 for vertical level 15 in K units. Finally, the water vapour field at the vertical level 15 in Kg_{H_2O}/Kg_{air} units is presented in Figure 9. The complete Master Field List of the NCAR CCM3 model and available options for the output of the model can be found in [1], p. 45-68.

15 Conclusions and future directions

The coupled simulation for climate system models provides an efficient tool for climate study. Moreover, the concept of coupling in modelling complex dynamic systems reflects one of the most general ways of implementing new effects and new information into mathematical models. However, the refinements of the approach based on coupling procedures for such complex systems as climate may continue indefinitely. The two natural ways to meet the arising challenge were discussed in this paper. First, it is natural to start with a *finite set of independent models* and try to couple them by informational message passing using certain physical principles such as conservation laws. The NCAR Climate System Model is of this type. We presented several numerical examples obtained on the basis of this approach. For the NCAR CSM the set of independent components consists of four mathematical models for atmosphere, ocean, land surface, and sea ice. Although the conservative properties of the whole system cannot be guaranteed in general, certain key com-

CLIMATE MODELLING WITH CCM3: FIELD U (lat,long) Vert level = 15 Time Step = 1

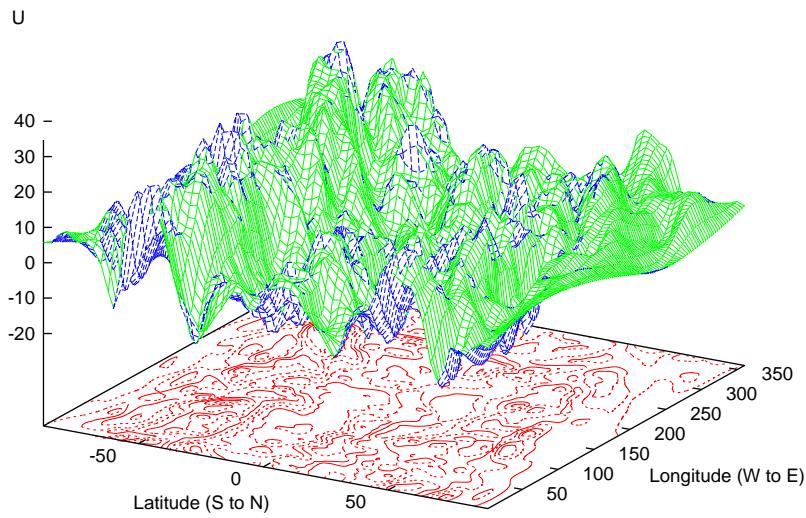


Figure 6: Zonal component of the wind (U, vertical level 15).

CLIMATE MODELLING WITH CCM3: FIELD V (lat,long) Vert level = 15 Time Step = 1

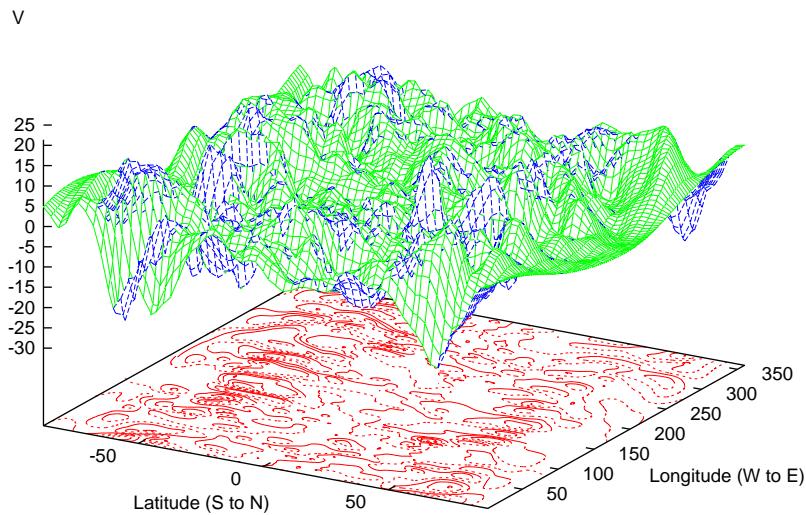


Figure 7: Meridional component of the wind (V, vertical level 15).

CLIMATE MODELLING WITH CCM3: FIELD T (lat,long) Vert level = 15 Time Step = 1

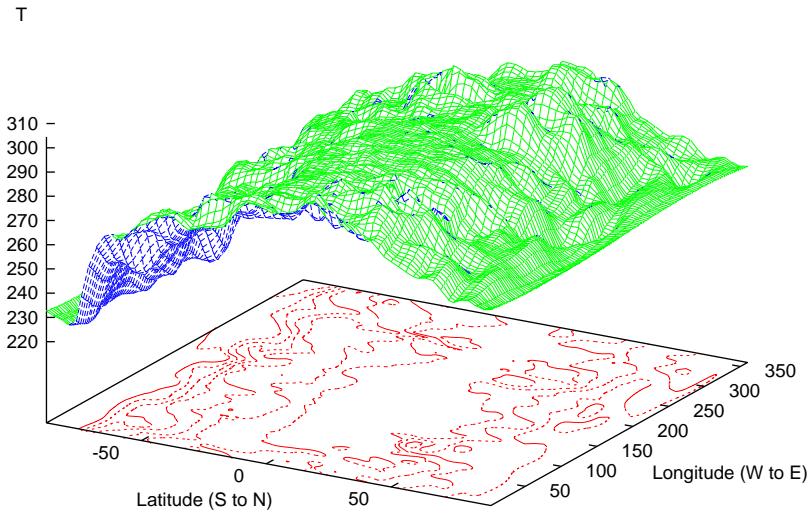


Figure 8: Temperature field at vertical level 15.

CLIMATE MODELLING WITH CCM3: FIELD Q (lat,long) Vert level = 15 Time Step = 1

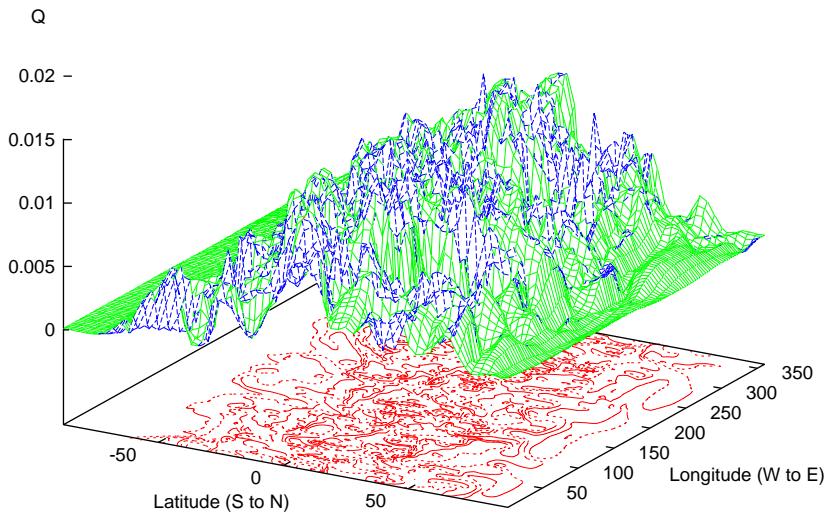


Figure 9: Water vapour at vertical level 15.

ponents, such as momentum, heat, freshwater, can be preserved *numerically*. The quality of such models depend on the dynamics of the error propagation that can be controlled by the Coupler Code. With additional information becoming available, this approach requires an increasingly complex code. In the end, we have to deal with the coupling phenomenon from the very beginning of the process of modelling. On the other hand, the sparse character of available informational datasets for such complex dynamic systems as climate makes *the concept of independency* for the model components natural, at least at the initial step of the process of modelling.

The dilemma between independence and coupling has led us to the necessity of considering another approach to the modelling of complex dynamic systems. We came to the conclusion that such an approach has to be based not only on a-priori information about the system (that is incomplete in its nature) but also on a-posteriori information. This allows us to construct numerical algorithms that can be analyzed in the traditional manner of using a-priori information, but the procedure of construction has to be based on a-posteriori information. In this case *the choice of the norm for the error control is model-specific, being influenced by the physical parameterization of the mathematical model*. For any given physical parameterization, conservation laws may be implemented only approximately into mathematical models of complex dynamic systems. As a result, the standard energy norms may not provide an appropriate choice for the error control in mathematical models of such systems. This idea is the basis for the future development of the presented work. Under a fixed degree of coupling and a given physical parameterization we need a scale of a-priori and a-posteriori estimates in a spectrum of norms to ensure the stability of the model. Numerical algorithms based on such estimates allow an adaptive error control within the chosen spectrum. In turn, this allows the construction of adaptive computational codes that can be effectively used in the study of complex dynamic systems with many transient states.

ACKNOWLEDGEMENT.

The author wishes to acknowledge the support of the School of Mathematics at the University of South Australia where a major part of this work was completed. The author thanks Dr Murray Dow from the Australian Supercomputer Facility for permission to use a vectorized version of the CCM3 model to obtain a series of computational results on the VPP-300 supercomputer and Dr Gerhard Gross from Portland State University for the help in graphical interface. The support of the leader of the Environmental Modelling Research Group at the Centre for Industrial and Applied Mathematics Professor J.Filar and the help of his collaborators, Mr P. Gaertner, Dr J. Day, and Dr S. Lucas is also gratefully acknowledged.

References

- [1] Acker, T.L., et al User's Guide to NCAR CCM3, *NCAR TN-421*, May, 1996.

- [2] Barenblatt, G.I., and Chorin, A.J. Scaling laws and vanishing viscosity limits in turbulence theory, *Department of Mathematics, University of California, Berkeley*, 1996.
- [3] Barenblatt, G.I., Chorin, A.J., and Prostokishin, V.M. Scaling laws for Fully Developed Turbulent Flow in Pipes: Discussion of Experimental Data, *Department of Mathematics, University of California, Berkeley*, 1996.
- [4] Briegbleb, B.P. Delta-Eddington Approximation for Solar Radiation in the NCAR Community Climate Model, *Journal of Geophysical Research*, Vol. 97, No.D7, 1992, 7603-7612.
- [5] Bryan, F.O., et al The NCAR CSM Flux Coupler, *NCAR TN*, May, 1996.
- [6] Chang, P., and Philander, S.G. A coupled ocean-atmosphere instability of relevance to the seasonal cycle, *Journal of the Atmospheric Sciences*, 51, No.24, 1994, 3627-3648.
- [7] Eriksson, K., and Johnson, C. Adaptive finite element methods for parabolic problems V: long-time integration, *SIAM J. Numer. Anal.*, 32, No. 6, 1995, 1750-1763.
- [8] Farrell, B.F. Small Error Dynamics and the Predictability of Atmospheric Flows, *Journal of the Atmospheric Sciences*, 47, No. 20, 1990, 2409-2416.
- [9] Fleming, W.H., and Soner, H.M. Controlled Markov Processes and Viscosity Solutions, *Springer-Verlag*, 1993.
- [10] Gelfand, I.M., and Vilenkin, N. Ya. Generalized Functions. Applications to Harmonic Analysis, *Academic Press*, Vol. 4, 1964, Chapter 4, Section 5.
- [11] Hess, S. L. Introduction to Theoretical Meteorology, *Holt, Rinehart & Winston*, 1959, New York.
- [12] Holloway, J.L. Jr., and Manabe, S. Simulation of Climate by a Global General Circulation Model, *Monthly Weather Review*, Vol. 99, No. 5, 1971, 335-370.
- [13] Holtslag, A.A.M., and Boville, B.A. Local Versus Nonlocal Boundary-Layer Diffusion in a Global Climate Model, *Journal of Climate*, Vol. 6, 1993, 1825-1842.
- [14] Honerkamp, J. Stochastic Dynamical Systems. Concepts, Numerical Methods, Data Analysis. *New York: VCH*, 1994.
- [15] Hopf, E. Statistical hydrodynamics and functional calculus, *J. Rat. Mech. Anal.*, 1, 1952, 87-142.
- [16] Jin, F.F., and Neelin, J.D. Modes of interannual tropical ocean-atmosphere interaction - a unified view, *Journal of Atmospheric Sciences*, 50, No.21, 1993, 3477-3503.
- [17] Johnson, J.R. A new paradigm for adaptive finite element method, in *The Mathematics of Finite Elements and Applications*, Ed. by J.R. Whiteman, 1994, 105-120.
- [18] Johnson, C., Rannacher, R., and Boman, M. Numerics and hydrodynamic stability: toward error control in computational fluid dynamics, *SIAM J. Numer. Anal.*, 32, No.6, 1995, 158-1079.
- [19] Kasahara, A., and Washington, W. NCAR Global General Circulation Model of the Atmosphere, *Monthly Weather Review*, Vol. 95, No. 7, 1967, 389-402.
- [20] Kasahara, A. Various Vertical Coordinate Systems Used for Numerical Weather Prediction, *Monthly Weather Review*, Vol. 102, 1974, 509-522.
- [21] Kiehl, J.T., et al Description of the NCAR Community Climate Model (CCM3), *NCAR TN -420*, 1996.
- [22] Kolmogorov, A.N. Local structure of turbulence in an incompressible fluid at a very high Reynolds number, *Dokl. Akad. Nauk. SSSR*, 30, 1941, 299-302.

- [23] Kreiss, H.O. Problems with different time scales, *in Acta Numerica*, 1992, 101-139.
- [24] Kurihava, Y. Numerical Integration of the Primitive Equations on a Spherical Grid, *Monthly Weather Review*, Vol. 93, No. 7, 1965, 399-415.
- [25] Lax, P.D. Hyperbolic systems of conservation laws and the mathematical theory of shock waves, *SIAM Publications*, Philadelphia, 1972.
- [26] Lax, P.D. The zero dispersion limit, a deterministic analog of turbulence, *Comm. Pure Appl. Math.*, 44, 1991, 1047-1056.
- [27] Lee, T.D. On some statistical properties of hydrodynamic and hydromagnetic fields, *Quarterly Appl. Math.*, 1952, 69-72.
- [28] Lin, S.-J., et al A class of the van Leer-type transport schemes and its application to the moisture transport in a General Circulation Model, *Monthly Weather Review*, 122, 1994, 1574-1592.
- [29] Ma, C.-C., et al Sensitivity of a Coupled Ocean-Atmosphere Model to Physical Parameterizations, *Journal of Climate*, Vol. 7, 1994, 1883-1896.
- [30] Manabe, S., Smagorinsky, J., and Strickler, R. Simulated Climatology of a General Circulation Model with a Hydrological Cycle, *Monthly Weather Review*, Vol. 95, No. 7, 1967, 389-402.
- [31] Manabe, S. Climate and the Ocean Circulation, *Monthly Weather Review*, Vol. 97, No. 11, 1969, 739-774.
- [32] Marchuk, G.I. Splitting and Alternating Direction Methods, *in Handbook of Numerical Analysis*, Vol. 1, Ed. by P.G.Ciarlet and J.L.Lions, 1990, Elsevier Science Publishers, 199-462.
- [33] McCreary, J.P., Jr., and Anderson, D.L.T. An Overview of Coupled Ocean-Atmosphere Models of El-Nino and the Southern Oscillation, *Journal of Geophysical Research*, 96, No.28, 1991, 3125-3150.
- [34] Melnik, V.N. Semi-implicit finite-difference schemes with flow correction for quasihydrodynamic models, *Engineering Simulation*, 1995, 856-865.
- [35] Melnik, V.N. Nonconservation law equation in mathematical modelling: aspects of approximation, *Proc. of the International Conf. AEMC'96*, Sydney, 1996, 423-430.
- [36] Melnik, V.N. Optimal probabilistic trajectories of deterministic finite-state machines, *School of Mathematics, University of South Australia*, TR 15, 1996, 1-25.
- [37] Melnik, R.V.N. Error Dynamics and Coupling Procedures in Mathematical Climate System Models, *accepted, to appear in Proceedings of the XVth World Congress on Scientific Computation, Modelling and Applied Mathematics, August 1997*, Berlin, Germany.
- [38] Monin, A. S. Weather Forecasting as a Problem in Physics, *The MIT Press*, 1972.
- [39] Muller, I., and Ruggeri, T. Extended Thermodynamics, *Springer-Verlag*, 1993.
- [40] Obukhov, A.M. Spectral energy distribution in turbulent flow, *Dokl. Akad. Nauk SSSR*, 32, 1941, 22-24.
- [41] Peshkov, V.A. Second sound in helium II, *Journal of Physics*, 3, 1944, 381.
- [42] Phillips, N.A. A Coordinate System Having Some Special Advantages for Numerical Forecasting, *Journal of Meteorology*, Vol. 14, 1957, 184-185.
- [43] Prandtl, L. Bericht über Untersuchungen zur ausgebildeten Turbulenz, *Zeitschr. angew. Math. Mech.*, 5, 1925, 136-139.

- [44] Prandtl, L. Zur turbulenten Stroemung in Rohren und laengs Platten, *Ergeb. Aerodyn. Versuch.*, Series 4, 1932, Goettingen.
- [45] Prokhorov, Yu.V. Convergence of random processes and limit theorems of probability theory, *Theory Probab. and Appl.*, 1, 1956, 157-214.
- [46] Randall, D.A., et al Analysis of snow feedback in 14 general circulation models, *Journal of Geophysical Research*, 99, No. D10, 1994, 20757-20771.
- [47] Rasch, P.J., and Williamson, D.L. The Sensitivity of a General Circulation Model Climate to the Moisture Transport Formulation, *Journal of Geophysical Research*, 96, No. D7, 1991, 13123-13137.
- [48] Rasch, P.J., et al A three-dimensional transport model for the middle atmosphere, *Journal of Geophysical Research*, 99, No. D1, 1994, 999-1017.
- [49] Read, P.L. Applications of chaos to meteorology and climate, in *The Nature of Chaos*, Ed. by T. Mullin, Clarendon Press, Oxford, 1994, 222-260.
- [50] Reynolds, C.A., Webster, P.J., and Kalnay, E. Random error Growth in NMC's Global Forecasts, *Monthly Weather Review*, 122, 1994, 1281-1305.
- [51] Richardson, L.F. *Weather Prediction by numerical Process*, Cambridge Univ. Press, London, 1922.
- [52] Sazonov, V. A note on characteristic functionals, *Theory Probab. and Appl.*, 3, 1958, 188-192.
- [53] Smagorinsky, J. General Circulation Experiments with the Primitive Equations, *Monthly Weather Review*, Vol. 91, No. 3, 1963, 99-164.
- [54] Shashkov, M. Conservative Finite-Difference Methods on General Grid, *CRC Press*, 1996.
- [55] Struwe, M. Variational Methods. Applications to Nonlinear PDEs and Hamiltonian Systems, *Springer-Verlag*, 1990.
- [56] Tripoli, G.J. A nonhydrostatic mesoscale model designed to simulate scale interaction, *Monthly Weather Review*, 120, 1992, 1342-1359.
- [57] UXP/V Fortran-90/VP. User's Guide (V10), *J2U5-0050-01EN*, Fujitsu, 1995.
- [58] UXP/V Fortran-90 Messages (V10), *J2U5-0060-01EN*, Fujitsu, 1995.
- [59] Vernotte, P. La veritable equation de la chaleur, *Comtes Rendus Hebd. Seances Acad. Sci.*, 247(23), 1958, 2103-2105.
- [60] Von Karman, T., Mechanische Aehnlichkeit und Turbulenz, *Nach. Ges. Wiss. Goettingen Math.-Phys. Klasse*, 1932, 58-76.
- [61] Wakata, Y., and Sarachik, E.S. Nonlinear Effects in Coupled Atmosphere-Ocean Basin Modes, *Journal of Atmospheric Sciences*, 51, No.6, 1994, 909-920.
- [62] Williamson, D.L., and Olson, J.G. Climate simulation with a semi-Lagrangian version of the NCAR Community Climate Model, *Monthly Weather Review*, 122, 1994, 1595-1610.
- [63] Wu, G., and Lau, N.C. A GCM simulation of the relationship between tropical-storm formulation and ENSO, *Monthly Weather Review*, 120, 1992, 958-977.

UNIVERSITY OF SOUTH AUSTRALIA



TECHNICAL REPORT SERIES

**On Consistent Regularities
of Control and
Value Functions**

by

(R.) **V Nick Melnik**

Report No. 1996/16

CENTRE FOR INDUSTRIAL
AND APPLIED MATHEMATICS
SCHOOL OF MATHEMATICS

Faculty of Information Technology

The Levels, South Australia 5095, Telephone (08) 8302 3343 Facsimile (08) 8302 5785

TECHNICAL REPORT SERIES

**On Consistent Regularities
of Control and
Value Functions**

by

(R.) V Nick Melnik

Report No. 1996/16

ON CONSISTENT REGULARITIES OF CONTROL AND VALUE FUNCTIONS

V.Nick Melnik

E-mail: matvnm@lv.levels.unisa.edu.au

Abstract

In this paper we deal with nonsmooth optimal control problems in the case when the control is allowed to be a discontinuous function. We analyse smoothness assumptions on an adjoint process in deterministic and stochastic cases. Possibilities of steep generalized space-gradients of the adjoint function imply the necessity of an approximation of the Hamiltonian. The key question of such an approximation is a relationship between the control and the value function. Under quite general assumptions it is proved that the performance measure for the original process is determined by the control function with possible discontinuities.

Key words: Hausdorff topological spaces, probabilistic weight functions, discrete control, approximations of the Hamiltonian, coupling regularities.

1 Introduction: mathematical models in topological spaces and ramifications from initial conditions given approximately.

Many problems in applications of mathematics can be formulated in the following generic form. Let $H(u, v)$ be a nonlinear operator defined on the product of the two topological spaces (T, τ) and (X, h) (which may coincide) with values in a topological space (\mathcal{E}, ω) , where T, X, \mathcal{E} are sets that form space-supports, whereas τ, h , and ω are systems of their subsets that define topologies of the corresponding spaces. Let us denote the set of all ordered pairs $\{(t, x) : t \in T, x \in X\}$ by $\Omega_0^0 = T \otimes X$. We assume that it is known that for a given (computed, estimated, observed, etc) value $v = v_0 \in \Omega_0^0$ the equation $H(u, v_0) = 0$ has a solution $u_0^\epsilon \in \Omega_0^0$. That is, for arbitrary small $\epsilon > 0$ the following condition holds:

$$H(u_0^\epsilon, v_0) \rightarrow 0 \text{ when } \epsilon \rightarrow 0^+. \quad (1.1)$$

Assume that the limit of u_0^ϵ exists and belongs to Ω_0^0 when ϵ tends to zero from the right,

$$\lim_{\epsilon \rightarrow 0^+} u_0^\epsilon = u_0, \quad u_0 \in \Omega_0^0, \quad (1.2)$$

In this case, from the analysis of (1.1), (1.2), a natural question which arises is to find such solutions of the equations

$$H(u, v) = 0 \quad (1.3)$$

for values of the parameter v that are close to v_0 in some sense and which “branch” from u_0^ϵ when $\epsilon \rightarrow 0^+$. The solution of (1.3) depends on the choice of the non-linear operator H as well as on the parameter v , that may be a functional, a function, or a numerical parameter. From the definition of

topology, it follows that $\emptyset \subset \tau$, $\emptyset \subset h$, and $\emptyset \subset \omega$. If we further assume that $u \in \mathcal{E}_1$, $v \in \mathcal{E}_2$, where \mathcal{E}_1 and \mathcal{E}_2 are given topological spaces (not excluding the possibility $\mathcal{E}_1 \equiv \mathcal{E}_2 \equiv \Omega_0^0$), this allows the possibility of a separation of the two elements u and v that in principle belong to the same set Ω_0^0 . Of course, as a partial case, it is possible that $\mathcal{E}_1 \equiv T$ and $\mathcal{E}_2 \equiv \mathcal{X}$. If it is a-priori given that at least one of these topological spaces is complete (typically this assumption is imposed on T), then the operator H can be specified in a more definite way using a generalization of the theorem on continuity of complex functions.

Theorem 1.1 [17] *If there exists a continuous mapping $f : \mathcal{E}_1 \rightarrow \mathcal{E}_2$ and a continuous mapping $\phi : \mathcal{E}_2 \rightarrow \mathcal{E}$ then the mapping $F = \phi(f) : \mathcal{E}_1 \rightarrow \mathcal{E}$ is also continuous.*

Theorem 1.1 provides a technical convenience that allows us to assume the possible existence of a continuous mapping $F : T \otimes \mathcal{X} \rightarrow \mathcal{E}$. Hence, the non-linear operator H in (1.3) may also, in principle, be defined by a continuous mapping. In this case, a constructive element in the determination of all solutions of (1.3) that branch from u_0^ϵ in a range of v close to v_0 , is introduced by a specification of topologies in the spaces \mathcal{E}_1 , \mathcal{E}_2 and \mathcal{E} . The idea of such specifications in a rigorous mathematical sense came from the fundamental work of Newton [26] who considered the determination of branching solutions of (1.3) that tend to v_0 when $u \rightarrow u_0^\epsilon$ for the special case when

$$\lim_{\epsilon \rightarrow 0^+} \frac{\partial H}{\partial y}|_{(u_0^\epsilon, v_0)} = 0 \quad (1.4)$$

assuming that H can be expanded in a series of positive integral powers of $(u - u_0^\epsilon)$ and $(v - v_0)$ when $\epsilon \rightarrow 0^+$. Under these conditions Newton's technique consists of seeking the solution of (1.3) in the form

$$v = v_0 + \alpha_1(u - u_0^\epsilon)^{\epsilon_1} + \alpha_2(u - u_0^\epsilon)^{\epsilon_2} + \dots \quad (1.5)$$

where $\epsilon_1, \epsilon_2, \dots$ is an increasing sequence of rational numbers. The basic idea of the Newton polygon method (which is often referred to as the parallelogram method) is intrinsically connected with the geometric interpretation of the problem. It is also known that fractional powers $\epsilon_1, \epsilon_2, \dots$ have a finite common denominator and that the Newton polygon method can be used to find all continuous solutions such that $v \rightarrow v_0$ provided $u \rightarrow u_0^\epsilon$ and (1.4) holds [38]. From the topological point of view we implicitly assume that any point in \mathcal{E}_1 , \mathcal{E}_2 , and \mathcal{E} is a closed set, and a finite set of such sets is again closed¹. Then the problem (1.3) becomes reducible to the determination of possible values of

$$\alpha_1, \epsilon_1, \alpha_2, \epsilon_2, \dots \quad (1.6)$$

in (1.5). The principal difficulty in the solution of such a reduced problem in the general case resides in the fact that the terms in the sequence (1.6) are not independent of the parameter ϵ . Thus, mathematically speaking we need a reinforcement of the first axiom of separability to be able to determine v from (1.5) with an a-priori given accuracy. To overcome this difficulty we may use the Hausdorff axiom of separability in the topological spaces \mathcal{E}_1 , \mathcal{E}_2 , \mathcal{E} . Since the Hausdorff property is a hereditary property², it implies the possibility of the assumption

$$\lim_{\epsilon \rightarrow 0} H(u_0^\epsilon, v) = H(u_0, v_0), \quad (1.7)$$

where $u_0 \in \mathcal{E}_1$. Then by the standard change of variables

$$u = u_0 + x, \quad v = v_0 + y \quad (1.8)$$

the problem of "branching" from the point (u_0, v_0) is reduced to the problem of "branching" from the origin $(0, 0)$. Hence, in principle one can seek such solutions of the equation $\tilde{H}(x, y) = 0$, (here

¹thus, in such topological spaces the first axiom of separability holds

²any subspace of a Hausdorff space is again Hausdorff

$\tilde{H}(x, y) = H(u, v)$) that branch from the zero of \mathcal{E}_1 for values of v that are close to the zero of \mathcal{E}_2 . Also, due to (1.1), (1.2) it is usually assumed that $\tilde{H}(0, 0) = 0$. The idea of such branching-from-zero solutions³ is intrinsic to topological spaces equipped with the Hausdorff property. Indeed, in topological spaces with the Hausdorff separability axiom, we can claim simultaneous realization of two equalities, satisfied exactly, that in general connect three topological spaces. These equalities are introduced by (1.3) and by a relationship between a-priori given H , u_0 , and v_0

$$H(u_0, v_0) = 0, u_0 \in \mathcal{E}_1, v_0 \in \mathcal{E}_2, H \in \mathcal{E}. \quad (1.9)$$

Although this approach has proved to be very useful in many applications, it has virtually inherited the classical mechanics idealization of temporal evolution as a motion of phase points along phase point trajectories. There are a number of situations where the validity of such an idealization is limited. On the whole, arbitrarily close initial conditions for models of the general type (1.3), (1.9) can give rise to exponentially diverging trajectories exhibiting qualitatively distinct physical behaviour [24]. In the general case, (1.3) may provide an approximation to sequences like (1.6), but it does not imply that a specified sequence may provide an appropriate approximation to the solution of the model (1.3). As a result, the quality of approximations depends decisively on a-priori information on ϵ and the specification of topologies of the corresponding spaces.

In many applications a small numerical parameter λ plays the role of the variable v in the model (1.3). In these cases it is natural to attempt to represent small solutions of (1.3) in the form of series of integral (or fractional) powers of λ . However, if the uniqueness of the solution that branches from the origin is not part of the information given a-priori, this approach encounters serious mathematical difficulties. When the series is sought as integral powers of λ then the search for a majorant series is problem specific, whereas a fractional-power-series representation does not guarantee that all small solutions may be taken into consideration by such a form. Mathematically, these difficulties can be gradually removed by applying the Lyapunov-Schmidt theory in the topological spaces equipped with the Hausdorff axiom. Typically, such an approach is limited by either stationary models or by mathematical models, in which time appears as a pure deterministic category, being a parameter that labels states of the system [24]. Such mathematical formalization allows us to reduce the original mathematical model (1.1), (1.2) to a model with equalities satisfied exactly in (1.3) and (1.9) which in general are interpreted with the probability 1. This leads to a certain reversibility of mathematical models in time⁴ that allows us to use a set-theoretic approach for the construction of mathematical models. The validity of the reduction of the mathematical model (1.1), (1.2) to the model (1.3), (1.9) is essentially based on a postulate proposed by J.W. Gibbs.

Postulate 1.1 [10, 21] *An appropriate description of a macroscopic system in thermodynamic equilibrium may be given by certain probability measures on the phase space of the system.*

The main difficulty in a rigorous justification of this postulate lies with the fact that conditions for the measure stability are not independent from the specification of the operator H (which can be given only approximately) and the definition of topologies in the spaces \mathcal{E}_1 , \mathcal{E}_2 , \mathcal{E} . At the same time, the possibility of translation (1.8) is defined by the conditions of measure stability. Indeed, (1.9) is a limiting case in which measure stability is implicitly assumed by equipping of the space topologies a-priori with the Hausdorff property. However, instead of (1.8), it is more realistic to consider a perturbed translation

$$u = u_0^\epsilon + x, \quad v = v_0 + y_\lambda, \quad (1.10)$$

where $\lambda > 0$ is typically a small parameter. As a result, without some a-priori information on $\epsilon > 0$ and $\lambda > 0$, the model (1.3), (1.9) cannot be justified rigorously. No matter how small the parameters ϵ and λ are assumed to be, the translation (1.10) is always dependent on properties of the topology

³referred to as "small solutions" in the Lyapunov-Schmidt theory (see [38] and references therein)

⁴which physically speaking implies the invariance of densities along trajectories [20]

which is a-priori introduced in Ω_0^0 . If such a topology is assumed to be Hausdorff, then a transition from the point $(0, 0)$ to an arbitrary fixed point (u, v) in the space Ω_0^0 can not necessarily be described by a continuous governing operator $H : \Omega_0^0 \rightarrow \mathcal{E}$. On the other hand, if the governing operator H is continuous then the availability of the Hausdorff property in Ω_0^0 is questionable ([33], p.309). In the final analysis, simultaneous assumptions of continuity of the operator H and the Hausdorff property of the topological space on which it is defined cannot be justified rigorously for any mathematical model. The complexity of the possible justification was realized at the time when the groundwork in statistical mechanics was laid. In particular, J.W. Gibbs wrote that “we are rarely justified in excluding the considerations of the antecedent probability of the prior events” [9].

If we use the Hausdorff property as an a-priori given argument it is possible to assign the same probability, 1, to two non-identical events, that eventually allows us to assume that the governing operator H associated with these events may be continuous. In theory, such an extension of the continuity assumption from classical to statistical mechanics⁵ still preserves time-reversibility of the constructed mathematical models. However it does not take into account the fact that the definition of the non-linear operator H in the model (1.3) is an inherently recursive approximation that is not independent of the definition of the topology. This mathematical idealization has turned out to be very useful in many practical applications including areas of statistical mechanics and classical irreversible thermodynamics. The success of applications of mathematical tools in statistical mechanics has been grounded on postulate 1.1, whereas the success of mathematical modelling in thermodynamics has been determined by the local-equilibrium hypothesis.

Hypothesis 1.1 *Any system out of equilibrium is assumed to depend locally on the same set of variables as when it is in equilibrium.*

During recent years experimental and theoretical physicists have shown a growing interest in going beyond the limits of this hypothesis (see [25,13] and references therein). Associated mathematical challenges require relaxation of some mathematical assumptions typically made when classical approaches are applied even in the non-relativistic case. This paper deals with such relaxations, bearing in mind the problems of optimal control theory.

The remainder of the paper is organized as follows:

- In Section 2 we analyse mathematical definitions of controlled dynamic rules in deterministic and probabilistic cases when the topology of the underlying spaces is equipped with the Hausdorff property. We argue that if the control is allowed to be a discontinuous function, then a *sequential regularization procedure* is necessary to ensure system stability.
- In Section 3 we show why the stability of the system may be violated when passing from an optimal control problem to the associated Hamilton-Jacobi-Bellman equation. It is emphasized that an appropriate approximation of the Hamiltonian of a controlled dynamic system, considered in a Hausdorff topological space, is not necessarily provided by a continuous mapping.
- Section 4 deals with the case when both control and the value function are allowed to come from the non-Euclidean Banach space $L^1(Q_0^0)$ where Q_0^0 is an approximation of Ω_0^0 . Under quite general assumptions it is shown that, in this case, control can always be chosen arbitrarily close to a specified time-averaged performance measure. The main result has been derived without reference to a possible continuity of the value function. The proposed proof is a properly constructed algorithm which ensures the measure stability in Q_0^0 .
- In Section 5 we discuss the validity of assumptions made for the derivation of theorem 4.1 and logical issues related to this.
- Conclusions and future directions are addressed in Section 6.

⁵according to J.W.Gibbs “conservation of extension in phase”

2 Deterministic and stochastic models of optimal control in Hausdorff spaces.

As above, let Ω_0^0 be a set defined as $\mathcal{T} \otimes \mathcal{X}$. We assume that the evolution of a dynamic system takes place in Ω_0^0 and such evolution can be controlled to achieve a certain goal. Let us denote a space-time domain of definition for admissible controls as Σ , and the set of all admissible controls as \mathcal{U} . We assume that the state dynamics of the system can be effectively defined by the two coupled mappings:

$$f : \Omega_0^0 \otimes \mathcal{U} \rightarrow \mathcal{B}_1 \quad (2.1)$$

and

$$u : \Sigma \rightarrow \mathcal{U} \subseteq \mathcal{B}_2, \quad (2.2)$$

where \mathcal{B}_1 and \mathcal{B}_2 are assumed to be Hausdorff spaces. The topological space \mathcal{B}_1 plays the role of the space \mathcal{E} introduced in Section 1.

If \mathcal{B}_1 is a Hilbert topological space, then it is possible to introduce an energetic inner product in \mathcal{B}_1 with respect to a linear, symmetric, and strongly monotone operator $L : D(L) \subseteq \mathcal{B}_1 \rightarrow \mathcal{B}_1$ as $(Lu|v) \forall u, v \in D(L)$. Hence, we conceptually assume that \mathcal{B}_1 is such that there exists an admissible sequence $u_k \in D(L)$ for all $u \in \mathcal{B}_1$, and the limit $\lim_{k \rightarrow \infty} (Lu_k|v_k)$ does not depend on the chosen admissible sequence [40].

The topological space \mathcal{B}_1 from (2.1) can be viewed as a generalization of such energetic spaces on spaces equipped with the Hausdorff property. If Ω_0^0 is a topological space with a topology π , then in order to specify the model (2.1), (2.2) it is important to have some a-priori information on the possibility of a closure of Ω_0^0 or its subset in some rigorous mathematical sense. If Ω_0^0 is a compact topological space, then equipping Ω_0^0 with the Hausdorff property provides a rigorous mathematical basis for a closure procedure on the basis of the Heine-Borel lemma arguments. In this case Ω_0^0 can be imbedded into a “richer” topological space with the Hausdorff property. Since compactness is an inner and integral property of Hausdorff spaces, it can always be used as a mathematical argument for a closure procedure, no matter how “rich” the actual topology π is, by requiring appropriate regularity assumptions on the mappings (2.1), (2.2). Such assumptions are typically based on the continuity arguments which in many cases allow us to ensure homeomorphic properties of the constructing mappings:

Theorem 2.1 *If there exists a continuous one-to-one mapping of a compact set X into a Hausdorff space Y then such a mapping is a homeomorphism.*

If Ω_0^0 is not a compact set then, in addition to the continuity arguments, one needs an exact specification of a point or a surface in a topological space to ensure that the mathematical model is well-posed. This leads to the main difficulty in gathering a-priori information about the problem itself, for in the general case such a specification is not topology-independent and is an inherently approximating procedure. Therefore, an a-priori equipping of the topology with the Hausdorff property, followed by the imbedding of the obtained Hausdorff space into a “richer” Hausdorff space, cannot ensure that the model with an a-priori given continuous mapping between the two spaces is well-posed. Coupling of the mappings (2.1) and (2.2) precludes a specification of topological properties of the sets Σ and Ω_0^0 with deterministic certainty, as well as with the probability exactly 1. The reverse statement is also true, and unknown topological properties of Σ and Ω_0^0 cause difficulties in the definition of the mappings f and u which are approximate by their nature.

However, the possibility of either of the following two cases always exists

$$\Sigma \subseteq \Omega_0^0 \text{ or } \Omega_0^0 \subseteq \Sigma, \quad (2.3)$$

and this allows an effective construction of models in mathematical applications.

- If $\varpi(\Sigma)$ and $\varpi(\Omega_0^0)$ are powers of these sets, then we have four logical choices to be considered:
- $\varpi(\Sigma) = \varpi(\Omega_0^0)$;
 - $\varpi(\Sigma) \succ \varpi(\Omega_0^0)$;
 - $\varpi(\Sigma) \prec \varpi(\Omega_0^0)$;
 - In neither of these sets there is a part that is equivalent to the other set.

The first case is a consequence of the Cantor-Bernstein theorem, which requires a one-to-one mapping of Σ into a subset of Ω_0^0 , and a one-to-one mapping of Ω_0^0 into a subset of Σ . The last case is usually excluded in the set theory by Zermelo-type arguments, whereas the two remaining cases are typical in the majority of applications. Each specific logical choice provides some *a-priori* constraints on the resulting mathematical models. Such constraints define the limitations of associated computational models from which particular algorithms can be derived (see also [37]).

Rigorously speaking, mappings (2.1) and (2.2) cannot be defined independently of each other for any *a-priori* choice of \mathcal{B}_1 , \mathcal{B}_2 , and the set inclusions (2.3) are responsible for the dependency between their topologies. As the first example, we consider a class of nonsmooth deterministic optimal control problem.

2.1. Deterministic control with *a-priori* given topology and initial conditions.

We restrict the theoretical generality of the problem, by bounded, Lebesgue measurable in time U -valued control functions. We consider the set of all controls that can be defined as $u_{\beta_2}^{\beta_2}(\cdot, \cdot) \in \mathcal{B}_2 \equiv L^\infty([t_0, T]; U)$, where $t_0 \leq t \leq T$ (with the possibility $T \rightarrow \infty$), and U is a subset in \mathcal{B}_2 defined as the control space, whereas β_1, β_2 are *a-priori* specified topologies in $\mathcal{B}_1, \mathcal{B}_2$ respectively. The dot-arguments reflect the dependency of the control on both time and state-space. As a partial case the topological spaces \mathcal{B}_1 and \mathcal{B}_2 may have the arithmetic Euclidean structure, that is

$$\mathcal{B}_1 \equiv R^n, \quad \mathcal{B}_2 \equiv R^m. \quad (2.4)$$

The set of admissible controls as a subset of \mathcal{B}_1 depends on the type of required constraints, and as soon as such constraints are specified we denote the set of all admissible controls by \mathcal{U} .

Given the above assumptions and notations, the classical deterministic problem of optimal control can be formulated in the following way. We assume that the process of evolution of a controlled dynamic system in Ω_0^0 can be appropriately described by a differential equation in terms of a space-time function $x(\cdot)$, the initial conditions and state constraints, that is

$$\frac{dx(t)}{dt} = \tilde{f}(t, x, u_{\beta_1}^{\beta_2}) \text{ a.e. in } T = (t_0, T), \quad \text{for } x \in \mathcal{X} \subseteq \mathcal{B}_1 \quad (2.5)$$

$$x(t_0) = x_0 \in \mathcal{X}, \quad u \in \mathcal{U}. \quad (2.6)$$

Our goal is to minimize the functional

$$J(u_{\beta_1}^{\beta_2}) = \int_{t_0}^T f_0(\tau, x(\tau), u_{\beta_1}^{\beta_2}(\tau, x(\tau))) d\tau + g(T, x(T)) \rightarrow \min \quad (2.7)$$

on the set \mathcal{U} of all admissible controls, where f_0 and g are given functions (the running and terminal costs respectively).

The model (2.5), (2.6), (2.7) may provide an approximation to the coupled mapping (2.1), (2.2). However, additional constraints on the set of admissible controls $\mathcal{U}(\Sigma)$ are required to construct such an approximation. For a given set of initial data (t_0, x_0) , these constraints imply some additional *a-priori* assumptions on the mapping f , and typically require the *a-priori* equipping of the topological product $\Omega_0^0 \otimes \mathcal{U}$ with the Hausdorff property. Then, the chain of logical steps for the model construction is typically based on the assumption of the possibility of the precise definition of initial conditions (2.6) and the existence of a continuous mapping between two subsets in $\Omega_0^0 \otimes \mathcal{U}$, one of which is a compact set allowing the use of the arguments on theorem 2.1.

In many applications these assumptions appear naturally, especially in those problems when the dynamic motion can be relatively easily represented by continuous phase trajectories. For example, this is usually the case when the topological spaces \mathcal{B}_1 and \mathcal{B}_2 are finite dimensional Euclidean spaces defined by (2.4), the set \mathcal{X} is an open set in R^n , and $T \equiv [t_0, t_1]$, $t_0 < t_1 \leq T$. In this case the topological product Ω_0^0 can be defined as $[t_0, t_1] \times \mathcal{X}$, and for finite time problems the closure of Ω_0^0 can be imbedded in a wider set $\bar{\Omega}_0^0 = [t_0, t_1] \times R^n$.

Now an approximation can be specified for the mapping (2.1), (2.2) using the a-priori assumption of its continuity

$$\bar{f} : \bar{\Omega}_0^0 \times U \rightarrow \mathcal{B}_1, \quad (2.8)$$

(for the infinite time horizon problems with $T \rightarrow \infty$ we can use the assumption $\bar{f} \in C(R^n \times U)$). Typically this consideration requires that for all $t \in [t_0, T]$ (or $[t_0, \infty)$) and $x_1, x_2 \in \mathcal{X} \subset R^n$ (or $x_1, x_2 \in R^n$) the following inequality holds,

$$|\bar{f}(t, x_1, u_{\beta_1}^{\beta_2}) - \bar{f}(t, x_2, u_{\beta_1}^{\beta_2})| < \epsilon(\delta)|x_1 - x_2|, \quad (2.9)$$

provided

$$u_{\beta_1}^{\beta_2} \in \mathcal{U} \text{ and } |u_{\beta_1}^{\beta_2}| < \delta. \quad (2.10)$$

Whenever \mathcal{U} is a compact set, (2.9) becomes the Lipschitz-type condition with constant ϵ . The assumption (2.9) is quite restrictive implying a-priori uniqueness for the solution of the following mathematical model:

$$\frac{dx(\tau)}{d\tau} = \bar{f}(\tau, x(\tau), u_{\beta_1}^{\beta_2}(\tau, \cdot)) \quad (2.11)$$

for any given control $u_{\beta_1}^{\beta_2}(\tau, \cdot)$, all $\tau \in [t, T]$, $t \in \bar{T}$ and the initial data $x(t) = x$, where the dot argument reflects the dependency of the control on the neighborhood states when time is fixed.

There are at least two practically important cases when such a consideration is inappropriate:

- a set \mathcal{U} is given approximately and no a-priori information is provided as for its compactness;
- initial conditions (2.6) in the governing equation (2.5) are given approximately.

Both situations are typical in many applications which makes the assumption of continuity of \bar{f} (as well as \bar{f}) inappropriate in general. The current field of optimal control theory essentially relies on the assumption that an admissible control will always remain such under the flow of temporal evolution. This mathematical assumption was induced by the understanding of an optimizer as an error-nulling device and represents a relatively narrow view of the concept of cybernetics [32]. Of course, such a simplified view has allowed the achievement of a number of successes in modelling some aspects of automobile driving, aircraft piloting, aerospace engineering, economics and in some other areas of application. However, the limitations of this view are being reached as the topic of supervisory control is being pursued [16].

Even if we assume the existence of an admissible sequence of controls in \mathcal{B}_1 , it does not necessarily mean that such a sequence is feasible in \mathcal{B}_2 . On the other hand, a possible continuity of $x(\cdot)$ does not implies continuity of $\bar{f}(t, x, u_{\beta_1}^{\beta_2})$ (and as a partial case the continuity of \bar{f} defined by (2.8)-(2.11)) and vice versa.

This leads to the major difficulty which is the construction of a mapping between x and \bar{f} (\bar{f}) that provides an appropriate approximation to the coupled mapping (2.1), (2.2). This difficulty is twofold. On the one hand, the well-posedness of the Cauchy problem (2.5), (2.6) is a highly desirable feature, not only from the mathematical point of view, but also from the point of view of physics [25]. On the other hand, in constructing mathematical models it is important to take into account that the assumptions of continuity together with the possibility of equipping of the topological space \mathcal{B}_1 with the Hausdorff property are not independent of the result of such a construction.

A consequence of an approximation of the coupled mapping (2.1), (2.2) by mappings \bar{f} (\bar{f}) and $u_{\beta_1}^{\beta_2}$ is that regularities of these mappings become coupled through the model stability. Certain a-priori regularities on $x(\cdot)$ imply a stabilizing trade-off between regularities of \bar{f} (\bar{f}) and $u_{\beta_1}^{\beta_2}$. Conversely,

a-priori regularities on \tilde{f} (\bar{f}) require a trade-off between the smoothness of x and $u_{\beta_1}^{\beta_2}$. This reflects the recursive structure of dynamic mappings that eventually leads to the question on the singular nature of the problem of initial conditions [11]. Mathematically, this can be formalized by the choice of one of the two possibilities in (2.3), and such a formalization allows the mathematical closure of either Ω_0^0 or Σ (or both). If the topology of \mathcal{X} is a-priori chosen, then the quality of approximation of the coupled mapping (2.1), (2.2) by the deterministic model (2.5), (2.6), (2.7) is decisively dependent on the choice of the sets \mathcal{U} and \mathcal{T} .

Let us assume the Lebesgue measurability of the approximating functions $u_{\beta_1}^{\beta_2} \in L^1(\Sigma)$ and $\bar{f} \in L^1(\Omega_0^0 \times \mathcal{U})$ in the deterministic model (2.1), (2.2), (2.3). Then the governing equation (2.5) can be interpreted as an integral equation using a consequence of the Chebyshev inequality [17]

$$x(t) - x(t_0) = \int_{t_0}^t \tilde{f}(\tau, x(\tau), u_{\beta_1}^{\beta_2}(\tau, \cdot)) d\tau + \int_{t_0}^t |\bar{\sigma}(\tau, x(\tau), u_{\beta_1}^{\beta_2}(\tau, \cdot))| d\tau, \quad (2.12)$$

where

$$\int_{t_0}^T |\bar{\sigma}(\tau, x(\tau), u_{\beta_1}^{\beta_2}(\tau, \cdot))| d\tau = 0. \quad (2.13)$$

The key a-priori assumption in the approximation of (2.1) and (2.2) by (2.12) and (2.13) is the possibility of equipping of the topological space \mathcal{X} with the Hausdorff property⁶. Since \mathcal{T} is also considered a Hausdorff topological space, the topological product Ω_0^0 of \mathcal{T} and \mathcal{X} inherits the Hausdorff property.

Under these assumptions let us consider two sets $X_1^{\mathcal{T}}$ and $X_2^{\mathcal{T}}$ in Ω_0^0 . Denote $\Upsilon_\epsilon(X_i^{\mathcal{T}})$ as an ϵ -neighborhood of $X_i^{\mathcal{T}}$, $i = 1, 2$. Then the Hausdorff separation of the set $X_1^{\mathcal{T}} \subset \Omega_0^0$ from $X_2^{\mathcal{T}} \subset \Omega_0^0$ is defined as [3]

$$Sep(X_1^{\mathcal{T}}, X_2^{\mathcal{T}}) = \inf\{\epsilon : X_1^{\mathcal{T}} \subseteq \Upsilon_\epsilon(X_2^{\mathcal{T}})\}.$$

If $\Upsilon_i(\tau_i) \subset \Omega_0^0$, $i = 1, 2$ are open neighborhood sets of all admissible states under control actions $u_{\beta_1}^{\beta_2}(\tau_i, \cdot) \in \mathcal{U}$, $i = 1, 2$, then it is assumed that $\forall \epsilon > 0$ in the state-space of the system we have

$$Sep(\Upsilon_1(\tau_1), \Upsilon_2(\tau_2)) < \epsilon \quad (2.14)$$

provided there exists $\delta(\epsilon, \Upsilon_1, \Upsilon_2) > 0$ such that

$$|\tau_1 - \tau_2| < \delta. \quad (2.15)$$

The conditions (2.14), (2.15) store the connection between the sets \mathcal{X} and \mathcal{T} through the intermediate influence of the set \mathcal{U} . This may not be explicitly mentioned when the Heine-Borel lemma arguments are applied to the set defined by δ in (2.15). Indeed, such arguments allow us to claim the existence of δ for any arbitrary given ϵ . However, they cannot guarantee stability of the resulting mathematical model for the described dynamic system. Stability conditions express a local dependency between Υ_i , $i = 1, 2$ and ϵ on the global scale of Ω_0^0 . Whenever the state-space of a dynamic system is equipped with the Hausdorff property such conditions are “frozen” in δ mathematically defined by (2.15). After the model has been constructed the question of its stability can be approached in the general case through the investigation of the dependency between β_1 , β_2 and t_0 . The complexity of this question has led to the attempts of relaxation of topological constraints on the initial data. To a large degree these attempts are connected to the development of stochastic mathematical models.

2.2. Hamiltonian approximations in deterministic and stochastic models of optimal control.

During the last decade a general framework based on an intrinsic connection between nonlinear partial differential equations in infinite dimensional spaces and optimal control in Hilbert spaces has

⁶in this case we will write $\Omega_0^0 \otimes \mathcal{U} \equiv \Omega_0^0 \times \mathcal{U}$

been extensively developed (see, for example, [27] and references therein). The development of this new approach is essentially being motivated by a desire to develop a rigorous mathematical basis for applications of the dynamic programming method (DPM) in optimal control problems. The main difficulty in practical applications of DPM stem from *a-priori* regularity assumptions on the value function⁷ that is not known from the control problem itself. On the other hand, the other fundamental approach in optimal control, the maximum principle method (MPM), encounters serious difficulties when steep space-time gradients occur. This produces a challenging problem on a connection between MDP and MPM in the general non-smooth case.

Let us recall the main steps in the application of DPM to deterministic models (2.5)-(2.7). We restrict terminal behaviour of a system by a target set \mathcal{K} assuming that $x(T) \in \mathcal{K}$ and at this stage limiting our consideration to arithmetic Euclidean spaces (2.4). When the initial conditions (t, x) and the target set are specified we define the performance measure as a time-averaging functional [15]

$$J(t, x; u_{\beta_1}^{\beta_2}) = \bar{g}(\tau, x(\tau)) + \int_t^\tau f_0(s, x(s), u_{\beta_1}^{\beta_2}(s, \cdot)) ds \quad (2.16)$$

where x is the solution of (2.5) for the specified initial conditions (t, x) , $x \in \mathcal{K}$, and τ is the “exit” time of $(s, x(s))$ from a closure of Ω_0^0 . The function \bar{g} is defined as

$$\bar{g}(t, x) = \begin{cases} g(t, x) & \text{when } (t, x) \in [t_0, T) \times \mathcal{X} \\ \hat{g}(x) & \text{when } (t, x) \in T \times \mathcal{X} \end{cases}$$

(we can formally set $\bar{g} = \infty$ whenever x does not belong \mathcal{K} [8]). This allows us to consider a family of optimization problems with different initial conditions (t, x) if we define the value function as the greatest lower bound of the functional (2.16) on the set of all admissible controls

$$V(t, x) = \inf_{u_{\beta_1}^{\beta_2} \in \mathcal{U}} J(t, x; u_{\beta_1}^{\beta_2}) \quad (2.17)$$

for all $(t, x) \in \bar{\Omega}_0^0$. If it is *a-priori* assumed that $V(t, x) > -\infty$ [7] and that the set \mathcal{U} is nonempty, then the basis for applying a DP approach is provided by the Bellman dynamic programming principle.

Theorem 2.2 [7] *For arbitrary initial conditions $(t, x) \in \bar{\Omega}_0^0$ and $r \in [t, T]$ the following formula holds*

$$V(t, x) = \inf_{u_{\beta_1}^{\beta_2} \in \mathcal{U}} \left[\int_t^{\min\{r, \tau\}} f_0(s, x(s), u_{\beta_1}^{\beta_2}) ds + \bar{g}(\tau, x(\tau)) \chi(\tau; r) + V(r, x(r)) \chi(r; \tau) \right], \quad (2.18)$$

where the set-characteristic function χ is defined as

$$\chi(x_1; x_2) = \begin{cases} 1 & \text{if } x_1 < x_2 \\ 0 & \text{otherwise} \end{cases}$$

Since the topological space \mathcal{B}_1 is a partially ordered set let us choose a chain⁸ \mathcal{M} in it. This chain is contained in a maximal chain of \mathcal{B}_1 (due to the Hausdorff theorem). In the general case it is not required that \mathcal{M} has to have the least upper bound unless \mathcal{B}_1 is a structure⁹) [17]. Hence in principle, any mapping defined with respect to the chain \mathcal{M} may only provide an approximation of a mapping

⁷even in linear cases this function is not necessarily differentiable

⁸a subset in which any two elements are comparable in the sense of an *a-priori* introduced partial ordering

⁹often referred to as a lattice, i.e. a partially ordered set for which any of its finite subsets has the greatest lower and the least upper bounds

defined with respect to the whole topological space \mathcal{B}_1 . If we assume that \mathcal{B}_1 is a lattice, then the quality of such an approximation influences the stability properties of the lattice as a whole, and this gives rise to difficulties in the definition of the least upper bound for the chain \mathcal{M} provided it exists.

In some cases such difficulties can be removed using the classical mechanical analogy related to the quantity which is a constant along the phase trajectory (x, p) of a system, where variables x and p denote space positions and generalized impulses of the system respectively. If we assume that such a quantity does not depend on time then it can be defined as a first integral of the Hamiltonian canonical system

$$\frac{\partial x}{\partial t} = \frac{\partial H}{\partial p}, \quad \frac{\partial p}{\partial t} = -\frac{\partial H}{\partial x}, \quad (2.19)$$

where $H(x, p)$ is the system Hamiltonian which provides a mathematical generalization of the mapping defined by (1.3). The other quantity of interest is the system Lagrangian L that allows us to relate time-averaging functionals to a canonical system (2.19). In the classical case this relationship is provided by the formula $p = \partial L / (\partial x \partial t)$.

In general, both quantities H and L are time dependent. When the motion of the system can be idealized as a motion of a material point along a phase space trajectory, the dynamics of the system can be described by the Hamiltonian principle of the least motion, allowing us to apply techniques of the calculus of variations. In classical mechanical applications, the definition of the quantity $H + L$ is straightforward. For example, for the motion in R^n this quantity is defined by the inner product between velocities and impulses

$$H + L = (\dot{x}|p) = \sum_{i=1}^n \dot{x}_i p_i, \quad (2.20)$$

and is easily generalizable on Hilbert spaces. This definition provides an appropriate approximation to the total energy of the system in many different applications of dynamic systems.

In attempts to generalize (2.20) on a wider class of dynamic systems one encounters two difficulties. Firstly, in the general case it may be inappropriate to define the Lagrangian as a function of only $(t, x(t), \dot{x}(t))$. Secondly, if the Lagrangian is taken as an integrand in the optimizing functional of an optimal control problem, then one needs to specify the topology of a set that may be embedded in a lattice produced by \mathcal{B}_1 . This set may play the role of the value domain of the system Hamiltonian.

Since this set is not required to be countable, the usual induction arguments may fail. This requires appealing to the Zorn lemma [40] under the a-priori assumption that such set can be partially ordered. Of course, the possible existence of the least upper bound of the set does not follow from such a consideration. However, if we assume that the value domain of the Hamiltonian is a Hausdorff topological space, it allows us to approximate the coupled mapping (2.1), (2.2) by the pair of "weakly" coupled mappings that define the Hamiltonian and the value function respectively:

$$H : \tilde{\Omega}_0^{0,1} \rightarrow \mathcal{B}_1 \quad (2.21)$$

and

$$V : \tilde{\Sigma} \rightarrow \mathcal{B}_2. \quad (2.22)$$

The quality of such an approximation decisively depends on an appropriate choice of the Hamiltonian and the specification of V on a subset of $\Omega_0^0 \times \mathcal{U}$, and eventually on the quality of the approximation of the initial conditions for the governing equation for the original optimal control problem. The point (t_0, x_0) , defined by the initial conditions of the model, may be a unique point of tangency of the sets $\tilde{\Sigma}$ and $\tilde{\Omega}_0^{0,1}$. In this case the oscillation of the value function on the set \mathcal{U} does not have to be defined explicitly¹⁰, and one can assume that the function V is merely a semi-continuous function assuming its equality to $+\infty$ at all points from which the target \mathcal{K} is unreachable [8,36].

¹⁰if a function \tilde{f} has infimum m_0 and supremum M_0 on a set X , then the difference $M_0 - m_0$ is called the oscillation of \tilde{f} on X

In principle, the application of this idea is virtually unlimited [36], provided the Hamiltonian of the system is defined. However, the dilemma of this approach consists in the fact that any *particular* specification of the Hamiltonian automatically imposes an *upper bound on the oscillation* of the function V on the set \mathcal{U} . Of course, in many practical applications this dilemma can be easily resolved when the evolution of a dynamic system can be idealized by the analogy with classical mechanics. As a result, this approach leads to the construction of mathematical models that give appropriate approximations of the system. For example, under quite general topological assumptions we can introduce the quantity

$$H(t, x, p) = \sup_{u_{\beta_1}^{\beta_2} \in \mathcal{U}} \{-p \cdot \tilde{f}(t, x, u_{\beta_1}^{\beta_2}) - f_0(t, x, u_{\beta_1}^{\beta_2})\} \quad (2.23)$$

that is called the Hamiltonian [7] using classical mechanics similarities. This allows us to use Theorem 2.1 to construct the model

$$\frac{\partial V}{\partial t} = H(t, x, D_x V), \quad V(T, \cdot) = \tilde{g}(\cdot) \quad (2.24)$$

known as the deterministic Hamilton-Jacobi-Bellman (HJB) equation, where in the general case $D_x V$ denotes the generalized gradient of the function V defined by (2.17). As in the original problem of optimal control the model (2.24) provides an approximation to the coupled mapping (2.1), (2.2) on the basis of a *competition* between the topological completeness of the domain of the definition of H and a-priori regularity assumptions on V . If the domain of definition of H is a-priori equipped with the Hausdorff property, then the quality of such an approximation depends decisively on the quality of the *approximation* of the function V on a subset of $\Omega_0^0 \times \mathcal{U}$. One of the possibilities to improve such an approximation is connected with stochastic mathematical models in which *careful examination of the neighborhood states of motion* is required. This requirement always competes with the requirement of conservation extension in phase postulated in statistical mechanics.

In the case when dynamics can be reasonably well described by the model (2.5)-(2.7), the main difficulty stems from a connection between $u_{\beta_1}^{\beta_2}$ and T expressed by the identity (2.13). Since a function $\tilde{\sigma}$ is unknown, many important results in the deterministic theory of optimal control have been obtained for the limiting case of $T \rightarrow \infty$. The other possible approach to this problem is based on Tichonov type models with singular perturbations, which require the investigation of dynamics in the limit of vanishing perturbations, $\epsilon \rightarrow 0^+$. The models, resulting from both approaches, are natural developments of the model (1.1), (1.2), which in the view of optimal control theory require information on a connection between $u_{\beta_1}^{\beta_2}$ and T .

Attempts to capture this connection in its generality lead to the limiting mathematical models for $T \rightarrow \infty$ simultaneously with $\epsilon \rightarrow 0^+$. However, each model of this type implies a-priori knowledge on a connection between topological properties of \mathcal{B}_1 and \mathcal{B}_2 (or $\tilde{\mathcal{B}}_1$ and $\tilde{\mathcal{B}}_2$), their subsets, or their approximations by other topological spaces. If such information is available, then the results of the Ljusternik-Schnirelman theory can be used to draw a conclusion on sets of critical points of functionals that define the mathematical model itself (see, for example, [36] and references therein). In the general case, the problem lies in the fact that the definition of topologies in \mathcal{B}_1 and \mathcal{B}_2 is not independent of the definition of the functionals that define the model itself. An approximate character of their connection is captured in evolutionary mathematical models that are obtained from the formal definition of the coupled mapping (2.1) and (2.2).

Which of the approximate models should be used is determined by the specific area of model application. For example, although neither $\tilde{\sigma}$ nor $x(\cdot)$ in (2.12) are known a-priori, and in the general case equality (2.13) is one of our a-priori assumptions, these assumptions can be reasonably well justified for a large number of dynamic systems. The development of physics beyond the scope of classical mechanics led to the necessity to analyze the functional $\int_{t_0}^t \tilde{\sigma}$ in (2.12) and consequently to relax the requirement (2.13). A fundamental idea for such a relaxation that allows us to approximate (2.1), (2.2) is connected with the following stochastic model with random coefficients

$$x(t) - x(t_0) = \int_{t_0}^t \tilde{f}(\tau, x(\tau), u_{\beta_1}^{\beta_2}(\tau, \cdot)) d\tau + \int_{t_0}^t \tilde{\sigma}(\tau, x, u_{\beta_1}^{\beta_2}(\tau, \cdot)) d\omega(\tau). \quad (2.25)$$

In spite of the general nature of this approximation, its quality in applications is determined by the regularity assumptions on $x(\cdot)$ and $\omega(\cdot)$. For example, we may assume that $x(\cdot)$ is a continuous sample path R^n -valued process and $\omega(\cdot)$ is a \mathcal{F}_t -adapted R^k -valued Wiener process on a probability space (Ω, \mathcal{F}, P) (the family of σ -algebra $\mathcal{F}, t \geq 0$ defines a filtration on this space). Then the problem of stochastic optimal control of the dynamics described by the governing equation (2.25) is to minimize the conditional mathematical expectation

$$E_{x,t}^{u_{\beta_1}^{\beta_2}}[J(u_{\beta_1}^{\beta_2})] \rightarrow \min, \quad (2.26)$$

where the functional $J(u_{\beta_1}^{\beta_2})$ is defined by (2.16).

As for the deterministic models, we require information on the connection between $u_{\beta_1}^{\beta_2}$ and T , which always comes implicitly from the a-priori regularity assumptions on the mappings which define the model. In principle, the situation is the same, when we extend the model (2.26) to a model with jumps by adding a “jump-term” to (2.25) defined by \mathcal{F}_t -Poisson random measures (see, for example, [19] and references therein).

If the process described by (2.25) and (2.26) is assumed to be a controlled (Markov) diffusion, then its dynamic programming equation (2.18) is reducible to the HJB equation, and this takes the form of the second order PDE

$$\frac{\partial V}{\partial x} = H(t, x, D_x V, D_x^2 V), \quad V(T, \cdot) = \bar{g}(\cdot) \quad (2.27)$$

under appropriate regularity assumptions (see [7] for details), where

$$H(t, x, p, Q) = \sup_{u_{\beta_1}^{\beta_2} \in \mathcal{U}} \left[-p \cdot \bar{f}(t, x, u_{\beta_1}^{\beta_2}) - \frac{1}{2} \sum_{i,j=1}^n (\bar{\sigma} \bar{\sigma}')_{ij} q_{ij} - f_0(t, x, u_{\beta_1}^{\beta_2}) \right], \quad (2.28)$$

and $Q = (q_{ij})$ is a symmetric, non-negative definite matrix. A procedure for the reduction of (2.25) and (2.26) to (2.27) and (2.28) preserves a connection between \bar{f} , $\bar{\sigma}$, and ω through topological properties of their functional space definition. From the statistical point of view, such a procedure is a construction of a canonical averaging ensemble which requires consistency with postulates of statistical mechanics. In general, such procedures cannot guarantee continuity of $x(\cdot)$, and thus, require the consideration of *singular stochastic control* models. Moreover, even if the original process is Markovian, it does not necessarily imply that it is a continuous process with transition probabilities which can be approximated arbitrarily well by a diffusion processes. However, in many practical applications such an approximation allows us to catch some important aspects of dynamic systems.

From a physical point of view, the necessity of relaxing requirement (2.13) has been strongly motivated by Eddington's idea of the “time arrow” [20]. In addition, such a relaxation has mathematical and technical convenience in its favour.

We consider a class of relaxed systems by considering the set Λ of all Borel measures on $\mathcal{T} \times \mathcal{X} \times \mathcal{U} \subset \Omega_0^0 \otimes \mathcal{U}$ such that

$$\lambda([0, s] \times \mathcal{X} \times \mathcal{U}) = s, \quad 0 \leq s < T \quad (2.29)$$

under the assumption that $\mathcal{T} \equiv [0, T]$ (or $\mathcal{T} = [0, \infty)$). The property (2.29) (the “fixed-time control-iteration”) allows us to construct mathematical models which, in principle, are reversible in time. This property subtends the Hausdorff property of an approximation to $\Omega_0^0 \otimes \mathcal{U}$. Then applying the Prochorov metric one can show that Λ is a compact metric space [27]. If we consider the Borel field \mathcal{B} on \mathcal{U} , and the σ -field $\sigma_t(\Lambda)$ generated by $\{\lambda([0, s] \times \mathcal{X} \times \mathcal{U}_0), s \leq t, \mathcal{U}_0 \in \mathcal{B}\}$ and finally introduce the σ -field generated by $\sigma_t(\Lambda)$, $t > 0$, then the space $\mathcal{P} = \mathcal{P}(\Lambda)$ of all probabilities on $(\Lambda, \sigma(\Lambda))$ (with weak convergence topology) is also a compact metric space due the Prochorov theorem. Now if $\omega(0) = 0$, μ is an \mathcal{F}_t -adapted Λ -valued random variable (relaxed control)¹¹ and $(\Omega, \mathcal{F}, \mathcal{F}_t, \mathcal{P})$ is a

¹¹for \mathcal{F}_t measurability of $\mu(X \times Y)$ it is required only that $X \in \mathcal{B}([0, t] \times \mathcal{X})$ and $Y \in \mathcal{B}(\mathcal{U})$

probability space, then any system of the form $(\Omega, \mathcal{F}, \mathcal{F}_t, \mathcal{P}, \omega, \mu)$ is a relaxed system. Since conditional mathematical expectation can be defined using progressively measurable Radon-Nikodim derivatives, appealing to the Radom-Nikodim theorem provides a way to a justification of the mathematical model (2.25), (2.26) for relaxed controlled systems.

Such models are less sensitive to a topological connection between the spaces \mathcal{B}_1 and \mathcal{B}_2 because for both spaces, the possibility of embedding into a Hausdorff topological space with the property (2.29) is assumed a-priori. This connection is still preserved through the definition of the process $\omega(\cdot)$ and the specification of the initial conditions of the problem. If the process $\omega(\cdot)$ is assumed to be continuous, such specification provides an *important class of limiting evolutionary models* with vanishing drift and normalized diffusion. This possibility follows from the second case in the comparison of the sets Σ and Ω_0^0 in (2.3).

The other important limiting case of a mathematical model for evolution is provided by the classical dynamical systems where we assume the existence of an evolutionary process with vanishing "diffusion" and a normalized "drift"

$$\dot{x} = f(t, x), \quad (2.30)$$

which is formally reducible to the autonomous system $\dot{x} = f(x^{n+1}, x)$, $\dot{x}^{n+1} = 1$ by a change of variables $x^{n+1} = t$ [30], the analogue of which for stochastic mathematical models, takes the form (2.29) [27]. The possibility of approximations based on the models (2.30) follows from the third case in comparison of the sets Σ and Ω_0^0 in (2.3). The first case in (2.3) corresponds to the situation when both processes $x(\cdot)$ and $\omega(\cdot)$ are continuous.

2.3. Coupling modes in mathematical models for evolution.

All three types of models considered in this section¹² assume the a-priori possibility of the regularization of mathematical models when the topological space (that has to approximate space-time of the dynamic system evolution) is equipped with the Hausdorff property. This consideration relies on the possibility of knowing the initial conditions for evolutionary mathematical models exactly. In turn, this transfers the complexity issues related to the definition of a balance between "drift" and "diffusion" components in the governing dynamic equation to the computational level where these issues are typically addressed.

An approximating mapping (possibly expressed on a finite lattice) between Ω_0^0 and Σ , which approximates the problem (2.1) and (2.2), implies that the control depends on both time and neighborhood states. Hence, whenever we assume the Hausdorff property of state-space and allow the control to be a discontinuous function a *sequential regularization procedure* is necessary to ensure the model stability. The complexity of this regularization has two sides.

- On the one hand, stochastic processes are not necessarily diffusion-dominated, and the problem of dynamic system control described by stochastic mathematical models may not necessarily be reduced to either elliptic or parabolic (possibly degenerate) PDEs. In the general case, the method of diffusion approximation may not provide an appropriate framework for the approximation of the original problem. This is shown by the growing number of physical examples (see [25,13] and references therein). Allowing the process $x(\cdot)$ to be a discontinuous function in stochastic singular control problems and assuming the control function to be from the class of functions with bounded variation, the regularity of the value function is virtually determined by the regularity of the interface, which is unknown from the control problem itself. This coupled regularity problem which in some cases can be resolved through different versions of the principle of smooth fit [2,34,7] or by the compactness arguments [18] in general, is problem specific. This leads to the situation when the quality of the mathematical model is completely defined by its consistency to the real-world phenomenon.
- On the other hand, deterministic processes may not necessarily be convection-dominated or purely autonomous (or such that they exhibit a strictly periodic behavior), and the problem of dynamic system control described by deterministic mathematical models may not necessarily be reduced to purely

¹²based on the first three logical choices in comparison of Σ and Ω_0^0

hyperbolic PDEs. Taking the limit of small diffusivity in the reaction-diffusion equation, exponential ill-conditioning may manifest itself. In this case, a numerical treatment of possible discontinuities in the density function is necessary [31].

Such competition between deterministic and stochastic features of mathematical models is essentially part of their effectiveness in solving real-world problems. This requires the study of phenomena connected with possible discontinuities of the density function in deterministic mathematical models as well as possible discontinuities in the function $x(\cdot)$ in stochastic mathematical models. An efficient way to study such phenomena can be provided by the analysis of the value function for the processes described by such mathematical models. The regularity of this function can be implemented in the model by the problem-specific information rather than by a-priori arguments of continuity.

3 Coupling regularities of adjoint function and control.

On the whole, the process of validation of mathematical models is defined by the possible existence of a separation principle between control issues of dynamics and statistical estimation aspects related to this dynamics. Such validation is based on some a-priori smoothness assumptions on functions that are not known from the mathematical model itself. Of course, in many important cases these assumptions can be justified making the model consistent with a certain (possibly very large) range of applications. Moreover, it is often possible to establish a scale of a-priori estimates that control effectiveness of the model applicability in practical problems (see, for example, [22] and references therein).

3.1. Control, boundary and mesh refinement.

However there are many other situations when regularities of unknown functions can not easily be predicted, and a natural experiment can not be easily undertaken. A control problem arises naturally from such situations, and yet the importance in the real application of controlling the boundary to obtain the physically stable situation results in the main difficulty in a rigorous mathematical formalization of such a problem. This gives rise to two main approaches to the control problems.

- In practice such control can be conducted effectively by minimizing the difference between the computed and partially-observed state variables that allow us in many cases to estimate values of those variables at other points as well as to approximately determine physical parameters that may not be distinct in observations [14].

- Alternatively, we can assume that in theory any region of interest in a specific problem can be arbitrary densely discretized. However, for a specific mathematical model such fine properties of the mesh may result in a discretized model which may not only be computationally inefficient but may also be practically infeasible.

Hence, using either of these approaches, it is crucial for the algorithm performance to predict the locations of the regions with large space-gradient. This requires some reasonable a-priori assumptions about the smoothness of unknown functions that eventually leads to the validation of the mathematical model on a class of real-world problems. Although it is often the case that such assumptions can be relatively easily justified for stationary models, in the time dependent problems the possible change of shapes and location of large-gradient regions requires an increase attention to the validation procedure for the mathematical model. The topology of the space becomes dependent on a-posteriori error estimations [12]. This in turn requires an adaptive mesh refinement using some a-priori defined indicators [28].

The definition of such indicators is a necessary step in the construction of the mathematical model itself and this gives rise to difficulties in approximating the system Hamiltonian. When the model has been constructed, an approximate character of the Hamiltonian is still reflected in the coupling regularities of control and value functions. In other words, we approximate the coupled mapping (2.1), (2.2) by a mathematical model that couples three main functions - control, value function, and the

system Hamiltonian. If the mathematical formalization of the Hamiltonian has been made, the key problem to be solved is the problem of coupling regularities of the control and value functions.

3.2. The Pontryagin maximum principle and normalizations of the Hamiltonian.

In the deterministic theory of optimal control the Pontryagin maximum principle (PMP) allows us to solve a wide range of important practical problems [30]. However, there are difficulties in the application of this principle when the value function has steep space-gradients, especially when their locations are not known a-priori. These difficulties can be overcome if we assume continuity (or at least semi-continuity) of the value function. This assumption allows us to reduce the original control problem to the problem of solving the associated HJB equation which follows from the Dynamic Programming Principle (DPP). When the value function has classical smoothness $V(t, x) \in C^{1,2}(\Omega_0^0)$ both approaches are connected by the method of characteristics which implies that

$$\psi(t) = \frac{\partial V}{\partial x}(t, x^*(t)), \quad (3.1)$$

where ψ is the adjoint function and x^* is the optimal trajectory from the optimal solution pair $(x^*, \{u_{\beta_1}^{\beta_2}\}^*)$.

If the model is stochastic, then for a diffusion approximation the dynamic programming formalism can still be put on a rigorous mathematical basis using viscosity solution theory. A non-smooth analogue of a connection between the PMP and DPP (3.1) involving the adjoint and value functions respectively, can be expressed in the form of embedding theorems.

Theorem 3.1 [41,42] *If $(x^*, \{u_{\beta_1}^{\beta_2}\}^*)$ is an optimal process with the Pontryagin adjoint function ψ , and $V(t, x)$ is a viscosity solution of the problem (2.24) (or (2.27) when the function $\bar{\sigma}$ in (2.25) does not depend on the control explicitly and $\omega(\cdot)$ is a Brownian motion¹³, then the inclusion*

$$D_x^- V(t, x^*(t)) \subset \{\psi(t, \cdot)\} \subset D_x^+ V(t, x^*(t)), \quad (3.2)$$

holds for any $t \in [s, T]$, $t_0 < s < T$, where $D_x^+ V$ and $D_x^- V$ denote the superdifferential and subdifferential of V respectively.

Since the Hamiltonian is dependent on the adjoint process it is important to know how the adjoint function, the Hamiltonian, and the value function are connected. In the deterministic case such a connection is expressed by the inclusion [41]

$$D_{t,x}^{1,-} V(t, x^*(t)) \subset \{H(t, x^*, \{u_{\beta_1}^{\beta_2}\}^*(t, \cdot), \psi(t, \cdot)), \psi(t, \cdot)\} \subset D_{t,x}^{1,+} V(t, x^*(t)) \quad (3.3)$$

which holds a.e. for $t \in [s, T]$.

In the stochastic case the embedding (3.3) may be violated, so this requires more careful examination of the dependency between the Hamiltonian and the adjoint function. The importance of such an examination lies with the fact that in using the DPP approach and specifying the form of the system Hamiltonian a-priori, the decisive role in the justification of resulting mathematical models belongs to the algorithm for the solution of the associated dynamic equation of the Hamilton-Jacobi-Bellman type [5,6,23,29]. The choice of an approximation to the Hamiltonian eventually defines the choice of the adjoint function and vice versa, and thus the regularities of the value function become a reflection of a connection between the Hamiltonian approximation and the adjoint process.

Let us analyze this connection with the example of deterministic models of optimal control (2.5)-(2.7) where in the classical smooth case the connection between PMP and DPP may be given by the formula (3.1). In the classical deterministic case the adjoint function $\psi(t, \cdot)$ is defined by the differential equation [30]

$$\frac{\partial \psi}{\partial t} = -\frac{\partial f_0}{\partial x} - \frac{\partial \tilde{f}}{\partial x} \psi \quad (3.4)$$

¹³in this case (3.2) holds with the probability 1

with the Cauchy type terminal condition

$$\psi(T, \cdot) = \tilde{g}(\cdot), \quad (3.5)$$

where $\tilde{g}(\cdot) = 0$ or $\psi(T, \cdot) = \partial\tilde{g}(T, x(T))/\partial x$ depending on the type of the problem we consider. In the general case, similar to the control function the adjoint function depends not only on time but on the topology of the state-space as well. Having defined the adjoint function, the definition of the Hamiltonian is typically assumed to be of the form [35,15]

$$H = f_0 + \tilde{f}\psi. \quad (3.6)$$

From the mathematical theory of optimal processes developed on the basis of the PMP [30], it follows that, for the process $(x(t), u_{\beta_1}^{\beta_2}(t, \cdot))$ to be optimal, it is necessary for an adjoint function (which is not identically zero) to exist such that the relationship

$$\min_{u_{\beta_1}^{\beta_2} \in U} H(t, x(t), u_{\beta_1}^{\beta_2}(t, \cdot), \psi(t, \cdot)) = H^*(t, x(t), \{u_{\beta_1}^{\beta_2}\}^*(t, \cdot), \psi(t, \cdot)) \quad (3.7)$$

holds a.e. in $[t_0, T]$.

On the other hand we can use the optimality principle of the Dynamic programming approach based on theorem 2.2 to obtain

$$V(t, x(t)) = \min_{u_{\beta_1}^{\beta_2}(\tau, \cdot): t \leq \tau \leq t + \Delta t} \left\{ \int_t^{t + \Delta t} f_0[\tau, x(\tau), u_{\beta_1}^{\beta_2}(\tau, \cdot)] d\tau + V(t + \Delta t, x(t + \Delta t)) \right\}. \quad (3.8)$$

In (3.8) the minimum-cost value function V is defined similarly to (2.17), that is as the minimum of the performance measure

$$V(t, x(t)) = \min_{u(\tau, \cdot): t \leq \tau \leq T} J(t, x; u_{\beta_1}^{\beta_2})$$

where $J(t, x; u_{\beta_1}^{\beta_2})$ in its turn is defined as the performance measure (2.16) changing s for τ and τ for T

$$J(t, x; u_{\beta_1}^{\beta_2}) = \tilde{g}(T, x(T)) + \int_t^T f_0(\tau, x(\tau), u_{\beta_1}^{\beta_2}(\tau, \cdot)) d\tau.$$

Then by the standard procedure from (3.8) we derive the HJB equation in the classical form (2.24), i.e. we have

$$0 = \frac{\partial V}{\partial t}(t, x(t)) + H^*\left(t, x(t), \{u_{\beta_1}^{\beta_2}\}^*, \frac{\partial V}{\partial x}\right), \quad (3.9)$$

where

$$V(T, x(T)) = \tilde{g}(T, x(T)), \quad (3.10)$$

and

$$H^*\left(t, x(t), \{u_{\beta_1}^{\beta_2}\}^*, \frac{\partial V}{\partial x}\right) = \min_{u_{\beta_1}^{\beta_2}(t, \cdot) \in U} H\left(t, x(t), u_{\beta_1}^{\beta_2}(t, \cdot), \frac{\partial V}{\partial x}\right) \quad (3.11)$$

whereas the Hamiltonian is defined as

$$H\left(t, x(t), u_{\beta_1}^{\beta_2}(t, \cdot), \frac{\partial V}{\partial x}\right) = f_0(t, x(t), u_{\beta_1}^{\beta_2}(t, \cdot)) + f(t, x(t), u_{\beta_1}^{\beta_2}(t, \cdot)) \frac{\partial V}{\partial x}. \quad (3.12)$$

Since PMP (3.7) gives only necessary conditions for optimality we cannot guarantee that the adjoint function ψ that satisfies (3.7) and is defined by the model (3.4), (3.5) is unique. This implies that in the general case the validity of passage from (3.4), (3.5) to (3.9)-(3.12) depends *decisively*

on the definition of the Hamiltonian. However, there is a striking difference between the definition of the Hamiltonian by the formulas (3.6), (3.12) and the definition (2.20) in classical mechanics. If the function f_0 defines the system Lagrangian¹⁴, then the formulas (3.6), (3.12) can be seen as those that “normalize” the quantity $H - L$ whereas the formula (2.20) “normalizes” the quantity $H + L$. Strictly speaking, it is not clear from *a-priori* reasoning that the coefficient near the function f_0 given *a-priori* in (3.6) and (3.12) should be equal to 1. It is this assumption that allows us to prove the uniqueness results for the value function in the types of models (2.24), (2.27) under relaxed smoothness assumptions (see [7] and references therein).

To explain this phenomenon let us substitute the Hamiltonian (3.6) into the second equation of the set of canonical equations (2.19) which in the theory of optimal control are known as the Euler-Hamilton canonical equations [30]. We assume that the initial condition $x(t_0) = x_0$ may be given precisely for the model

$$\frac{\partial x}{\partial t} = \frac{\partial H}{\partial \psi}, \quad (3.13)$$

and that the terminal condition (3.5) for the adjoint model

$$\frac{\partial \psi}{\partial t} = -\frac{\partial H}{\partial x} \quad (3.14)$$

may also be given precisely. Then we have the following relationship between f_0 , ψ and \tilde{f}

$$\frac{\partial \psi}{\partial t} = -\frac{\partial f_0}{\partial x} - \psi \frac{\partial \tilde{f}}{\partial x} - \tilde{f} \frac{\partial \psi}{\partial x}. \quad (3.15)$$

From the comparison (3.4) and (3.15) we conclude that if the Hamiltonian is defined with the normalizing factor 1 near the Lagrangian¹⁵ then the adjoint function must have the property

$$\tilde{f} \frac{\partial \psi}{\partial x} = 0. \quad (3.16)$$

This property has a natural physical interpretation, that is if the average velocity of a controlled process is high, the adjoint process should be characterized by small space gradients, and conversely steep gradients in the adjoint process necessarily should imply a slow controlled dynamics.

This mathematical formalization of a connection between slow and fast motions may be inappropriate for complex dynamic systems that are often characterized by steep space-time gradients of functions unknown *a-priori*. It leads to the situation when an adequate approximation of the Hamiltonian can be made in general only on the basis of the problem-specific information.

A level of unification in the mathematical theory of optimal control can be obtained when the Hausdorff assumption is imposed on the topological space of dynamic system state-space. In some cases this allows the dynamic programming equation to be reduced to a partial differential equation of the HJB type. This assumption is natural when it is applied to deterministic finite dimensional systems for which the continuity of the function $x(\cdot)$ may be taken for granted. In such deterministic situations the forcing term \tilde{f} of the system dynamics, defined by the set of *statistical equalities*

$$x(t + \Delta t) - x(t) = \tilde{f}(t, x(t), u_{\beta_1}^{\beta_2}(t, x))\Delta t, \quad (3.17)$$

can be reasonably well approximated by a differential equation (2.5) provided that the point $x(t_0)$ is specified. To generalize approximations of (3.17) obtainable from the model (2.5), we can use stochastic mathematical models or, alternatively, some averaging procedures for deterministic models. In both situations we eventually split the forcing term \tilde{f} into several parts using *problem-specific information*.

¹⁴recall that in classical mechanics the quantity $\int_{t_0}^T L dt$ is called the action functional

¹⁵or near the running cost if we use the terminology of optimal control theory

A wide class of mathematical models obtained from such splitting is provided by equation (2.25). Let us assume that $t_0 \leq t < t + \Delta t \leq T$. Then from (2.25) we have

$$x(t + \Delta t) = x(t) + \int_t^{t+\Delta t} \bar{f}(\tau, x(\tau), u_{\beta_1}^{\beta_2}(\tau, \cdot)) d\tau + \int_t^{t+\Delta t} \bar{\sigma}(\tau, x(\tau), u_{\beta_1}^{\beta_2}(\tau, \cdot)) d\omega. \quad (3.18)$$

This model allows an exchange of information between two splitting terms \bar{f} and $\bar{\sigma}$ which, in turn, allows the assumption (2.13) to be relaxed. This assumption is typical for deterministic models. For stochastic models neither

$$\int_{t_0}^T |\bar{\sigma}(\tau, x(\tau), u_{\beta_1}^{\beta_2}(\tau, \cdot))| d\omega = 0 \quad (3.19)$$

nor

$$\int_{t_0}^T |\bar{\sigma}(\tau, x(\tau), u(\tau, \cdot))| d\tau = 0 \quad (3.20)$$

is assumed a-priori. In the general case no matter how small $\Delta t > 0$ is assumed to be, the local conditions

$$\int_t^{t+\Delta t} |\bar{\sigma}(\tau, x(\tau), u_{\beta_1}^{\beta_2}(\tau, \cdot))| d\omega = 0, \quad \int_t^{t+\Delta t} |\bar{\sigma}(\tau, x(\tau), u_{\beta_1}^{\beta_2}(\tau, \cdot))| d\tau = 0 \quad (3.21)$$

may be also violated in principle for an arbitrarily chosen t from the interval $(t_0, T - \Delta t)$. In such cases the connection between PMP and DPP, in the sense of the relationship (3.1), loses its meaning and should be replaced by imbedding results similar to (3.2).

The problem of their connection has its ultimate roots in the solution of the problem (1.3) by the method (1.5) which essentially uses the geometric interpretation of the problem. When applying the DPP to the derivation of the HJB type equations we are required to justify an expansion of the unknown function $V(t + \Delta t, x + \Delta x)$ about the point (t, x) which automatically implies some a-priori smoothness assumptions on the function V . Let us initially assume that $V(t, x) \in C^{2,2}$. (This is an assumption that is certainly excessive for the majority of optimal control applications.) In the deterministic case the quality of the approximation of the original problem (2.5)-(2.7) by the model (3.9)-(3.12) is determined by the remainder term of the form [15]:

$$R = \frac{1}{2} \{ A(\Delta x)^2 + B\Delta x\Delta t + C(\Delta t)^2 \},$$

where A and C define second derivatives with respect to x and t respectively, and B denotes the mixed derivative of the unknown function V . If the approximation is required to be at least of the second order with respect to Δt we have to set

$$R = o(\Delta t) \text{ i.e. } \lim_{\Delta t \rightarrow 0} R/\Delta t = 0,$$

where o is the Landau symbol. This implies a certain connection between the partial derivatives of the function V

$$\lim_{\Delta t \rightarrow 0} \{ \Delta x [A \frac{\Delta x}{\Delta t} + C \frac{\Delta t}{\Delta x}] \} = - \lim_{\Delta t \rightarrow 0} B \Delta x.$$

The validity of this depends on the topological properties of a space-time region of interest which in turn, the control function is dependent on. Hence, we have to reformulate the optimality principle (3.8) in the form of a connection between control and value functions before the actual approximation of the Hamiltonian. Alternatively, we can allow an additional degree of freedom for the Hamiltonian by introducing the Hamilton-Pontryagin function [30,39]

$$H(t, x, u, \psi, a_0) = -a_0 f_0 + \bar{f} \psi, \quad (3.22)$$

where a_0 is a parameter of normalization. In R^n the Hamilton-Pontryagin function is a direct generalization of the classic mechanics Hamiltonian (2.20) and can be written as

$$H(t, x, u, \psi, a_0) = -a_0 f_0 + \sum_{i=1}^n \psi_i f_i. \quad (3.23)$$

If it is possible to justify a-priori that $a_0 > 0$ then the normalization condition for the Hamiltonian takes the form $a_0 = 1$. Otherwise a parametric normalization with respect to the quantities $(a_0, \psi_1, \dots, \psi_n)$ in (3.23) (or with respect to a_0 and ψ in (3.22)) leads to a *boundary problem* of the PMP for the optimal control [39]. In the general case, the a-priori assumption that the coefficient near the function f_0 equals 1 is inappropriate. Similarly, in stochastic optimal control problems an approximation of the Hamiltonian by the formula (2.28) cannot be rigorously justified unless it is known that such normalization of the Hamiltonian is valid. In many cases such validation can be provided by the problem-specific information.

By assuming the possibility of a probabilistic character of evolution for a controlled dynamic system, we simultaneously require an *a-priori* deterministic condition for the normalization of the Hamiltonian approximation when the procedure of model construction is being undertaken. Since the continuity assumption for the function $x(\cdot)$ is not applicable in the general stochastic case, this implies the increasing importance of the analysis of neighborhood states of evolution. The question on how the exchange of information between such states takes place can be eventually answered by the adjustment of regularities of the control and the value function in mathematical models which we apply. In general, such an adjustment can be performed efficiently using the information which relates the mathematical model to the real-world situation.

In the next section, we propose the mathematical formalization of such an adjustment. We assume that both the control and value functions are from the Banach space $L^1(Q_T)$, where $Q_T = \{(x, t) : t_0 \leq t \leq T, x_0 \leq x \leq x(T)\}$ is a space-time region of interest which can be thought to be a topological approximation of the set Ω_0^0 . Although the theoretical possibility that $T \rightarrow \infty$ is not excluded by Theorem 4.1, any specific mathematical model of optimal control with the infinite horizon requires a justification of the Hamiltonian normalization. In turn, this requires a certain trade-off between averaging and discounting procedures that is achievable at least in principle by relating the model to the real-world situation.

4 Regularization algorithm for mathematical models of optimal control.

Let $f_0 \in L^1(Q_T)$. We assume that $t_0 \leq t \leq t^0 \leq T$ and denote the topological space that approximates Ω_0^0 by

$$Q_0^0 = \{(t, x) : t_0 \leq t \leq t^0, x_0 \leq x \leq x^0(t^0)\}. \quad (4.1)$$

Let us assume that a mathematical model of optimal control has been constructed and $(x(\cdot), u_{\beta_1}^{\beta_2}(\cdot, \cdot))$ is the process described by this model, where dots represent the dependency of the process on time, neighborhood states, and the goal functional defined, for example, by (2.7) or by (2.26). Then the governing equation for the system dynamics can be considered in the form of (2.5)(as a partial case by (2.11)) or (2.25) respectively.

To reflect the dependency of the control function on the neighborhood states at a given moment of time τ we will write that

$$u_{\beta_1}^{\beta_2}(\tau, \cdot) = u_{\beta_1}^{\beta_2}(\tau, x(h(\tau))),$$

where $h(\tau)$ is a recursive function of time referred to as the function of the neighborhood states when the model is specified. This leads to the interpretation of control as a family of functions that form a control population, and consequently to the coupling between the topology of Q_0^0 and the definition of the function h which is not known a-priori from the mathematical model itself.

We assume that for any two control populations $\hat{u}_{\beta_1}^{\beta_2}, \tilde{u}_{\beta_1}^{\beta_2} \in \mathcal{U}$ that are realizable with the same goal function f_0 the Lipschitz-type condition holds in $L^1(Q_0^0)$, thus

$$\|f_0(t, x, \hat{u}_{\beta_1}^{\beta_2}) - f_0(t, x, \tilde{u}_{\beta_1}^{\beta_2})\|_{L^1(Q_0^0)} \leq L \|\hat{u}_{\beta_1}^{\beta_2} - \tilde{u}_{\beta_1}^{\beta_2}\|_{L^1(Q_0^0)} \quad (4.2)$$

(here L may be dependent of the topologies β_1 and β_2). Then we define the performance measure for the mathematical model by the Lebesgue integral as

$$V^h(x, t, u_{\beta_1}^{\beta_2}) = \int_t^{t^0} f_0(\tau, x(h(\tau)), u_{\beta_1}^{\beta_2}(\tau, x(h(\tau))) d\tau. \quad (4.3)$$

Let us assume that for the model under consideration the function h , as a function of time, can be specified, at least in principle, in the topological space Q_0^0 as a whole. We introduce the following definition.

Definition 4.1 A pair of positive functions χ_1 and χ_2 that are dependent on the function h is called the conjugate pair of probabilistic weights in Q_0^0 if it satisfies the inequality

$$\chi_1 + \chi_2 \leq 2.$$

Let us further assume that there exist two points (t, x) and (t', x') (which may coincide) in the topological space Q_0^0 such that $t_0 < t \leq t' < t^0$. We define the topological spaces Q_L and Q_R by the following inequalities respectively

$$Q_L = \{(t, x) : t_0 \leq t \leq t', x_0 \leq x \leq x'(t')\}, \quad (4.4)$$

and

$$Q_R = \{(t', x') : t \leq t' \leq t^0, x \leq x' \leq x^0(t^0)\}, \quad (4.5)$$

where all values x and x' may be dependent on the function h .

Such an interpretation of optimal control problems allows us to relate the mathematical model to the problem specific-information which is formalized through the function h . This provides the basis on which the choice of the algorithm should be made [1]. Since $u_{\beta_1}^{\beta_2}$ is a population of all realizable controls with the domain of definition that can be imbedded in Q_0^0 , the values of $u_{\beta_1}^{\beta_2}$ depend on the topology T (provided X is specified we approximate the topological product Ω_0^0 by Q_0^0) as well as on the choice of measures ω in the resulting topological approximation.

Hence, globally in the region Q_0^0 , the control is a function of time, realizable states, topology and measure

$$u_{\beta_1}^{\beta_2}(t, x) = u(t, x; T, \omega).$$

At the same time, locally in the neighborhood of any specified point, the control is a function of time and states. Since this function is dependent on the definition of the neighborhood states, which in turn are dependent on h , in general we will write $u_{\beta_1}^{\beta_2} = u^h(t, x)$. Of course, at any given point $(t, x) \in Q_0^0 \approx \Omega_0^0 = T \otimes X$, we may have many control functions that may belong to different control populations. They may be different in the sense of the definition of their neighborhood state functions h , however their values at a specified point of Q_0^0 are expected to be the same, given the model has been constructed.

To measure the quality of control with respect to the goal function we introduce the operator of controlled measures by the formula

$$\Phi^h(t, x) = u^h(t, x) - \int_t^{t^0} f_0(\tau, x(h(\tau)), u^h(\tau, x)) d\tau, \quad (4.6)$$

In general, this operator may be unbounded subject to the definition of the goal function f_0 and the function of neighborhood states h .

The main result of this section provides a constructive approximation of measures to guarantee boundedness properties of the operator $\Phi^h(t, x)$. It is shown below that, by the appropriate choice of measures in Q_0^0 , the L^1 -norm of the operator (4.6) can be made arbitrarily close to zero under quite mild a-priori assumptions on the functions u^h and V^h . As a result of our construction, the operator $\Phi^h(t, x)$ becomes arbitrarily close to an absolutely continuous function.

Theorem 4.1 Let $u^h(t, x) \in L^1(Q_0^0)$ and $V^h(t, x) \in L^1(Q_0^0)$, where Q_0^0 is defined by (4.1). Let us assume that for any $\epsilon > 0$ there exist such functions $\chi_1(M) > 0$, $\chi_2(N) > 0$ of integer numbers M and N respectively that

$$\chi_1(M) + \chi_2(N) \leq 2, \quad (4.7)$$

and that the following inequalities for the Lebesgue measure ω defined in Q_0^0

$$\omega(Q_L) < \delta_L, \quad \omega(Q_R) < \delta_R, \quad (4.8)$$

hold where subregions Q_L and Q_R in Q_0^0 are defined by (4.4) and (4.5) respectively and

$$0 < \delta_L < \frac{\epsilon \chi_1(M)}{N+1}, \quad 0 < \delta_R < \frac{\epsilon \chi_2(N)}{M+1}, \quad (4.9)$$

Then

$$\|\Phi^h(t, x)\|_{L^1(Q_0^0)} < \epsilon. \quad (4.10)$$

Proof.

The proof of the theorem consists of two parts which correspond to the forward and backward procedures of the model construction respectively.

I. Forward procedure. First we assume that a moment t_0 in the initial condition of the model¹⁶ is given with the probability exactly 1 (if the model is based on the governing equation (2.5), then with the deterministic certainty of (2.6)). No other assumptions with the same probability are made a-priori. We introduce the sequence of sets $Q_{t_0^0}, Q_{t_0^1}, \dots$ by the formula

$$Q_{t_0^i} = \{(t, x) \in Q_0^0 : i \leq |\Phi^h(t, x)| < i+1\}, \quad i = 0, 1, \dots \quad (4.11)$$

Then for the Lebesgue measure ω in Q_0^0 we have

$$\int_{Q_0^0} |\Phi^h(t, x)| d\omega = \sum_{i=0}^{\infty} \int_{Q_{t_0^i}} |\Phi^h(t, x)| d\omega \quad (4.12)$$

due to the σ -additivity property of the Lebesgue integral.

Let us consider two cases.

Case (a): the deterministic step followed by the probabilistic approach.

Let the function f_0 and the measure ω be given in Q_0^0 . Then for arbitrarily small $\epsilon > 0$ and arbitrarily large integer number N' let δ_L be such that

$$0 < \delta_L < \frac{\epsilon}{2(N'+1)}. \quad (4.13)$$

Since (4.7) implies $\chi_1(M) < 2$, we have from (4.9) that

$$\omega(Q_L) < \delta_L, \quad (4.14)$$

where

$$Q_L = \bigcup_{i=0}^{N'} Q_{t_0^i}. \quad (4.15)$$

Then from (4.8), (4.9) and (4.11)-(4.15) we conclude that

$$\int_{Q_L} |\Phi^h(t, x)| d\omega < \frac{\epsilon}{2}. \quad (4.16)$$

¹⁶for example, with the governing equation (2.25)

If N' is a-priori fixed then this allows us to approximate the process described by the model, at least up to the time $t_0^{N'}$, which defines a lower time-bound for the limit of the model predictability. The value $t_0^{N'}$ depends on the topology \mathcal{T} and the measure ω which are subject to our a-priori choice. Such a choice implies probabilistic features of our knowledge on the evolution of the process from $t_0^{N'}$ to t^0 . The possible existence of the process related to the evolution from $t_0^{N'}$ to t^0 [4] can be formalized mathematically by introducing in the model a probabilistic weight function $\bar{\chi}_1(M)$ and by assuming a possibility of such evolution by a-priori choice of the topology \mathcal{T} and the measure ω in the topological space Q_0^0 . Let us define a remainder set by $Q_{t_0}^{N'+1} = Q_0^0 \setminus Q_L$. Then we choose N' in such a way that

$$\int_{Q_{t_0}^{N'+1}} |\Phi^h(t, x)| d\omega = \sum_{i=N'+1}^{\infty} \int_{Q_{t_0}^i} |\Phi^h(t, x)| d\omega < \frac{\epsilon}{2} \bar{\chi}_1(M). \quad (4.17)$$

The dependency of the probabilistic weight function $\bar{\chi}_1$ on M reflects the fact of possibility of model improvement when additional information on the problem becomes available. From the probabilistic point of view we expect that

$$\text{if } M \rightarrow \infty, \text{ then } \bar{\chi}_1(M) \rightarrow 1. \quad (4.18)$$

Case (β): the probabilistic step followed by the deterministic approach.

Alternative reasoning is based on the probabilistic assumption (4.18). If (4.18) holds then given \mathcal{T} and ω for any $\epsilon > 0$ we can choose N' such that

$$\int_{Q_{t_0}^{N'+1}} |\Phi^h(t, x)| d\omega = \sum_{i=N'+1}^{\infty} \int_{Q_{t_0}^i} |\Phi^h(t, x)| d\omega < \frac{\epsilon}{2}. \quad (4.19)$$

This step challenges the problem on the limit of predictability for backward evolution. Specifically, since $t_0^{N'}$ (implicitly defined by the set $Q_{t_0}^{N'+1}$) depends on the topology \mathcal{T} (that a-priori introduced in Q_0^0) and the measure ω , we have to formalize mathematically a possible reversibility of evolution from t_0 to $t_0^{N'}$. Such a formalization is performed by introducing in the model a probabilistic weight $\hat{\chi}_1(M)$ that satisfies the assumptions (4.8), (4.9) of the theorem. As a result we have

$$\omega(Q_L) < \delta_L \text{ for } 0 < \delta_L < \frac{\epsilon \hat{\chi}_1(M)}{2(N'+1)}. \quad (4.20)$$

Then from (4.15) we conclude

$$\int_{Q_L} |\Phi^h(t, x)| d\omega = \sum_{i=0}^{N'} \int_{Q_{t_0}^i} |\Phi^h(t, x)| d\omega < \frac{\epsilon}{2} \hat{\chi}_1(M). \quad (4.21)$$

II. Backward procedure. Let us assume now that the moment t^0 is given, from the set Q_0^0 , or may at least be given in principle with probability exactly 1 (or with the deterministic certainty of (2.24) or (3.10)). This type of assumptions is typical when backward evolution equations are formulated with the terminal (Cauchy type) conditions in a similar way to the model (2.27).

Analogous to the forward procedure this situation gives two cases to be considered.
Case (α): the deterministic step followed by the probabilistic approach.

Let

$$0 < \delta_R < \frac{\epsilon}{2(M'+1)}, \quad (4.22)$$

where ϵ and M' can be an arbitrarily small positive real number and an arbitrarily large integer number respectively. Similar to (4.11), (4.12) we use the σ -additivity property of the Lebesgue integral to conclude that

$$\int_{Q_0^0} |\Phi^h(t, x)| d\omega = \sum_{j=0}^{\infty} \int_{Q_{t_j}^0} |\Phi^h(t, x)| d\omega \quad (4.23)$$

where

$$Q_{t_j^0} = \{(t, x) \in Q_0^0 : j \leq |\Phi^h(t, x)| < j + 1\}. \quad (4.24)$$

On the basis of (4.22) and (4.8), (4.9) we have

$$\omega(Q_R) < \delta_R, \quad (4.25)$$

where

$$Q_R = \bigcup_{j=0}^{M'} Q_{t_j^0}. \quad (4.26)$$

This allows us to conclude that

$$\int_{Q_R} |\Phi^h(t, x)| d\omega = \sum_{j=0}^{M'} \int_{Q_{t_j^0}} |\Phi^h(t, x)| d\omega < \frac{\epsilon}{2}. \quad (4.27)$$

In view of the fact that M' can be arbitrarily large, we fix it to satisfy the following inequality with the probabilistic weight $\bar{\chi}_2(N)$

$$\int_{Q_{t_{M'+1}^0}} |\Phi^h(t, x)| d\omega = \sum_{j=M'+1}^{\infty} \int_{Q_{t_j^0}} |\Phi^h(t, x)| d\omega < \frac{\epsilon}{2} \bar{\chi}_2(N), \quad (4.28)$$

where $0 < \bar{\chi}_2(N) < 2$. By (4.28) we formalize mathematically the possibility of recovering a backward evolution from $t_{M'}^0$ to t_0 for the specified mathematical model. To reflect the possibility of the model improvement it is expected that

$$\text{if } N \rightarrow \infty, \text{ then } \bar{\chi}_2(N) \rightarrow 1. \quad (4.29)$$

Case (β): probabilistic step followed by the deterministic approach.

If we a-priori assume that the limiting situation (4.29) holds for the mathematical model, then we can always choose M' in such a way that the following estimate

$$\int_{Q_{t_{M'+1}^0}} |\Phi^h(t, x)| d\omega = \sum_{j=M'+1}^{\infty} \int_{Q_{t_j^0}} |\Phi^h(t, x)| d\omega < \frac{\epsilon}{2} \quad (4.30)$$

will be satisfied for any arbitrarily small ϵ . However, to justify the mathematical model we need an assumption of reversibility of evolution from $t_{M'}^0$ to t^0 . Since the definition of $t_{M'}^0$ is given implicitly through the definition of the set $Q_{t_{M'+1}^0}$ and is not independent of the measure ω and the topology in Q_0^0 , we choose a probabilistic weight function $\hat{\chi}_2(N)$ to satisfy the assumptions of the theorem for any arbitrarily large M' chosen to satisfy (4.30). Hence, the choice of δ_R such that inequality

$$0 < \delta_R < \frac{\epsilon \hat{\chi}_2(N)}{2(M' + 1)} \quad (4.31)$$

holds, implies that

$$\omega(Q_R) < \delta_R. \quad (4.32)$$

Therefore

$$\int_{Q_R} |\Phi^h(t, x)| d\omega = \sum_{j=0}^{M'} \int_{Q_{t_j^0}} |\Phi^h(t, x)| d\omega < \frac{\epsilon}{2} \hat{\chi}_2(N). \quad (4.33)$$

To combine both above procedures we introduce the following functions
• for the forward procedure

$$\chi_1(M) = \begin{cases} \hat{\chi}_1(M) & \text{for the case}(\alpha), \\ \bar{\chi}_1(M) & \text{for the case}(\beta), \end{cases}$$

- and for backward procedure

$$\chi_2(N) = \begin{cases} \bar{\chi}_2(N) & \text{for the case } (\alpha), \\ \hat{\chi}_2(N) & \text{for the case } (\beta). \end{cases}$$

These functions are coupled by the condition (4.7) that allows us to define the conjugate pair of probabilistic weights (χ_1, χ_2) which characterizes the mathematical model under consideration. Then from estimates (4.16), (4.17), (4.19), (4.21) and (4.27), (4.28), (4.30), (4.33) we derive (4.10) and this completes the proof. ■

Remark 4.1 *The case (α) of the forward procedure is essentially an approximation of the boundary of the set Q_L whereas the case (β) is an approximation of the boundary of the supplement of Q_L to the whole set Q_0^0 , that is an approximation to the boundary of the set $Q_0^0 \setminus Q_L$. Therefore, the approaches of both cases are equivalent in the sense that they attempt to approximate the boundary between Q_L and $Q_0^0 \setminus Q_L$ separated by a surface that characterizes system dynamics at $t = t_0^{N'}$. The a-priori unknown topology of this surface in the general case leads to the limiting considerations $t \rightarrow \infty, N' \rightarrow \infty$, which cannot be justified within the forward procedure itself.*

Remark 4.2 *Specifying t^0 either explicitly or implicitly (by the set Q_0^0) we use the backward procedure to approximate the boundary of $Q_0^0 \setminus Q_R$ or the boundary of Q_R in the cases (α) and (β) respectively. In general, the surface separating these two sets is not reducible to a point, though such a reduction can be always effectively performed under the assumption that the initial conditions in the mathematical model are given either precisely or with the probability exactly 1. If such an assumption is made a-priori, then the resulting mathematical models manifest a competition between the Markovian character of evolution and classical conservation laws [21]. In the framework of theorem 4.1 this corresponds to the situations when (4.18) and (4.29) both hold.*

5 Logical issues related to the probabilistic weight functions.

Although the possibility of simultaneous realization of (4.18) and (4.29) is not excluded by theorem 4.1, we have not required the assumptions

$$\chi_1(M) \leq 1, \quad \chi_2(N) \leq 1 \tag{5.1}$$

in the proof. Nevertheless, if (5.1) holds, it is possible to interpret the probabilistic weight functions as a pair of conjugate (or coupled) random functions that are associated with the model itself. Due to the requirement of positivity of χ_1 and χ_2 neither part of such a conjugate pair vanishes although each of them can be arbitrarily small.

On the other hand, the possibility of equalities in (5.1) is induced by the assumption that a point on the phase space trajectory of a dynamic system may be specified at least in principle with the probability exactly 1. Physically, this assumption requires invariance of the density along the trajectory and is often referred to as conservation of extension in phase. Theorem 4.1 includes this possibility, yet the realization of this possibility excludes one logically possible case in the comparison of two arbitrary sets in (2.3). In order to include all four possibilities, we have introduced a recursive dependency of control on the structure of the a-priori defined topological space and measures that are used. Such a consideration allows us to take into account incompleteness of information available a-priori, and as a result better reflects the reality of modelling in mathematical control theory.

When constructing mathematical models of optimal control it is important to take into account that the Hamiltonian (Lagrangian) of dynamic systems is always an approximation which is dependent on

- approximation of initial conditions for the system;
- approximation of system-environment boundary interface.

Such dependency can be relaxed by appropriate assumptions on the topology of the state space and initial conditions of the system. Such relaxation is based on the two logical steps that form the core of mathematical models construction:

- the notion of the empty set;
- the definition of a point or a surface with probability exactly 1.

Typically, in the theory of optimal control we require both of these steps to be undertaken simultaneously. Theorem 4.1 provides a constructive way for the sequential interchange of information between these two arguments. Using conjugate pairs of probabilistic weights, it is possible to analyse mathematical models of optimal control on the basis of the associated partial differential equations without a-priori assumptions related to the Hamiltonian normalization on the boundary interface (similar to (2.23) or (2.28)). In many applications of the theory of optimal control the assumption of the existence of a point (t, \mathbf{z}) , for which the following probabilistic equality holds,

$$P\{|x(t + \tau) - x(t)| < \epsilon_1 \wedge |u(t + \tau, \cdot) - u(t, \cdot)| < \epsilon_2\} = 1, \quad (5.2)$$

can be justified for sufficiently small τ and arbitrarily small $\epsilon_1 > 0$, but ϵ_2 not necessarily small. Although this equality is natural in classical problems of mechanical control when dynamic motion is idealized by a continuous phase space trajectory, the possibility of its realization becomes questionable when we approach the solution of singular control problems.

Similarly, it is often possible to reduce the dynamic programming equation to a partial differential equation with the given approximation to the Hamiltonian H and the unknown value function V . In such cases it is often straightforward to assume that there exists sufficiently small τ and $\hat{\epsilon}_1$ such that the equality

$$P\{|V(t + \tau, \cdot) - V(t, \cdot)| < \hat{\epsilon}_1 \wedge |H(t + \tau, \cdot) - H(t, \cdot)| < \hat{\epsilon}_2\} = 1 \quad (5.3)$$

holds for some value $\hat{\epsilon}_2$ (not necessarily small). In the general case, neither control changes nor the Hamiltonian increment in that small period of time τ are required to be small, and the values of ϵ_2 and $\hat{\epsilon}_2$ can, in principle, be arbitrarily large.

The a-priori continuity assumption imposed on at least one of the functions appearing in (5.2) and (5.3) leads to the possibility of theoretical justification of the mathematical model. In the context of Theorem 4.1 this assumption is equivalent to the “freeze” of one of the functions \bar{x}_1 , \hat{x}_1 , \bar{x}_2 or \hat{x}_2 at a constant level, and the investigation of the dynamics induced by the resulting approximation. The quality of such approximations is eventually defined by a comparison between

$$|\epsilon_2 - \hat{\epsilon}_1| \text{ and } |\hat{\epsilon}_2 - \epsilon_1|. \quad (5.4)$$

With increasing values of ϵ_2 and $\hat{\epsilon}_2$ such approximations become inappropriate when the absolute value of the difference between ϵ_1 and $\hat{\epsilon}_1$ also increases. In order to improve such approximations the regularities of the control function u and the value function V should be adjusted by a sequential procedure based on estimates of the quantities (5.4). The freezing of one of the functions in (5.2), (5.3) determines the boundary interface between two sets that often includes *internal boundary conditions in a recursive manner* which requires the solution of a multiphase regularity problem. As a result, the construction of the model in general and the definition of the topology of the boundary interface in particular ultimately defines the choice of the method for an efficient solution of the problem.

The a-priori assumptions (4.18) and (4.29) may play a decisive role in mathematical justification of the model in those cases when an efficient algorithm for solving a problem described by the mathematical model is not known. Finally, these assumptions allow us to claim that two distinct states of a dynamic system that correspond to different times of its evolution can be determined with the same probability (exactly 1) using a mathematical model. The theoretical possibility of the existence of such a model is covered by Theorem 4.1 as a limiting case.

6 Concluding remarks and future directions.

Each algorithm is model-specific in the sense that its quality essentially depends on the consistency of the mathematical model to the real-world situation [37]. Since *constructing* a model is an art whereas *deriving* an algorithm is a science, it would be reasonable to combine these two processes.

This idea suggests future directions of this work. Since the Hamiltonian of any dynamic system can be given only approximately in general, continuity arguments may not be appropriate for construction of mathematical models with the unknown value function in Hausdorff topological spaces. The partial differential equations resulting from such arguments *do* acquire hyperbolic features as is expected from the general principles of extended thermodynamics [25,13]. The results on the derivation of algorithms from the computational models that have been obtained on the basis of partial differential equations that contain a hyperbolic mode (generalized energy equations [21]) in stochastic, nonsmooth and purely deterministic cases will be the subject of a separate publication. Some preliminary results in this field can be found in [21]. The underlying numerical procedures are essentially based on the Markov Chain approximation method [19], that allows us to define an approximation to the mapping H between the two topological spaces $\Omega_0^0 = \mathcal{T} \otimes \mathcal{X}$ and \mathcal{E} . For any mapping H , given a-priori, the possible existence of a homeomorphism between these topological spaces can not be established without additional a-posteriori information. Therefore, the approach based on an approximation of such a mapping provides a natural, and the most general, framework for the mathematical modelling of dynamic systems and evolutionary processes. This framework leads one to view the problem of control as much broader than the current field of optimal control theory [16,32].

Acknowledgements

I wish to acknowledge the support of the School of Mathematics at the University of South Australia. I would like also to thank Dr John Boland for his helpful comments and suggestions of improvement and Dr Jemery Day for helpful assistance at the final stage of preparation of this paper.

References

- [1] Anderssen, R.S. *Global optimization*, in Optimization, University of Queensland Press, 1972, 26-48.
- [2] Benes, V.E., Shepp, and Witsenhausen, H.S. *Some solvable stochastic control problems*, Stochastics, 4, 1980, 181-207.
- [3] P. Diamond, P. Kloeden and A. Pokrovskii, *Degree theory on finite lattices: discretizations of dynamical systems*, Numer. Funct. Anal. and Optimiz. vol.16, no. 1&2, 1995, 43-52.
- [4] D. Dubois and H. Prade, *Possibility Theory*, Plenum Press, New York, 1988.
- [5] Falcone, M., Ferretti, R., *Discrete time high-order schemes for viscosity solutions of Hamilton-Jacobi-Bellman equations*, Numerische Mathematik, 67, 1994, 315-344.
- [6] Falcone, M., Giorgi, T. and P. Loreti, *Level sets of viscosity solutions: some applications to fronts and rendez-vous problems*, SIAM J. App. Math., Vol.54, No.5, 1994, 1335-1354.
- [7] Fleming, W. and H. Soner, *Controlled Markov Processes and Viscosity Solutions*. Springer-Verlag, 1993.
- [8] Frankowska, H. *Lower semicontinuous solutions of Hamilton-Jacobi-Bellman equations*, SIAM J. Control and Optimiz., vol.31, no.1, pp.257-272, 1993.
- [9] J. W. Gibbs, *Elementary Principles in Statistical Mechanics*, Dover, New York, 1960.
- [10] Goldstein, S. *Sufficient Conditions to Single out the Gibbs Measure from other Time-Invariant Measures*, in Long-Time Prediction in Dynamics. Ed. by C.W.Horton, Jr., L.E.Reichl, and V.G.Szebehely, John Wiley & Sons, 1983, 71-78.
- [11] Hawking, S., *The edge of spacetime*, in The New Physics, Ed. Paul Davis, Cambridge; New York: Cambridge University Press, 1989, 61-69.
- [12] Johnson, C. *A New Paradigm for Adaptive Finite Element Methods*, in The Mathematics of Finite Elements and Applications. Ed. by J. R. Whiteman, John Wiley & Sons, 1994, 105-120.
- [13] Jou, D., Casas-Vazquez, J. and G. Lebon, *Extended Irreversible Thermodynamics*, Springer-Verlag, 1993.
- [14] Kawahara, M., Anju, A. *Control, identification and estimation in finite elements in fluids*, The Third World Congress on Computational Mechanics, V.1, 1994, 300-305.
- [15] Kirk, D.E. *Optimal Control Theory*, Englewood Cliffs, N.J.: Prentice-Hall, 1970.
- [16] Kirilik, A., Miller, R.A., and Jagacinski, J., *Supervisory Control in a Dynamic and Uncertain Environment: A Process Model of Skilled Human-Environment Interaction*, IEEE Trans. on Systems, Man, and Cybernetics, vol.23, no.4, 1993, 929-951.
- [17] A. Kolmogorov and S. Fomin, *Fundamentals of Theory of Functions and Functional Analysis*, Moscow: Nauka, 1989, (or by Dover Publications, New York, 1975).
- [18] Krylov, N.V., *Controlled Diffusion Processes*, Springer-Verlag, 1980.
- [19] Kushner, H., and P. Dupuis, *Numerical Methods for Stochastic Control Problems in Continuous Time*. New York: Springer-Verlag, 1992.
- [20] Mackey, M.C. *Time's Arrow: The Origins of Thermodynamic Behaviour*. Springer-Verlag, 1992.
- [21] Melnik, V.N., *Nonconservation Law Equation In Mathematical Modelling: Aspects of Approximation*, Proceedings of the International Conference AEMC'96, Sydney, pp.423-430, 1996.
- [22] Melnik, V.N., *Convergence of the operator-difference scheme to generalized solutions of a coupled field theory problem*, to appear in Journal of Difference Equations and Applications, 1997.

- [23] Milner, F.A., Park, E.J. *Mixed finite element methods for Hamilton-Jacobi-Bellman-type equations*, IMA Journal of Numerical Analysis, 16, 1996, 399-412.
- [24] Misra, B., and I. Prigogin, *Time, Probability, and Dynamics*, in Long-Time Prediction in Dynamics, Ed. by C.W.Horton, Jr., L.E.Reichl, and V.G.Szebehely, John Wiley & Sons, 1983, 21-43.
- [25] Muller, I. and T. Ruggeri, *Extended Thermodynamics*, Springer-Verlag, 1993.
- [26] Newton, I., *A tretise of the method of fluxions and infinite series, with its application to the geometry of curve lines*, in V.1, Mathematical Works, New York: Johnson Reprint Corp., 1964-1967.
- [27] M. Nisio, *Optimal control for stochastic partial differential equations and viscosity solutions of Bellman equations*, Nagoya Math. J., vol.123, pp.13-37, 1991.
- [28] Okuda, H., Yagawa, G. and Yashiki, T. *Three-dimensional incompressible flow analysis using adaptive finite element method*, The Third World Congress on Computational Mechanics, V.1, 1994, 167-168.
- [29] Osher, S., Shu, C.-W., *High-Order essentially nonoscillatory schemes for Hamilton-Jacobi equations*, SIAM J. Numer. Anal., Vol.28, No.4, 1991, 907-922.
- [30] Pontryagin, L.S. et al, *The Mathematical Theory of Optimal Processes*, New York: Gordon and Breach Science Publishers, 1986.
- [31] Reyna, L.G., Ward, M.J., *On exponential ill-conditioning and internal layer behavior*, Numer. Funct. Anal. and Optimiz., 16 (3 &4), 1995, 475-500.
- [32] Rouse, W. B., *Systems Engineering Models of Human-Machine Interaction*, New York: North Holland, 1980.
- [33] Russell, B., *Human Knowledge. Its Scope and Limits*. London: Allen & Unwin, 1966.
- [34] H. M. Soner and S. E. Shreve, *Regularity of the value function for a two-dimensional singular stochastic control problem*, SIAM J. Control and Optimiz., vol.27, no.4, 1989, 876-907.
- [35] Sontag, E.D. *Mathematical Control Theory*, Springer-Verlag, 1990.
- [36] Struwe, M., *Variational Methods*, Springer-Verlag, 1990.
- [37] A. Sucharev, *Minimax Models in the Theory of Numerical Methods*, Kluwer Academic Publishers, 1992.
- [38] Vainberg, M.M. and Aizengendler, P.G. *The Theory and Methods of Investigation of Branch Points of Solutions*. in Progress in Mathematics, Vol.2, Ed. R.V. Gamkrelidze, New York: Plenum Press, 1968, pp.1-72.
- [39] Vasiliev, F.P. *Numerical Methods for the solutions of extremal problems*, Moscow, Nauka, 1988.
- [40] Zeidler, E. *Applied Functional Analysis*, Springer-Verlag, 1995.
- [41] Zhou, X.Y., *Maximum principle, dynamic programming, and their connection in deterministic control*, Journal of Optimization Theory and Applications, Vol.65, No.2, 1990, 363-373.
- [42] Zhou, X.Y., *The connection between the maximum principle and dynamic programming in stochastic control*, Stochastics and Stochastics Reports, Vol.31, 1990, 1-13.

UNIVERSITY OF SOUTH AUSTRALIA



TECHNICAL REPORT SERIES

**Optimal Probabilistic
Trajectories of Deterministic
Finite-State Machines**

by

(R.) V Nick Melnik

Report No. 1996/15

CENTRE FOR INDUSTRIAL
AND APPLIED MATHEMATICS
SCHOOL OF MATHEMATICS

Faculty of Information Technology

The Levels, South Australia 5095, Telephone (08) 8302 3343 Facsimile (08) 8302 5785

TECHNICAL REPORT SERIES

**Optimal Probabilistic
Trajectories of Deterministic
Finite-State Machines**

by

(R.) V Nick Melnik

Report No. 1996/15

Optimal Probabilistic Trajectories of Deterministic Finite-State Machines*

V.Nick Melnik

E-mail: matvnm@lv.levels.unisa.edu.au

Abstract

In this paper computational aspects of the mathematical modelling of dynamic system evolution have been considered as a problem in information theory. The construction of such models is treated as a decision making process with limited available information. The solution of the problem is associated with a computational model based on heuristics of a Markov Chain in a discrete space-time of events. A stable approximation of the chain has been derived and the limiting cases are discussed. An intrinsic interconnection of constructive, sequential, and evolutionary approaches in related optimization problems provides new challenges for future work.

Index terms: decision making with limited information, optimal control theory, hyperbolicity of dynamic rules, generalized dynamic systems, Markov Chain approximation.

1 Introduction

Many mathematical problems in information theory and optimal control related to dynamic system studies can be formulated in the following generic form. A decision maker (DM, i.e. problem solver, modeler or observer) receives information about a system from observations, measurements, or computations in the form of a data stream that can be formalized mathematically as a sequence

$$(x_0, x_1, \dots). \quad (1.1)$$

We assume that such a sequence has at least two elements and that each element of the sequence is labeled by its own time t . Hence, referring to the element x_t of the sequence, we assume that the total amount of information about the system that corresponds to the time interval $(0, t)$ of its behaviour has been received, or at least can be received in principle. Under the above

*Preliminary results of this paper were first presented at the 32nd ANZIAM Conference, Masterton, New Zealand, in February, 1996.

assumptions we can introduce a set T_t of permissible strategies for each time t . Then, observing the sequence (x_0, \dots, x_t) , the decision maker can choose a strategy that is defined by the inclusion

$$s_t \in \bigcup_{\tau=0}^t T_\tau. \quad (1.2)$$

Typically we reduce the problem of constructing a map between elements x_t and s_t defined by (1.1), (1.2) to a simpler problem allowing the set of permissible strategies for all times of consideration to be fixed and to be given *a-priori*. Namely, we can idealize actions of the decision maker as follows. We can assume that the DM can select a strategy s_t at each time t from a given set U_T . Of course, the validity of such a simplification ultimately depends on the Axiom of Choice excluding the logically possible case of incomparability of two arbitrary sets that correspond to two different times t and t' [49,32]. However, on the other hand, such a simplification permits the development of a set-theoretic approach to dynamic system evolution, and simplifies the mathematical formalizations of complex optimization problems. In fact, we can introduce a loss function $l(\cdot, \cdot)$ as a function of two variables, states x_t and strategies s_t , which are both characterized by the same time t . A desire to minimize time-averaging characteristics of this function can be formalized through the optimization problem

$$F(l) \rightarrow \min, \quad s_t \in U_T. \quad (1.3)$$

Here, the objective functional F may be, for example, the Cesaro-type sum

$$F(l) = \frac{1}{T} \sum_{k=0}^n l(x_{\tau_k}, s_{\tau_k}), \quad (1.4)$$

where $\tau_k \in (0, T) \forall k \in (0, 1, \dots, n)$ and T is assumed to be given. The limiting problem in the spirit of classic ergodic theorems arises when we investigate the limit behaviour

$$\lim_{T \rightarrow \infty} F(l)$$

with $F(l)$ given by (1.4). Objective criteria may also be formulated in an integral form. For example, for

the Boltz problem in optimal control theory we have the form of the functional in (1.3)

$$F(l) = g(x_T) + \int_0^T f_0(\tau, x_\tau, s_\tau) d\tau \quad x_T \in K, \quad (1.5)$$

where $T \in (0, \infty)$ is assumed to be given, and K is a given target set¹. We can also consider a class of problems with infinite time horizon using discounting cost procedures. All these examples provide important partial cases of the general problem (1.1), (1.3).

Of course, to complete the formulation of the problem (1.1), (1.3) mathematically, we have to specify in what sense the sequence $\{x_t\}$ in (1.1) should be understood. One possible specification can be provided by an assumption that x_t may be appropriately described by a given stationary ergodic distribution. Then a typical assumption imposed on functions s_t from U_T is Lebesgue-measurability on the interval $(0, t)$. Under the above mentioned assumptions, associated theoretical issues are often addressed using the theory of Markov processes [19]. Starting from the work of Bellman [5,6], the theory has been extensively developed, and a number of efficient algorithms have been proposed. Discrete dynamic programming ideas have been essentially generalized for the continuous case during the past decades [18,19], and many new results that appeared recently indicate the continued research interest in these topics [19,35]. It should be noted, however, that many results in this area rely (explicitly or implicitly) on the assumption that a measurable function of strategies $s_t \in U_T$ may be effectively approximated using past states $x_{t'}, 0 < t' < t$. If such an assumption is made, the attainability of the minimum in (1.3) becomes the subject of a corresponding smoothness assumption on the loss function [40]. On the other hand, regularity of this function is strongly dependent on complete information about the past states, and eventually on model data and parameters. Since the initial data for the model can be only known approximately, the whole stream of information available to the decision maker at time t can be interpreted, at best, as an approximation of system dynamics. The quality of such an approximation at time t is defined by the "informational completeness" of the data stream

$$(s_0, x_0; s_1, x_1; \dots; s_{t'}, x_{t'}; \dots) \quad (1.6)$$

when $t' \rightarrow t$. To complete the step corresponding to time t in this process, one can assume that the strategy s_t may be chosen from the same set U_T . Then, the next stream element x_t may be received with a given accuracy, at least in principle, if we also assume that element x_0 in (1.1) may be given with infinite precision. Of course, in the reality of mathematical modelling the

¹The functions f_0 and g are called running and terminal costs respectively. If $f_0 \equiv 0$ we have Mayer's problem whereas for $g \equiv 0$ the problem is referred to as the Lagrange problem.

latter assumption cannot be rigorously justified [43]. However, if strategies are chosen at each step to satisfy a certain subgoal, the described process provides the possibility of evaluating the quality of satisfaction of a subgoal that corresponds to time t . If the process is finite then we can refer to the last subgoal as a top-level goal [33]. The latter can be satisfied by satisfying subgoals at each step appealing to multicriteria analysis of the underlying problem.

The main problems in such analysis stem from the coupling of the sequence of subgoals to the definition of the top-level goal in the form of a functional of the loss function $l(\cdot, \cdot)$. Mathematically speaking, we should be able to define a mapping between fixed-time subgoal functions and an averaged-time goal functional. Such a definition is closely connected with the definition of optimal strategies which we do not know *a-priori*. However, if it is known that $s_t \in U_T$, then it is reasonable to choose strategies based on knowledge not only of time t , but also on states x_t . If we assume further that x_t "accumulates" all past information about the system, then the concept of a Markov Chain comes by itself. Because of uncertainty in knowledge base (1.1), such an accumulation cannot be understood in a purely deterministic way [8]. The origin of such uncertainty is induced by the strategy s_0 in the data stream (1.6). However, mathematically such uncertainty can be formalized if instead of constraints (1.2) we consider "relaxed" constraints

$$s_t \in U_T, \quad (1.7)$$

assuming that the set U_T is given *a-priori* for the whole time-set of interest. Then, instead of the data stream (1.6), we can consider an *informationally reduced stream*:

$$(x_0, (s_1, x_1); \dots; (s_{t'}, x_{t'}); \dots), \quad (1.8)$$

where all strategies satisfy the constraints (1.7). An additional assumption of continuity of the sequence (1.1) in time allows a convenient mathematical framework for justification of models based on an approximation of (1.6) by (1.8). Such a classical idealization of temporal evolution by continuous trajectories of phase points, induced by classical mechanics, can be applied only within certain limited contexts, and involves serious difficulties in many areas of mathematical modelling. The main problems are caused by the fact that there are many dynamic systems for which arbitrary close initial conditions can give rise to qualitatively distinct (including exponentially diverging) types of trajectories [43]. Such strong trajectory instability requires other approaches in the description of dynamic system evolution. Under a probabilistic approach, deterministic invariance of phase points along trajectories is replaced by the invariance of the density along trajectories. Physically, such a "conservation of extension in phase" (due to J. Gibbs [37]) eventually requires a con-

struction of Gibbs distribution functions using a probabilistic description of states. Mathematically speaking, this problem can be seen as a problem of a "closure" of the reduced informational stream (1.8) with respect to all possible states. Such a closure can be performed if we assume Lebesgue integrability of the function

$$\eta(\omega) = \begin{cases} -\omega \log \omega, & \omega > 0 \\ 0, & \omega = 0, \end{cases} \quad (1.9)$$

over the set Σ of all possible states, where $\omega = f(t, x_t)$ is the density function. From an information theory perspective, this logical step, which in the end requires answering the question of system stability, is equivalent to a transformation from the classic Shannon entropy [51,47] to the Boltzmann-Gibbs entropy [37]. Under such a transformation we formally identify a (thermo)dynamical system with a measure space [37]. If Σ is fixed and the measure is defined as a Lebesgue measure, then for any time-set $(0, T)$ (including the possibility of $T \rightarrow \infty$) the validity of the above transformation requires an *a-priori* assumption of lower semi-continuity [53] of the recursive function

$$\zeta(\omega) = f_n(f_{n-1}(\dots f_1(\omega)\dots)) \quad (1.10)$$

as a function of density, where a theoretical possibility of $n \rightarrow \infty$ is permitted. If we assume that such a function exists, then in principle, the only possible uncertainty in the model (1.3), (1.8) for any $t = T$ is induced by the definition of x_0 and $\zeta(f(x, T))$. Such is indeed the case in optimal control theory where the recursive function ζ plays the role of the value function. In fact, if we know *a-priori* that the top-level goal can be described appropriately by a continuous function $F(l)$, then the associated optimal control problems can be studied through a nonlinear backward evolution PDE known as the Hamilton-Jacobi-Bellman equation with Cauchy-type terminal conditions ([11,19] and references therein). If an algorithm for the numerical solution of the latter problem exists, it can be in principle represented in the form of the informational stream

$$((\zeta_{x_T}, s_T); (\zeta_{x_{T-\Delta t}}, s_{x_{T-\Delta t}}); \dots (\zeta_{x_t}, s_{x_t}); \dots), \quad (1.11)$$

when $t \rightarrow 0^+$ and $\Delta t > 0$. The main theoretical difficulty in the rigorous justification of algorithmic rules constructed according to (1.11) is the existence of the limit of s_x , when $t \rightarrow 0^+$. If we assume that such a limit exists, then we should be able to evaluate the quantity

$$s_0 = \lim_{t \rightarrow 0^+} s_{x_t} \quad (1.12)$$

on the basis of ζ_{x_T} (which is assumed to be given) and some logical rules. In reality, the recursive function of density (1.10) at a fixed moment of time may be given only approximately. Such an approximation defines a degree n of the underlying recursion (1.10), and in turn defines a basic structure of a finite lattice on which the system dynamic can be approximated [14].

Hence, in general, information on an approximation of the same dynamic system can be provided in two possible ways:

- using the sequence (1.1), and
- using a subsequence of (1.11) that is $(\zeta_{x_T}, \zeta_{x_{T-\Delta t}}, \dots)$.

Due to intrinsic uncertainty in the definitions of x_0 and ζ_{x_T} , neither of these approximations considered separately from the other can guarantee the adequacy of the approximation to the real system. However, we can draw certain conclusions on the system dynamics by analyzing both of the sequences *simultaneously*. The complexity of such analysis is due to the necessity of a coupled investigation of the same system in two different scales. Mathematically, such scales are induced by the two limiting types of system behaviour with respect to the time-component: $t \rightarrow \infty$ and $\Delta t \rightarrow 0^+$. They are connected by the definition of the recursive degree for the system density, and ultimately, on the definition of the top-level goal in (1.3). Splitting up such a goal into subgoals provides an efficient method for the analysis of the system dynamics. In turn, such analysis gives a way to derive a sequential approximation of the system Hamiltonian, ensuring a stable model of system dynamics.

The remaining part of the paper is organized as follows. In section 2 basic preliminaries are recalled for the formulation of optimal control problems as problems in information theory. Section 3 is devoted to consideration of deterministic and stochastic dynamic rules. Examples are given to show that if such rules are specified, then an informationally consistent formulation of control problems requires an analysis of system stability. Section 4 deals with deterministic and probabilistic algorithmic machines and analyzes problems involved in their application. Section 5 gives a link between the questions discussed in the previous sections and discrete optimization problems using their common physical and informational basis. In sections 6 and 7 mathematical models are constructed and computational models derived to analyse dynamic system evolution using the Markov Chain approximations. A stable approximation for the hyperbolic model is obtained and the algorithm has been given. Computational aspects of Discrete Markov Decision Processes are discussed in section 8. The main conclusions are summarized in section 9.

2 Preliminaries.

Let us define the state space of the system by Σ and the Borel σ -algebra induced ² by Σ as $\mathcal{B}(\Sigma)$. Then, no matter what the time-partition in $[0, t]$ is, $0 \leq \tau_1 <$

²the least σ -algebra that contains all open subsets of Σ

$\tau_2 < \dots < \tau_n < \tau$, $\tau \in (0, t)$, we assume that $\forall X \in \mathcal{B}$:

$$P(x_\tau \in X | x_{\tau_1}, \dots, x_{\tau_n}) = P(x_\tau \in X | x_{\tau_n}) \quad (2.1)$$

almost surely³. That is, the data stream x_t under the strategy of the time partition has the Markovian property. Of course, continuity of the data stream x_t in t does not follow from the condition (2.1). Furthermore, even if x_t is a continuous function of time, it does not, on any account, mean that strategies form a continuous function of time as well. In general, we have a multicriteria optimization problem induced by the partition of time and the analysis of the sequence (1.6). However, the difficulty in evaluating the limit (1.12) prompts several ways to further simplify the problem. One of the direct ways is to assume a-priori continuity of the sequence (1.1) in time. Then we can reformulate the multicriteria optimization problem arising in analysis of (1.6) as an optimal control problem (1.3) with respect to a continuous function of time $F(l)$ and some dynamic rules that define the sequence (1.1).

Alternatively, we can analyze the sequence (1.6) using Discrete-Markovian-Decision-Processes (DMDP). The theory of DMDP is well-developed under the assumption of the possibility of complete information in (1.6). During recent years new challenging problems have stimulated further development in the theory of DMDP [34,25,17]. In brief, one of the most interesting problems in this field is induced by the question of data perturbations in the informational stream (1.6). Indeed, when perturbations of a Markov Chain change its ergodic structure, the stationary distribution of the perturbed system may not be a continuous function [50,1]. Hence it is reasonable to assume that system dynamics depend on some parameters of the Markov Chain and due to the imprecision of available information we can study system dynamics using in general Singularly Perturbed Markov Chains. In this framework evolution of a system is coupled to its Markov Chain parameters. An example of this type DMDP was provided in [13] where non-diffusion stochastic models were studied. We assume that in general the parameter of the Markov Chain is allowed to jump, and the jumping rate may be dependent on the state function x_t . The corresponding systems described by x_τ at time τ are called piecewise-deterministic stochastic systems. Such systems have been extensively studied during recent time by theoretical physicists [29], and indicate growing interest in hyperbolic dynamic rules of nature [44,30].

Mathematically speaking, we define a finite-state Markov Chain μ_τ with the state space \mathcal{M} . The chain is regarded as a parametric process for the dynamics of the system which is described by a state function x_τ and a parameter μ_τ . The parameter μ_τ may undertake

³with respect to corresponding σ -algebra [19]

a jump on the interval $(0, t)$ at times $\tau_1 < \dots < \tau_n$, and the jumping rate is a function of time τ , state of the system x_τ , the “before-jump” value of the parameter μ_1 and the “after-jump” value of the parameter μ_2 of the Markov chain. Hence we define a function of jump rates as

$$j \stackrel{\text{def}}{=} j(\tau, x_\tau, \mu_1, \mu_2). \quad (2.2)$$

It allows us to regard the process (x_τ, μ_τ) as a Markov process with the state space $\tilde{\Sigma} = \Sigma \otimes \mathcal{M}$. It should be emphasized that the system itself x_τ may not have Markovian behaviour. Thus, difficulties arise in constructing a mapping that relates the function (2.2) to states x_τ of the system. Ultimately, such difficulties stem from the problem of mathematical formalization of the concept of perturbations, which are usually regarded as a small and external-to-the-system source. Of course, in the real world modelling, statistics of the source is unknown a-priori, which precludes assumptions based on an ϵ -additivity of perturbations. In general, such assumptions may not be adequate for the transition law of the Markov Chain as well as for the Hamiltonian of the system as a whole.

3 Dynamic Rules and Control Problems.

Eventually, due to the approximate character of available information about the informational stream (1.6), any mathematical model can provide at best a description of a perturbed rather than an unperturbed dynamic system. Hence, if the mathematical model of a dynamic system has been constructed, in derivation of a computational algorithm we should adapt the choice of strategies s_t in our approximation of (1.6) to the character of such perturbations. Another way of putting it is that the model and the algorithm should be informationally consistent, reproducing the informational stream (1.6), and giving an approximation with a reasonable degree of accuracy.

3.1 Differential equations and inclusions.

To include the possibility of perturbations into models let us start from the definition of a mapping

$$f(t, x_t, s_t) : T \otimes \Sigma \otimes U_T \rightarrow \mathcal{R}, \quad (3.1)$$

where T is a given set of time. When x_t is assumed to be continuous the dynamics of a deterministic system can be appropriately described in almost everywhere sense by the differential equation

$$x'_t = f(t, x_t, s_t), \quad x|_{t=0} = x_0^t, \quad s_t \in U_T, \quad (3.2)$$

where x_0^t is an element of a given set X_t , defined as an ϵ -neighbourhood of an idealized point x_0 . In general, the mathematical model (3.1), (3.2) can provide a description of a perturbed rather than unperturbed dynamic system. This is the case even if we formally

exclude s_t from the right hand part of the model or introduce some optimizing criteria. The next example is to demonstrate the possibility of instability in the perturbed model under any arbitrary small level of perturbations.

Example 3.1. Let us analyze unperturbed and perturbed dynamics of a homogeneous linear system:

$$(a) \dot{x} = Ax, \quad (b) \dot{x}_\epsilon = A_\epsilon x_\epsilon. \quad (3.3)$$

Here we assume that the matrix A is given and $A_\epsilon = A + \Delta$, whereas $\|\Delta\| \leq \epsilon \|A\|$ is the absolute error for perturbations of the matrix elements. If we assume that the initial conditions for the model (3.3) may be given precisely, then the problem of stability for the model is equivalent to the investigation of the ϵ -spectrum of the original matrix A . The ϵ -spectrum of a matrix is defined as the union of all spectra of perturbed matrices for a certain level of error [23]. In general, for any arbitrary matrix A there exists a special connection between its spectrum and its resolvent under ϵ -perturbations. The problem consists of the fact that without restrictions on ϵ , an absence of practical dichotomy can be anticipated. More precisely, there might exist such $\epsilon = \epsilon(\delta)$ that A_ϵ with $\|\Delta\| \leq \epsilon$ can have in the left-half plane the number of eigenvalues different from the number of points of the matrix A spectrum. If the matrix A is defined as follows

$$A = (a_{ij}) = \begin{cases} -1, & j = i \forall i = 1, 2, \dots, 20, \\ 10, & j = i + 1 \forall i = 1, 2, \dots, 19, \\ 0, & \text{otherwise}, \end{cases}$$

and the matrix of perturbation is defined as

$$\Delta = (\delta_{ij}) = \begin{cases} \epsilon = 10^{-18}, & i = 20, j = 1, \\ 0, & \text{otherwise}, \end{cases}$$

then though the matrix A has one negative eigenvalue -1 of multiplicity 20, the eigenvalues of the perturbed matrix ($\sqrt[20]{10}-1$) lie in the right-hand plane, indicating instability in the perturbed model. Of course, similar examples can be constructed for any $\epsilon > 0$ no matter how small it is assumed. ■

We note that example 3.1 deals with the perturbation of the right-hand part of the model, but not with the initial condition. The latter was assumed to be fixed for both perturbed and unperturbed models. The idea of "frozen" initial conditions for a family of the perturbed right-hand parts leads to the mathematical models in which dynamic rules are defined by differential inclusions. In fact, on the basis of the point-valued map f , we can define a set-valued map [2,19]

$$\mathcal{F}(t, x_t) \stackrel{\text{def}}{=} \{f(t, x_t, s_t)\},$$

where s_t is assumed to be defined by another set-valued map. Of course, the set-valued map for the definition

of s_t is coupled to the definition of the optimizing functional $F(l)$ in (1.3). Hence, when describing dynamic rules by the differential inclusion

$$x'_t \in \mathcal{F}(t, x_t) \quad (3.4)$$

in an almost-everywhere sense, a family of perturbed mathematical models (1.3), (3.4) defines an optimal control problem. In the models of this type we have a natural contradiction. On the one hand the quality of this model has to be defined with respect to the stability of the system dynamic. On the other hand, such stability depends on the definition of s_t , which is an unknown function in the mathematical model. Hence, eventually the quality of the model depends on the definitions of the mapping (3.1) and initial conditions. In the end such definitions depend on the problem of evaluating the limit (1.12). If the initial conditions of the model are fixed, then an example of instability for the mapping (3.1) may in principle be constructed for any specified sequence s_t . This type of instability is usually referred as computational instability. The example 3.1 clearly shows that theoretical issues of stability should primarily be addressed if "precise" initial conditions are assumed. In optimal control theory we do not require the sequence s_t to be specified explicitly, and therefore, the problem of the model stability can be formally circumvented by some appropriate regularity assumptions on the mappings \mathcal{F} and F . The remaining theoretical problem is to prove that if the mapping (3.1) is well-defined then $s_0 \in U_T$, where s_0 is defined by the limit (1.12), whereas x_0 may not be given precisely. The complexity of this problem led to the constructing mathematical models of optimal control using recursive functions of density (1.10). In theory such approaches require analysis of a subsequence of (1.11) that consists of the values of the recursive function ζ

$$(\zeta_{x_T}, \zeta_{x_{T-\Delta}}, \dots, \zeta_{x_1}, \dots), \quad (3.5)$$

when $t \rightarrow 0^+$. Such analysis is typically performed for $\Delta t \rightarrow 0^+$, and essentially uses the assumptions that x_0 and ζ_{x_T} in (1.1), (3.5) may be given either precisely, or at least with equal probabilities.

First let us consider a deterministic optimal control problem where ζ_{x_t} plays the role of the value function. For the Boltza problem (1.3), (1.5), (3.1), (3.4) we can introduce the performance measure

$$J(t, x_t, s_t) = \int_t^T f_0(\tau, x_\tau, s_\tau) d\tau + g(x(T)). \quad (3.6)$$

If we define the value function as

$$V(t, x) \stackrel{\text{def}}{=} \inf_{s_t \in U_T} J(t, x_t, s_t), \quad (3.7)$$

then using appropriate regularity assumptions and dynamic programming principle [19,20], the original optimal control problem can be studied through the

Hamilton-Jacobi-Bellman (HJB) equation

$$\begin{aligned}\dot{V}(t, x_t) + H(t, x_t, D_x V(t, x_t)) &= 0, \\ V(T, \cdot) &= g(\cdot),\end{aligned}\quad (3.8)$$

where the Hamiltonian H is defined as

$$H(t, x_t, \delta) \stackrel{\text{def}}{=} \sup_{s_t \in U_T} \{-\delta \cdot f(t, x_t, s_t) - f_0(t, x_t, s_t)\}. \quad (3.9)$$

The rigour in mathematical justifications of the models (1.3), (1.5), (3.1), (3.4) and (3.6)-(3.9) is grounded in the following logical rule. Provided x_0 is given precisely, the forward-evolution model (1.3), (1.5), (3.1), (3.5) can be studied through the backward-evolution model (3.6)-(3.9) for any given function g from a specified topological space. The definition of topology for such a space requires the definition of a set in which physical states of the system can be embedded. Mathematically, the problem is usually considered with respect to Euclidean spaces (either finite dimensional [19] or infinite dimensional [28]). It allows us to use the logical rule in the reverse order: provided g is specified in a topological space, the backward-evolution model can, in principle, recover the forward-evolution of the system for any given initial condition x_0 .

We note that the definitions of x_0 and g are coupled to the definition of the system Hamiltonian by the specification of a topological space. An assumption that the topological space satisfies the Hausdorff separability axiom allows us to complete the chain of logical arguments in the mathematical justification of the original optimal control problem. The only problem remaining with such reasoning is that of system stability. This question is associated with the question of stability of measures defined with respect to the system's state-space, which is typically a-priori assumed to be Hausdorff. Formally, this assumption corresponds to the choice of such a function ζ in (1.10) for which $n \rightarrow \infty$. Therefore, eventually the quality of the backward evolution model (3.6)-(3.9) depends on the definition of a set X_ϵ from which we "puncture" a point x_0 when $\epsilon \rightarrow 0^+$. In the end, the question is reducible to the existence of an optimal strategy s_0 for such an operation, and evaluation of the limit (1.12). Since such a strategy is known neither with a deterministic certainty nor with the probability 1, it is reasonable to estimate the quality of the backward-evolution models with respect to a set X_ϵ , where ϵ may be small, but always assumed to be positive. Then the model (3.6)-(3.9) cannot be considered other than a *perturbed mathematical model*. Since $\epsilon > 0$, the instability of the system can be anticipated, unless the strategies from the set U_T are chosen consistently with the states of the system from the set Σ . Such consistency is defined by the definition of the system Hamiltonian in a chosen topological space, which is eventually defined by the

mapping (3.1). In this sense the Hamiltonian can be regarded as a higher degree recursion of this mapping. Since the function $f(t, x_t, s_t)$ may be discontinuous in general, so may the Hamiltonian function, unless it can be represented as an infinite degree recursion of f . The assumption of positiveness for ϵ precludes such a situation, which seems to correspond to all physically conceivable situations. However, it implies a hyperbolicity in the underlying mathematical model [44,30]. The hyperbolic nature of mathematical models in optimal control theory caused by the splitting the informational string (1.6) into two: (1.1) and (3.5). A simultaneous consideration of these strings implies their approximation by the perturbed informational strings

$$(x_0^\epsilon, x_1^\epsilon, \dots, x_{t'}^\epsilon, \dots), \quad (3.10)$$

$$(\zeta_{x_T}^\epsilon, \zeta_{x_{T-\Delta t}}^\epsilon, \dots, \zeta_{x_t}^\epsilon). \quad (3.11)$$

After the approximation, neither of the two equalities

$$\lim_{\epsilon \rightarrow 0^+} \lim_{t' \rightarrow T} x_{t'}^\epsilon = \lim_{t' \rightarrow T} \lim_{\epsilon \rightarrow 0^+} x_{t'}^\epsilon \quad (3.12)$$

$$\lim_{\epsilon \rightarrow 0^+} \lim_{t \rightarrow 0} \zeta_{x_t}^\epsilon = \lim_{t \rightarrow 0} \lim_{\epsilon \rightarrow 0^+} \zeta_{x_t}^\epsilon, \quad (3.13)$$

can be guaranteed in general. The lack of equalities in (3.12), (3.13) is caused by possible singularities in transformations from s_0 to x_0^ϵ and from s_T to $\zeta_{x_T}^\epsilon$. Nevertheless, for any arbitrary $\epsilon > 0$, the informational string (1.6) can be eventually approximated as

$$(s_0, x_0^\epsilon; \zeta_{x_0}^\epsilon; s_1, x_1^\epsilon; \zeta_{x_1}^\epsilon; \dots; s_{t'}, x_{t'}^\epsilon; \zeta_{x_{t'}}^\epsilon; \dots), \quad (3.14)$$

when $t' \rightarrow t$, $\forall t \in (0, \infty)$. Hence, the quality of approximating (1.6) by (3.14) is defined by the sequential character of approximation for the function ζ , which in optimal control theory plays the role of the value function that depends on an approximation of the system Hamiltonian (or Lagrangian).

3.2 Stochastic rules.

Let us consider a dynamic system described in terms of the stochastic differential equation

$$\begin{aligned}dx &= f(\tau, x_\tau, s_\tau)d\tau + \sigma(\tau, x_\tau, s_\tau)d\omega(\tau), \\ x(0) &= x_0,\end{aligned}\quad (3.15)$$

where f and σ in (3.15) denote drift and diffusion terms respectively, and ω is a Wiener process. As a functional F in (1.3) we choose:

$$F(l) = E_{tx} \left\{ \int_l^T f_0(\tau, x_\tau, s_\tau)d\tau + g(x(T)) \right\}. \quad (3.16)$$

Then the problem is to find

$$\inf_{U_T} F(l), \quad (3.17)$$

where $F(l)$ is defined by (3.16) under the dynamic rules (3.15), and (3.17) provides a typical example of

a stochastic optimal control problem. The use of the Bellman's principle

$$V(t, x) = \inf_{U_T} E_{tx} \left\{ \int_t^{t+\Delta} f_0(\tau, x_\tau, s_\tau) d\tau + V(t + \Delta, x(t + \Delta t)) \right\} \quad (3.18)$$

can formally reduce the problem to the dynamic programming equation

$$\begin{aligned} \min_{x_t \in U_T} [\Lambda_{x_t} V(t, x) + f_0(t, x_t, s_t)] &= 0, \\ V(T, \cdot) &= g(\cdot). \end{aligned} \quad (3.19)$$

The definition of the value function in (3.18) is analogous to that in (3.7) when we consider the conditional expectation of the performance measure (3.6). Note also that in the equation (3.19) the linear operator of backward evolution Λ is well-defined only if the limit

$$\Lambda V(t, x) = \lim_{\Delta \rightarrow 0^+} \frac{E_{tx} V(t + \Delta, x_{t+\Delta}) - V(t, x_t)}{\Delta} \quad (3.20)$$

exists for each $x \in \Sigma$ and $t \in I \subset [0, T]$, except of $t = T$ itself. In the end, the existence of the limit (3.20) is subject to the definition of $V(0, x_0)$. As in the deterministic case, such a definition depends on the definition of a set X_ϵ , and thus eventually requires the definition of s_0 . To put it differently, for a justification of the limit in (3.20) we need existence of two limits induced by (1.10) and (3.11), namely

$$\lim_{n \rightarrow \infty} f_n(f(t, x_t)) \text{ and } \lim_{\epsilon \rightarrow 0^+} \zeta_x^\epsilon, \quad \forall t \in [0, T].$$

The latter may be assumed *a-priori* rather than justified rigorously. However, even under such an assumption the procedure of transformation from the model (3.15)-(3.17) to the model (3.19), (3.20) remains an essentially sequential heuristic procedure.

The heuristic nature of the model (3.19), (3.20) can be circumvented by using the diffusion approximation method for the original optimal control problem (3.15)-(3.17). As a result, we arrive at the form of HJB equation

$$\dot{V} + H(t, x_t, D_x V, D_x^2 V) = 0, \quad V(T, \cdot) = g(\cdot), \quad (3.21)$$

where the Hamiltonian H is defined as

$$\begin{aligned} H(t, x_t, \delta, \Pi) &\stackrel{\text{def}}{=} \sup_{s_t \in U_T} \{-\delta \cdot f(t, x_t, \delta) - \\ &\quad \frac{1}{2} \text{tr}[\pi(t, x_t, \delta)\Pi] - f_0(t, x_t, \delta)\}. \end{aligned} \quad (3.22)$$

Here $\pi = \sigma\sigma'$, and Π is a symmetric nonnegative definite matrix (for details, see [19]). Note that a reduction of the problem (3.15)-(3.17) to a partial differential equation by the rescaling of a Markov Chain is accompanied by a loss of information about the dynamic system itself. Indeed, the original dynamics x_t intrinsic

to the model may or may not be Markovian in general. Though the Markovian property has to be preserved for the process (s_t, x_t) , it may be violated after the rescaling procedure, which requires a conservation of the Markovian structure from x_t .

3.3 General rationale for the optimization of singular perturbed dynamics.

For all described dynamic rules, regularities of mappings that define the Hamiltonian of the system and the value function are coupled by a specific mathematical model, and eventually depend on the topology of the space (in which investigation of the model is being conducted) and the initial conditions of the model. In principle, a-priori regularity assumptions on the Hamiltonian allow the recovery of information about the regularity of the sought-for solution. Results of this type provide a rigorous mathematical justification of the models for which the form of the Hamiltonian is specified. During the past years the theory has been extensively developed in this direction for deterministic and stochastic optimal control problems (see [11,45,28,19] and references therein).

Since the Hamiltonian of the system can be given only approximately, whereas regularity for the sought-for solution is not a-priori knowledge being the subject of our assumptions, it seems to be reasonable to couple the model and algorithm for its solution using an approximation of the informational string (1.6). Mathematically speaking, we do not assume a-priori "smoothness" of the "transition" between s_t and x_t^ϵ for an approximation of the informational stream (1.6), even if $\epsilon \rightarrow 0^+$. It implies a consideration of *singular stochastic problems* in which the function x_t^ϵ is allowed to be discontinuous (the first problems of this type were studied in [3,4]). In general, since a "transition" between s_t and x_t^ϵ ($T \in (0, \infty)$) may be discontinuous, we cannot use the *principle of smooth fit* (see [52] and references therein) to claim continuity of the recursive function of density ζ_x^ϵ , when $t \rightarrow T$ (possibly $T \rightarrow \infty$). If our objective is a probabilistic attainability of the following limits

$$\lim_{\epsilon \rightarrow 0^+} x_t^\epsilon = x_t, \quad \lim_{\epsilon \rightarrow 0^+} \zeta_x^\epsilon = \zeta_x, \quad (3.23)$$

then regularities of the limiting functions x_t and ζ_x become the subject to our *a-priori* assumptions, which in turn bring the possibility of singularities in such dynamic processes as "strategy-state" (s_t, x_t) and "strategy-state-density" $((s_t, x_t); \zeta_x)$. It reduces the problem of analysis of the sequences (1.1) and (3.5) to the analysis of the perturbed informational strings (3.10), (3.11), which formally allows us to include the parameter of perturbation ϵ into the model. We can assume, for example, that the dynamics of the system can be effectively described by "fast" and "slow" com-

ponents [57]:

$$\begin{cases} \epsilon \dot{z} = f_1(z_t, y_t, s_t, t, \epsilon), & z(0, \epsilon) = z_0, \\ \dot{y} = f_2(y_t, z_t, s_t, t, \epsilon), & y(0, \epsilon) = y_0. \end{cases} \quad (3.24)$$

If we choose a functional F in (1.3) as

$$F(l) \stackrel{\text{def}}{=} J_\epsilon = g(y_T, z_T) + \int_0^T f_0(\tau, y_\tau, z_\tau, s_\tau) d\tau, \quad (3.25)$$

then the problem (1.3), (3.24), (3.25) is an optimal control problem for the singular perturbed dynamics. In general, neither y_t nor z_t are required to have the Markovian property. The role of the string (s_t, x_t) in this case plays that of the sequence $(s_t, (y_t, z_t))$, in the sense that the sequence (y_t, z_t) is dependent on Markov Chain parameters, and thus the whole process $(s_t, (y_t, z_t))$ can be seen as a Markov Chain approximation. We can also interpret the sequence (y_t, z_t) when $\epsilon \rightarrow 0^+$ as the definition of a recursive function of density ζ_x , with increasing degree of recurrence as $n \rightarrow \infty$. Then the model (1.3), (3.24), (3.25) will be well-defined if we define a set X_ϵ of initial conditions with a specified level of error. Hence, as above, the definition of the pair (y_0, z_0) is eventually dependent on the definition of s_0 in the informational string (1.6). It implies an approximation of the informational string (1.6) induced by singular dynamic rules using sequential decision schemes.

4 Algorithmic Machines.

- Probabilistic finite-state finite-action machines under singular perturbation.

First, let us consider a probabilistic finite-action machine that analyzes a Discrete Markov Decision process. Mathematically, the analysis can be formalized as a set of four-tuple

$$\mathcal{M} = \{x_t \in X; \bar{s}_t \in \mathcal{U}; \gamma_t \stackrel{\text{def}}{=} \gamma(x_t, \bar{s}_t); p_{tt'}^\epsilon \stackrel{\text{def}}{=} p(x' = x_{t'}, |(x_t, \bar{s}_t)), x' \in X, t' \geq t\}, \quad (4.1)$$

where $p_{tt'}^\epsilon$ is the perturbed probability of the transition from the state x_t to the next state x' , γ_t is an immediate reward, \mathcal{U} is a finite set of actions, X is a finite set of states, and T is a set of all times for which states from X are realizable. In general, the disturbance law of the transition probabilities in (4.1) is not known a-priori. We may assume, however, that

$$\sum_{x' \in X} p(x'| (x_t, \bar{s}_t)) = \sum_{t' \in T} p(x'_{t'} | (x_t, \bar{s}_t)) = 1. \quad (4.2)$$

We also observe that every strategy s_t induces a perturbed P_t^ϵ rather than an unperturbed transition matrix. Hence, assuming the flow of time *ad-infinitum*,

we can define the Cesaro-type limit matrix

$$P^\epsilon(\alpha) \stackrel{\text{def}}{=} \lim_{t \rightarrow \infty} \frac{1}{t} [P_0^\epsilon + \sum_{k=1}^n P_{\tau_k}^\epsilon(\alpha)], \quad (4.3)$$

where $0 \leq \tau_1 \leq \tau_2 \leq \dots \leq \tau_n < t$ with the possibility of $n \rightarrow \infty$. A strategy α in (4.3) denotes a sequence that consists of elements s_t . Of course, using the reward function $\gamma(\cdot, \cdot)$, we can construct classes of optimization problems in a similar way to what we have done with respect to the loss function in section 1. For example, we can consider the limit Markov control problem

$$J^\epsilon(\bar{s}, s_t) \rightarrow \max s_t \in U_T, \quad (4.4)$$

where

$$\begin{aligned} J^\epsilon(\bar{s}, s_t) \stackrel{\text{def}}{=} \liminf_{t \rightarrow \infty} \frac{1}{t} [E_\alpha(\gamma_0, \bar{s}) + \\ \sum_{k=1}^n E_\alpha(\gamma_k, \bar{s})], \end{aligned} \quad (4.5)$$

and $\bar{s} \in \mathcal{U}$, $\alpha \in U_T$. We note that the definition of the matrix P_0^ϵ in (4.3) and the quantity $E_\alpha(\gamma_0, \bar{s})$ in the problem (4.4), (4.5) eventually depends on our definition of the first pair (s_0, x_0) in the informational stream (1.6), which may be given only approximately. Hence, it is reasonable to assume that the transition law matrix P_0^ϵ has Markovian structure under specified n if the exact equality in (4.2) holds. To put it differently, for any finite n the structure of P_0^ϵ depends on the topological structure of sets X and T , thus when X and T are specified such dependency remains in force even if $n \rightarrow \infty$. In the general case, it precludes the definition of the matrix P_0^ϵ as a fixed finite dimensional matrix with the probability 1 [16]. As a result, stability analysis of the associated optimization models requires consideration of a family of matrices P_0^ϵ under a specified level of error. Recall that a similar situation holds when dynamic rules are given. Then, we need the whole set X_ϵ under a specified level of error to perform analysis of stability. Without such a "relaxation" of probabilistic requirements on the initial conditions of the model, for any arbitrary small $\epsilon > 0$ an example of practical instability can always be constructed.

- Deterministic Finite-State Finite-Memory Machines.

Now let us consider another type of algorithmic machine. Deterministic finite-state machines in the case of finite memory are defined as the triple [40]

$$\mathcal{D} = (\Sigma_m, f_2, f_1), \quad (4.6)$$

where Σ_m is a finite set of machine states, and f_1 is a mapping $\Sigma_m \otimes \Sigma_m \rightarrow \Sigma_m$ which defines the machine-next-state function. The set Σ_s is a finite set of system states. More precisely, we assume that Σ_s can be formalized as a sequence (1.1) as a result of observations, computations, measurements etc. This sequence

"feeds" the machine (4.6). The mapping $f_2 : \Sigma_m \rightarrow U_T$ defines the output function with a set of strategies U_T . Hence, starting from the state $\hat{s}_0 \in \Sigma_m$, the machine (4.6) produces strategies (s_1, s_2, \dots) while going through a sequence of its states $(\hat{s}_1, \hat{s}_2, \dots)$ according to the recursive rules

$$\hat{s}_t = f_1(x_{t-1}, \hat{s}_{t-1}), \quad s_t = f_2(\hat{s}_t). \quad (4.7)$$

Excluding the current state of the machine \hat{s}_t from (4.7), we find a function of strategies as a second degree recursion of the sequence (x_{t-1}, \hat{s}_{t-1})

$$s_t = f_2(f_1(x_{t-1}, \hat{s}_{t-1})). \quad (4.8)$$

Hence, having knowledge of the previous state of the machine and a corresponding letter of the alphabet Σ , we can define the current strategy using the recursive function (4.8). This model does not require any formal association with a statistical model, and does not even assume the existence of the latter [40]. The informational data stream produced by such machine is

$$((x_0, \hat{s}_0), s_1, (x_1, \hat{s}_1), \dots) \quad (4.9)$$

From (4.9) we conclude that the starting information to compute the first strategy is a pair (x_0, \hat{s}_0) . We also observe that the main drawback of such a deterministic model is the requirement to fix the strategy immediately when the state of the machine D is given. Loosely speaking, some relaxation time between the transition $\hat{s}_{t-1} \rightarrow \hat{s}_t$ should be incorporated into the model to allow strategy correction. Indeed, such time is implemented into probabilistic finite-state finite-action machines by probabilities of the transition from one state of the system to another under certain actions of a controller or DM. However, if we know *a-priori* that

$$P(\hat{s}_{t-1} \rightarrow \hat{s}_t | x_{t-1}, (s_t, x_t)) = 1, \quad (4.10)$$

or time for such a transition is defined by a given time-interval, then the sequential decision scheme based on deterministic finite-state finite-memory machines is quite natural. If such information is not available *a-priori*, then probabilistic finite-state finite-action machines appear to be useful in the analysis of system dynamics.

In the next sections we develop a technique to find a reasonable compromise between the two approaches described above.

5 The perturbation parameter as a fuzzy border between deterministic and probabilistic descriptions of system dynamics.

Major complexity in the mathematical modelling of dynamic systems arise from the *a-priori* unknown character of the disturbance law. On one hand, the implicit

assumption of deterministic models on the existence of an associated optimal algorithm (like an assumption (4.10)) can be hardly justified in modelling complex processes and phenomena. On the other hand, the main difficulty in effective applications of probabilistic models arises from the question of how common is the ergodicity of the Hamiltonian flow on the energy surface [24]. As was pointed out, perturbations can qualitatively change the ergodic structure of the underlying dynamic system. The examples of Markov Chains with discontinuities in the stationary distribution of the perturbed system can be found, for example, in [50,1]. Furthermore, for any decomposition of such a chain into a finite number of independent ergodic subclasses (under the assumption $\epsilon \rightarrow 0^+$) examples of system instability can be constructed for arbitrary small ϵ .

5.1 Degree of recurrence in mathematical models for evolution.

An idealization of "unperturbed" mathematical models obtained in the limit of vanishing perturbations $\epsilon \rightarrow 0^+$ can often help to better understand real-world phenomena and processes. However, it should be realized that such an idealization has limited applicability, and depends on quite restrictive mathematical assumptions related to

- homogeneity of the environment of the system, and
- uniformity of density which characterizes the system or its parts.

Since for any model of a dynamic system with specified dynamic rules the parameter of perturbation ϵ may be small but always positive, rescaling procedures for the associated (with the optimization model) Markov Chain may not provide an adequate approximation to the system dynamic. Such procedures may eventually ignore the neighborhood structure of the chain. If such a rescaling (for example the diffusion approximation) has been performed, then the original problem can be reformulated as an inverse problem with respect to a recursive function of density (1.10). The complexity of the solution of the inverse problem is determined by the degree of recurrence n and the topology of the space where investigation is being conducted. Moreover, if the topology is *a-priori* specified then the regularity assumptions on the function f_n allow us to recover the information on the regularity of the function ζ , at least in principle for any arbitrarily big n , following certain logical rules. In the models like (3.8), (3.9) and (3.21), (3.22), f_n plays the role of the Hamiltonian function. Such models can be regarded as discrete optimization problems if we interpret the function f_n as one that defines the top-level goal, whereas all functions $f_i, i = n-1, \dots, 1$ are supposed to define certain subgoals. The definition of the density function pro-

vides constraints for such a problem of multicriteria optimization. From the physical point of view such problems require finding the minimum of the Hamiltonian of the system on the energy surface, and can be formulated as follows: given a finite (typically large) number n of subsystems of a big system, minimize an approximation to the system Hamiltonian on an approximating set of its energy surface.

Now recall the definition of system entropy in statistical physics as a quantity that uncertain to an additive constant and is dependent on the choice of units, defined by the Liouville measure [36]

$$\sigma = - \int f \log [(2\pi\hbar)^s f] dp dq. \quad (5.1)$$

Here s is the degree of system freedom, p and q are momentum and position variables. If we assume that the whole system entropy can be defined through the entropies of its subsystems as $\sigma = \sum_i \sigma_i$, then for any probability distribution $\mathbf{p} = (p_1, p_2, \dots, p_n)$ its associated information can be defined as the Shannon entropy [51,47]:

$$\sigma_s(\mathbf{p}) = - \sum_{i=1}^n p_i \log p_i. \quad (5.2)$$

The constant n in (5.2) can be approximated with respect to the required accuracy ϵ and is ultimately coupled to the definition of s in (5.1). In the limit of "vanishing perturbations" $\epsilon \rightarrow 0^+$ and "maximum knowledge" $n \rightarrow \infty$, the Shannon entropy can be generalized to the continuous case of the Boltzmann-Gibbs entropy. The latter transformation requires a justification of system stability. From the physical perspective mathematical idealization of two simultaneous limits $n \rightarrow \infty$ and $\epsilon \rightarrow 0^+$ requires an estimation of the degree of system freedom in the definition (5.1). In this sense such an idealization is problem specific, and always requires analysis of the measure stability.

5.2 Discrete optimization and evolution of thermodynamic systems.

Any specific algorithm for the solution of the problem of modelling dynamic system evolution is affected by the form of the function f_n (as a Hamiltonian approximation on the energy surface) and by the neighbourhood structure of the system evolution. In this sense an algorithm is always coupled to the problem specific information. In discrete optimization such algorithms can be conditionally divided into three main categories [10,48]:

- **constructive algorithms (CAs)** that require construction of decreasing and embedded in each other subsets of a given finite set of states Σ ,
- **sequential algorithms (SAs)** that attempt to construct a path through Σ , and

- **evolutionary algorithms (EAs)** that manipulate sets of solutions in Σ .

Let us assume that, for any given state x_t from Σ that characterizes the whole system, there is a neighbouring set of states N_{x_t} , where transitions from x_t are allowed. Then CAs usually apply a "greedy" policy - when starting from $x_0 \in \Sigma$, they choose at stage n an x_{n+1} such that

$$\mathcal{E}(x_{n+1}) = \min \{\mathcal{E}(t) : t \in N_{x_n}\}, \quad (5.3)$$

where \mathcal{E} is an energy functional. Mathematically speaking, we expect that given \mathcal{E} and an accuracy $\epsilon > 0$, we can find a solution, at least in principle, when $n \rightarrow \infty$. However, it is well-known that as a result of such policy CAs may relatively easily be trapped in a local minimum of \mathcal{E} . If \mathcal{E} is assumed to be continuous and Σ is a "reach" enough set, then in general the degree of recursion in (1.10) tends to infinity and we theoretically face infinitely many optimization problems (5.3). By now it is clear that without an appropriate analysis of the structure N_{x_n} , success of such algorithms cannot be guaranteed. As we pointed out earlier, such analysis has to be conducted with respect to given ϵ .

The main advantage of SAs is based on the fact that they do not exclude the theoretical possibility of occasional acceptance of new states that may increase the energy functional [41]. We also assume that an "initial" solution $x_0 \in \Sigma$ may be given (for example, obtained by a CA). Moving to a neighboring solution $x' \in \Sigma$, the structure of the neighborhood of the solution should be carefully analyzed to avoid the difficulty of CAs.⁴ The basic idea for such an analysis came from statistical physics. The growing complexity of the solution of deterministic equations of motion for a system of many subsystems (such as particles) has led to the idea of ensemble averaging instead of classic-mechanical averaging in time. As the number of subsystems increases dramatically, the Monte-Carlo and particle-type simulations [27] eventually remain the only algorithmic procedures that can be applied in theoretical generality. However, such procedures may encounter serious difficulties in non-equilibrium thermodynamics [46]. In a search for alternative approaches to the ensemble averaging, many useful ideas have been generated during recent years. The intrinsic ability of Markov Chains to form a canonical Gibbs ensemble numerically has lead to growing interest to the subject [19,35]. Using the principles of statistical physics we can assign to each state $x_t \in \Sigma$ the probability

$$p_T(x_t) = \frac{\exp(-f(x_t)/T)}{\sum_{x_t \in \mathcal{A}} \exp(-f(x_t)/T)}, \quad (5.4)$$

where $f(x_t) = \mathcal{E}(x_t)/k$. The quantity $\mathcal{E}(x_t)$ can be interpreted as the potential energy of each state (or sub-

⁴There are classic examples of SAs like the steepest-descent method that have potentially the same problems as CAs.

system) in phase space that belongs to an ensemble. The probability that a system belongs to the ensemble is proportional to $\exp[-\mathcal{E}/(kT)]$ where k is the Boltzmann constant. We observe that the smaller $T > 0$ is, the more evident is the tendency of the Gibbs distribution defined by (5.4) to be concentrated on states x_t with small values of $f(x_t)$. Hence, if we could simulate the *cooling* of the system, a state of minimum energy may, in principle, be obtained provided that the Markov Chain converges (in distribution) to the Gibbs distribution (stationary) law. This allows us to consider CAs as a partial case of this general interpretation when a Markov Chain is run for $T \rightarrow 0^+$. Another extreme case of the “high T limit” ultimately leads to the idea of *dynamic continuity*. In such a case all states are assigned the same probability, and evolution is thought as moving from a state to its neighbors *uniformly*. The computational implementation of the above idea is provided by the simulated annealing algorithm first proposed in [31]. For a real physical system, temperature may be lowered too rapidly, and the system may be trapped in a local energy minimum. However, the choice of $T_n = c/\log n$ with a sufficiently large c can theoretically guarantee the system’s “escape” from the local minimum [21]. In practice, the algorithm works as follows. If for the time-index n x_{t_n} is given, then from the set $N_{x_{t_n}}$ we choose state t , calculate $\Delta f = f(t) - f(x_{t_n})$, and set

$$x_{t_{n+1}} = \begin{cases} t, & \text{with probability } p = \exp(-\bar{\Delta}/T_n), \\ x_{t_n}, & \text{with probability } 1-p, \end{cases}$$

where $\bar{\Delta}$ is Δ when Δ is positive and zero otherwise. Of course, the choice of the *neighborhood structure* is crucially important for the algorithm’s performance. If the neighborhood is chosen too small, then the resulting simulated Markov Chain may move very slowly around Σ in the search of the minimum. On the other hand, if the neighborhood is chosen too large, then the process eventually performs a “blind” random search throughout Σ . It samples randomly from a large portion of the state space, and every next possible state is chosen practically uniformly over the whole set Σ . As an extreme case it may happen that $N_{x_t} = \Sigma$. The conclusion which has to be drawn from the above consideration is that the choice of neighborhood should be adapted to the approximation of the energy functional (or system Hamiltonian) in the search for a compromise between these two extremes.

The first step towards such an adaptation is realized in EAs. Typically, EAs deal with a population of solution instead of a single partial solution, as in CAs or SAs. The most important advantage of EAs consists of allowing an exchange of information between solutions in the current population (a cooperation step during the “generation cycle”). The main problems for EAs are related to the self-adaptation step when the solution’s

internal structure may be changed without interaction with other members of the population. When there are a lot of replicates of the same solution in a population, EAs may converge prematurely, which is usually called a diversity crisis. In such situations EAs are not competitive with the best versions of SAs.

Let us summarize the definitions of strategies in the above three classes of discrete optimization algorithms:

$$\begin{cases} s_t = \mathcal{F}_1(x_{t-\Delta t}, \mathcal{E}) & \text{for CA,} \\ s_t = \mathcal{F}_2(x_{t-\Delta t}, N_{x_{t-\Delta t}}, \mathcal{E}) & \text{for SA,} \\ (X_n^\epsilon, s_t) = \mathcal{F}_3(N_{x_{t-\Delta t}}, \mathcal{E}) & \text{for EA.} \end{cases} \quad (5.5)$$

Here $\Delta t > 0$ is a relaxation time coupled to the algorithm performance when $\epsilon > 0$, and X_n^ϵ is a population of solutions for the n th generating cycle. Functions $\mathcal{F}_i, i = 1, 2, 3$ are algorithm-specific. In general, they can be regarded as recursive functions of energy functionals, and the set of initial approximations X_ϵ for the specific algorithm:

$$F_i = f_n, (f_{n-1}(\dots(f_1(X_\epsilon, \mathcal{E})\dots))). \quad (5.6)$$

At any specified moment of time t , the definition of strategy s_t implies a coupling rule between ϵ and n_i . The definition of such a coupling leads to the well-posedness of the problem. In this sense, the well-posedness of limiting models based on the assumptions $\epsilon \rightarrow \infty$ and $n_i \rightarrow \infty$ is totally dependent on *complete information* about the initial conditions of the system, and a precise definition of the energy functional.

The process of constructing mathematical models is always a competition between (i) an approximation of the system-environment boundary interface (which involves the system’s internal time [42]), and (ii) the conservation laws for integral characteristics of the system (which involves modeler’s time [39]). As a result of such a competition, the resulting mathematical models *simulate* coupling of the system to its environment, and can be considered as models of *neither isolated nor closed systems*. A formal expression of the competition is provided by the physical concept of relaxation time. Having captured in the mathematical model the notion of information formally, its numerical expressions can be used in decision making with uncertainty, characterised by the adequacy of the simulation of the system-environment coupling. In general, numericological methods can be used effectively only if an appropriate model has been constructed. Hence, the quality of an algorithm depends decisively on an adequate reflection of the system-environment coupling in the mathematical model. If constructing a model is an art rather than a science, then the latter formally begins from the derivation of an algorithm from the model [54].

In concluding this section, it should be emphasized that the quality of a mathematical model for dynamic sys-

tem evolution is decisively dependent on (i) the approximation of the initial conditions for the system, and (ii) the approximation of the system-environment boundary interface. To minimize such dependency, the solution of a sequence of optimization problems can be used as an alternative to the limiting rescaling procedures approach. Such an approach seems to be more physically reasonable, since *a-priori* information about the system can be given only as a certain probabilistic distribution which allows us to select a new distribution according to certain principles [15,47].

6 Coupled Mathematical Models of Macro and Micro evolution.

The complexity in identifying a "hard boundary" interaction between system and its environment is eventually determined by the degree of recurrence in the definition of the system Hamiltonian. Such a definition should be given with respect to the upper bound of error ϵ in the identification of the set of initial conditions X_ϵ . Since, in general, perturbed and unperturbed models might give rise to *qualitatively distinct types of descriptions of system behaviour* for any arbitrary $\epsilon > 0$, the perturbation parameter alone cannot be an appropriate characteristic of the model's uncertainty. We observe that perturbations are an important part of the system dynamics which cannot be appropriately formalized in mathematical models unless we regard the mathematical modelling of dynamic system evolution as a *decision making process with limited information* from the very beginning of the modelling process. Additional information about the system becomes available in time at stages due to the model-associated computations, observations and measurements. Hence, to approximate the dynamic system evolution, it is essential to take into consideration the fact that initial information about the system can only be given approximately. A mathematical formalization of such approximations is a challenging problem that requires new approaches.

On one hand, the idea of sequential approximation and the hyperbolicity of the underlying differential equations is an intrinsic element of recent investigations in physics foundations [44,30]. On the other hand, rescaling procedures allow us to construct mathematical models which are essentially parabolic by their nature. Moreover, the latter have proved to be a very useful tool for investigating the laws of nature. Although such rescaling procedures are always connected with the loss of some information, a justification of parabolic approximations of dynamic system evolution may be obtained if we assume that there exists a system density f on the Gibbs phase space Γ such that its associated index of probability is given by $\log f$. In general it allows us to consider the definition of entropy

in the Gibbs form as

$$H(f) = - \int_{\Gamma} \eta(f) \mu_{\mathcal{E}}(dx_i) \quad (6.1)$$

instead of the definition (5.1), where η is defined by (1.9). Such a formal identification of a (thermo)dynamical system with a probability space is based on the Gibbs conjecture. Namely, we assume that the appropriate description of a macroscopic system in thermodynamic equilibrium may be provided by certain probability measures on the phase space of the system. Although this conjecture has never been rigorously proved [24,39], the passage from (5.1) to (6.1) is not without certain gains. It provides a convenient framework for the development of a mathematical theory for dynamic systems allowing the formulation of the concept of ergodic theory that expresses at least some aspects of irreversible thermodynamic evolution [43]. However, the introduction of a recursion function ζ using the Lebesgue measure $\mu_{\mathcal{E}}(dx_i)$ does not answer the question of stability for a "projection" of the Liouville measure (for a system with a certain degree of freedom (5.1)) onto the energy surface using a sequence of the Gibbs measures that deal with micro-canonical ensembles. As we explained above, from the physical point of view we should approximate the system Hamiltonian on the energy surface, which is also subject to an approximation. Hence, mathematically speaking, to rigorously justify models arising from application of the Gibbs conjecture, we should be able to construct both the forward-evolution model and its associate for the backward-evolution as we explained it in section 3. Gibbs was the first who arrived at the concept of mixing, and who noticed that the very use of probabilities in the description of physical states implies a time asymmetry [43]. In turn, the latter implies reversibility of distribution functions in a mathematical sense, as well as a forgetfulness property with respect to the initial conditions of the system in the flow of time. Such a reversible time-asymmetry in the mathematical theory of dynamic systems is in contrast with the irreversible character of evolution implied by the second law of thermodynamics and Eddington's time arrow. The complexity of the mathematical formalization of evolution irreversibility was well understood by J.Gibbs, who wrote [22],

it should not be forgotten when ensembles are chosen to illustrate the probabilities of events in the real world, that while the probabilities of subsequent events may often be determined from the probabilities of prior events, it is rarely the case that probabilities of prior events can be determined from those of subsequent events, for we are rarely justified in excluding the considerations of the antecedent probability of the prior events.

Almost a century ago he clearly pinpointed that the main difficulty in a mathematical formalization of the backward evolution models lies in the complexity of a probabilistic description of the initial conditions for the dynamic system, even if the probability of a terminal event is assumed to be given a-priori. At the same time he proposed an approach that allows the effective construction of a framework for a formal separation of the "observer" from the "modeler", and the system from its environment. Such a construction plays a resolving role in mathematical modelling and computational experiments. In fact, if the conjecture is accepted, the "modeler" (at least in principle) can perform a task in the "best" possible way, and the idea to exclude the "observer" from the *intermediate process of computations* (except at the very beginning and the very end of this process) becomes natural [58]. Then the whole time-set of the evolution of a dynamic system may be associated exclusively with the "modeler" as an "error-nulling" optimizing device. The existence of such a device depends on the existence of an error-free model of dynamic systems, that in turn eventually depends on the definition of a sequence of switching events or a time-partition, when the "modeler" may become the "observer" and vice versa.

Starting from this idea we can introduce the notion of a Generalized Dynamic System (GDS) where the decision maker (modeler/observer or problem solver) is considered as an intrinsic part of the model [39]. The basic steps of such a model construction are as follows: first, we consider the mathematical model of a dynamic system

$$e_{n+1} = H(v_\epsilon, e_n), \quad n = 0, 1, \dots \quad (6.2)$$

as a mapping that couples two space-time events of the system evolution by a function of the perturbed velocity v_ϵ and the system's Hamiltonian or its approximation H . Then, we specify a sequence of events (e_0, e_1, \dots) by temporal evolution. In practice such a specification is always an approximation for both the probabilistic and deterministic approaches. We assume that the basic features of dynamic rules that govern a system can be appropriately described by a velocity function v_1 . Furthermore, we allow the possibility of a "correction" of these dynamic rules by another dynamic which is specified by another velocity function v_0 . Formally, v_1 can be seen as a higher, but a-priori unknown, degree of recursion of the function v_0 . As a result, we arrive at the two coupled sequences

$$(x_0, x_1, \dots) \text{ and } (h_0, h_1, \dots). \quad (6.3)$$

When $n \rightarrow \infty$ and $\epsilon \rightarrow 0^+$ we expect that the sequences (6.3) merge, producing events that can be characterized by the limit of the model (6.2). Since neither the degree of recursion nor the level of perturbations are known a-priori, we formalize the dynamics of the system by

the two equations

$$\begin{cases} x_{t+1} = H_1(v_t, x_t), \\ h_{\tau+1} = H_0(v_0, h_\tau), \end{cases} \quad (6.4)$$

where H_1 is an approximation to H and H_0 is an operator for sequential corrections of such an approximation. If we assume that in principle system dynamics can be described arbitrary accurately, then the first equation of the system (6.4) in the long run should be practically independent on v_0 . Such a limiting case corresponds to viewing perturbations as a force, "continuously" external to the system. However, in general, both functions v_0 and v_1 are perturbation-dependent. Thus, the system (6.4) provides the possibility of looking at the coupling between the velocity of the perturbed system and perturbations of its environment. It is assumed that in general such coupling can be looked at in two different space-time frames of reference, macroscopic and microscopic.

One possible direction in the development of the theory of dynamic systems was provided by the celebrated Gibbs conjecture which we mentioned above. This led naturally to the idea of the control of dynamics described adequately (for example, in the almost-everywhere sense) by the first equation of the system (6.4) or its consequences, some of which we have considered in previous sections. Under this approach mathematical formalization of the decision rules need some a-priori assumptions on the smoothness of the function (or functions) that provides (or provide) an approximation to the recursive function H_1 . It is precisely these assumptions which formally allow the use of the perturbation theory in the investigation of underlying dynamic problems. In this way we "localize" the problem of scale interactions into a perturbation parameter ϵ which stores information about the complexity of the problem no matter how big the degree of recursion n really is. From this point of view it seems reasonable to look at the classical system of the theory of singular perturbations (like (1.3), (3.24), (3.25)) as those that may be obtained as a partial case of (6.4) by some appropriate rescaling procedures. More precisely, if ϵ is interpreted as a force, which is external to the system, then in the limit of $\epsilon \rightarrow 0^+$ the classic models in the theory of singular perturbations may be regarded as an infinite-recursion decision rule.

In the general case, however, the model (6.4) provides an interpretation of perturbations as an intrinsic to-the-system force. In this case it is reasonable to assume that both functions v_0 and v_1 are dependent on ϵ for any interval of time. Moreover, since the only available a-priori information on ϵ is its positiveness, we need to introduce a mapping to describe the behaviour of ϵ while the system evolves. To put it differently, in order to perform at least in principle an infinite-recursion procedure when $\epsilon \rightarrow 0^+$ and $n \rightarrow \infty$,

we need some *learning rules* to be introduced into the model. In [39] it was shown that under quite general assumptions the optimal control problem (1.3), (1.5), (3.2) is reducible to the hyperbolic-type equation (generalized energy equation):

$$(1 + v_1) \left[\frac{\partial \mu}{\partial x} + \frac{1}{v_1} \left(\frac{\partial \mu}{\partial t} + f_0 \right) \right] = 0, \quad (6.5)$$

that has a unique generalized solution (in the sense of an integral identity). The unknown function was assumed to be Lebesgue integrable, that is $\mu \in L^1(Q)$, where Q is the space-time region of interest. In the general case this function is referred to as the *decision maker* function. The interpretation of the equation (6.5) as a partial case of the system (6.4) can be formally given as follows. We consider a mathematical model that consists of two parts: (i) an idealized equation for a phase point in the system's time (with a trajectory $h(\tau)$) associated with the center of the system gravity, and (ii) the macro-model of dynamic system micro-evolution in the decision-maker time "external" to the system (in terms of the decision-maker function μ)⁵. Such a model of a Generalized Dynamic System couples two different space-time scales with the perturbed velocity function v_ϵ in its two different manifestations, micro-velocity v_0 , and macro-velocity v_1 :

$$\begin{cases} \dot{h}(\tau) = v_0(\tau, h, \mu), & (a) \\ \frac{\partial \mu}{\partial t} + v_1(t, x, \mu) \frac{\partial \mu}{\partial x} = 0. & (b) \end{cases} \quad (6.6)$$

Hence, the model is constructed in such a way that both parts of the perturbed velocity functions v_0 and v_1 inherit their dependency on the decision-maker function. If two events (between which GDS evolution has to be studied) are specified, then a pair of functions $(h(\tau), \mu(t, x))$ give the solution to the problem. An approximation of such events can be given using a probabilistic connection between the micro and macro levels of the system description in the form of the complementarity principle

$$v_0(h, \tau, \mu)|_{\tau=\tau_0} \cdot \mu(t, x, v_1)|_{t=t_0} = 1. \quad (6.7)$$

If the smaller velocity v_0 is assumed, then the bigger μ at the initial moment of time should be chosen. Hence, formally by (6.7), we postulate the existence of the system in a space-time of events with the probability 1 at the initial moment t_0 of absolute DM-time for any arbitrary small values of v_0 . Since τ_0 may be given only approximately, any approximation that follows from (6.6),(6.7) enables us to identify such an approximation with a Perturbed GDS (PGDS). In the limit of vanishing perturbations ($\epsilon \rightarrow 0^+$) the model (6.6),(6.7) (PGDS evolution) formally converts into the model for Unperturbed GDS (UGDS) evolution and merges with the model (6.2). Therefore, in principle

⁵We started from the consideration of the following equations $\dot{h} = v_0(\tau, h, \mu)$ and $\dot{x} = -v_1(t, x, \mu)$.

the model (6.5) can be obtained from (6.6), (6.7) using (6.7(b)) as a corrector for the equation (6.7(a)). Such a corrector induces the presence in the equation (6.5) of the goal function f_0 . The main difficulty behind such a formal procedure is how to construct an appropriate corrector. From the probabilistic point of view this difficulty was dealt with by Gibbs. Of course, there do not exist two non-identical events (related to the present state of the system evolution, and its future or past behaviour) described by any mathematical model with the same probability exactly equal to 1. In reality, all constructions of mathematical models for dynamic system evolution start from a countable base in space-time of events of PGDS evolution. At the next step, we approximate (6.2), and this "fuzzifies" the deterministic concepts of evolution in the probabilistic descriptions of events. It should be noted, however, that a randomness of GDS evolution is induced by inherent approximations in the model construction and is not an independently established fact by itself. The lack of rigour in the description of a dynamic system by purely probabilistic models stems from the fact of such an approximation. On the other hand, the main difficulty in applications of deterministic models is in the construction of effective correctors to describe adequately dynamic rules. In both situations the success of modelling is defined by the quality of an algorithm, which should be derived from the model using the concept of system stability.

7 Computational Models as Markov Chain Approximations.

As soon as dynamic rules (with or without control) define a model for system evolution as a function of time x_t^ϵ , such a function becomes subject to intrinsic uncertainty for arbitrary small intervals of time. This is a natural reflection of the approximate character of mathematical models which can be in principle characterized by the degree n of recursion for such a function with respect to the function of density. Since such a degree can be rarely given *a-priori*, we can approach the problem solution by imposing an upper bound on ϵ . It seems to be natural that in applications to the real world, mathematical models of dynamic systems have to be understood as perturbed rather than unperturbed models. Of course, they will remain as such in the foreseeable future. In general, it precludes assumptions on the forgetfulness property for density distributions, and as a result the Markovian property for the perturbed system dynamics x_t^ϵ . Behind the complexity of the problem is the question of the system's stability. The idea which will be developed in what follows is to construct a Markov Chain approximation simultaneously with an approximation of the system (that depends on Markov Chain parameters) to guarantee its stability. Hence the Markov Chain shall play the role of

a learning rule for the system under an approximation of the perturbed system's velocity by its approximation v_1 in the macroscopic DM frame of reference. As a result of such a construction and the Markov theorem on the generalized law of big numbers, the pair of functions $(h(\tau), \mu(t, x))$, which describes the process of GDS evolution, shall possess the Markovian property. Furthermore, it is proposed to approximate this process by a pair of discrete functions $(\xi_n^{\tau h}, \mu_n^{\tau h})$ ⁶, where $\xi_n^{\tau h}$ is an associated (with the microscopic frame of reference) Markov Chain state.

Let us consider the PGDS described by the form of the generalized energy equation (6.5)

$$\frac{\partial \mu}{\partial t} + v_1(t, x, \mu) \frac{\partial \mu}{\partial x} = \tilde{f}_0(t, x, \mu). \quad (7.1)$$

The approximation of the initial condition for this model is specified in the DM-time scale as

$$\mu(x, t)|_{t=t_0} = \delta(\epsilon), \quad (7.2)$$

where ϵ depends on the approximation of the function v_0 in (6.7). Hence, formally, the model (7.1), (7.2) can be seen as a macro-model for GDS evolution. However, microscopic features of the dynamics⁷ are taken into account by the possibility of coupling between the parameter of system perturbations ϵ and the decision-maker function μ . In what follows, a technique which is based on the construction of a hybrid-type algorithm [10] for the solution of this problem will be developed. The main results concern the derivation of a learning heuristic procedure that combine the effective features of (5.5), (5.6). To simplify the derivation, I explain the main ideas in the one-dimensional case, denoting a characteristic length of the system as h and assuming that $h \ll T - t_0$. Let us consider the evolution of the system defined by the dynamic rules (7.1), (7.2) in a square region of the macroscopic frame of reference

$$G = \{(x, t) : x_0^{\epsilon} \leq x < X_t, t_0^{\epsilon} \leq t < T_x\}, \quad (7.3)$$

where absolute DM-times of initial ($t_0^{\epsilon} = t_0$) and terminal ($T_x = T$) events, as well as a position $x_0^{\epsilon} = x_0$ of the system, are specified. If GDS evolution takes place in \bar{G} under a certain level of perturbations $\epsilon > 0$, then for this region the function v_1 depends on the DM-function μ . This depends on v_0 being subject to approximation from the initial moment of DM-time. Hence, we shall approximate the function v_1 with respect to our approximation of the function v_0 in a recursive manner. First we introduce the discrete grid in the region (7.3)

$$\omega_{\tau h} = \{(x_i, t^j) : x_{i+1} = x_i + h, t^{j+1} = t^j + \tau, i = \overline{0, n-1}, j = \overline{0, m-1}, t^m = T\}, \quad (7.4)$$

⁶ compared to random processes with Markov Chain parameters in the continuous absolute time in [13, 19] and references therein

⁷ induced by (i) an approximation of system-environment boundaries at $t = t_0$ and (ii) corrections of the function v_1 by v_0

and consider an elementary space-time cell $c_{ij} = [x_i, x_{i+1}] \otimes [t^j, t^{j+1}] \subset \bar{G}$. The nodes of the grid (7.4) connect events relevant to the system evolution. We shall refer to the whole set of such events in \bar{G} as a set of macroscopic events. Let t^j and t^{j+1} be two moments of absolute time (defined by DM) that correspond to two subsequent macroscopic events e_j, e_{j+1} of system evolution. Since the process (x_t, μ_t) is assumed to be Markovian, these events can be specified by two pairs of discrete functions $e_j = (\xi_j^{\tau h}, \mu(x_i, t^j)), e_{j+1} = (\xi_{j+1}^{\tau h}, \mu(x_{i+1}, t^{j+1}))$, where $\xi_j^{\tau h} = x_i^j$ and $\xi_{j+1}^{\tau h} = x_{i+1}^{j+1}$ are states of the associated Markov Chain⁸. To preserve basic macroscopic features of the system, the values of jumps $\Delta \xi_j^{\tau h} = \xi_{j+1}^{\tau h} - \xi_j^{\tau h}$ of this chain should be subordinated to the corresponding approximation of system-environment boundaries. For example, let time spent to cover the characteristic length h of the system be τ . Then, we formally express the idea of subordination in the definition which follows, where we consider the limiting case $\tau \rightarrow 0$ of such a subordination.

Definition 7.1 Let $e_j = (\xi_j^{\tau h}, \mu(x_i, t^j)), e_{j+1} = (\xi_{j+1}^{\tau h}, \mu(x_{i+1}, t^{j+1}))$ be two subsequent macroscopic events of GDS evolution that happen with the probability 1. Then the GDS velocity function between the macroscopic events e_j and e_{j+1} can be defined in an elementary space-time cell $c_{ij} \subset \bar{G}$ as

$$v(t, x) = \lim_{\tau \rightarrow 0} \frac{E^{\tau h}(x_i, \mu^j) \Delta \xi_j^{\tau h}}{\tau}. \quad (7.5)$$

The numerator under the limit in (7.5) is referred to as the velocity of the Markov Chain between two subsequent macroscopic events.

The definition of the velocity function as the most probable jump of the associated Markov Chain (the jump which minimizes the energy of the transition) gives a way to construct a stable approximation of the Hamiltonian of GDS evolution. We relate the macroscopic behaviour of the system to its microscopic characteristics defined in an elementary space-time cell c_{ij} . As a result, in any such cell⁹ the GDS velocity defined by (7.5) is always greater than or equal to 1. Hence, if the process is approximated in c_{ij} , the Courant-Friedrichs-Lowy (CFL) stability condition [12] ($\tau \leq h$) is satisfied automatically, regardless of the actual values of the velocity function in c_{ij} .

Remark 7.1 In the limiting case $h \rightarrow 0$, definition 7.1 loses its meaning and a macroscopic system degenerates into a point. Mathematically, however, this situation is

⁸To simplify the notations, numeric indexes near τ and h are omitted.

⁹it cannot degenerate into a point due to the existence of the macro-level

well-defined as $n \rightarrow \infty$ ($m \geq n$):

$$\lim_{n \rightarrow \infty} v(t, x) = v_\epsilon, \quad (7.6)$$

which returns us to the model (6.2).

Although formally, definition (7.5) coincides with the ordinary definition of the velocity function under the assumption of continuity (an number of microscopic events between e_j and e_{j+1}), the latter is subject to application only in the case when both of the following claims are justifiable:

- knowledge of the "exact" Hamiltonian;
- knowledge of the initial conditions with "infinite precision".

Neither of these two can be guaranteed even for a simplified dynamic motion [18,19]. Whereas in the classical definition of the velocity function we relate microscopic points in the macroscopic frame of reference, (7.5) establishes a correspondence between two macroscopic events on the probabilistic basis of microscopic events between them. Hence, *the GDS velocity is a measure of changes which take place on the microscopic level with respect to the macroscopic behaviour of the system*. If we assume that such changes are vanishing, $\lim_{\epsilon \rightarrow 0^+} v_\epsilon = v_1$, then we can expect (see (7.6)) that

$$\lim_{\epsilon \rightarrow 0^+, n \rightarrow \infty} v(t, x) = v_1. \quad (7.7)$$

We call the mathematical idealization of evolution described by the model (6.2) with the limiting velocity defined by (7.7) an Infinite Length Unperturbed Markov Chain (ILUMC). The reality of perturbations ($\epsilon > 0$) implies an approximation $v_\epsilon \approx v_1$ that leads to the computational idealization of an Infinite Length Perturbed Markov Chain (ILPMC). The approximate relationship

$$\lim_{n \rightarrow \infty} v(t, x) \approx v_1 \quad (7.8)$$

reflects our endeavors to describe the evolution of PGDS. In general, mathematical modelling of GDS evolution according to (7.8) implies an approximation of the macroscopic velocity function with respect to an inevitable approximation of the function of micro-velocity. Such an approximation can be seen as the choice of a countable base in a topological space that induces a transformation from a space-time of events of PGDS to a discrete space-time of macroscopic events of this system evolution. This assumes a passage from the grid of macroscopic events $\omega_{h\tau}$ defined by (7.4) to a new grid, nodes of which are computational models of these events defined by a topology base in the macroscopic frame of reference.

A consideration of the space-time as a causal discrete set was the subject of many publications (see, for example, [7,9] and references therein). Recently some new

theoretical results on dynamic system discretizations on lattices have been obtained [14]. Below we formalize these ideas with respect to our models using the Markov Chain approach. First, the state space of the initial macroscopic event e_0 has to be specified with respect to

- absolute time of the decision maker, and
- an approximation of the system-environment boundary at the initial moment of such absolute time.

In the case of a one-dimensional approximation we define this space in the macroscopic frame of reference as

$$\begin{aligned} \Xi(i; 0) &= \{x_i, \quad i = 0, 1, 2, 3, \dots, N; \\ N &= 2n, \quad n = \lceil (T - t_0)/h \rceil \}. \end{aligned} \quad (7.9)$$

We assign to each state of Ξ a particular probability weight p_i^0 , which can be defined on the basis of the micro-velocity approximation with the property of decreasing probabilities $1 \geq p_0^0 = p_1^0 > p_2^0 > \dots > p_N^0 \geq 0$ (the theoretical limit of "infinite precision" is not excluded). Thus, to define the state space of a macroscopic event, we include a theoretical possibility of GDS evolution in each cell of the grid of macroscopic events. If $h_i = h$, $i = 0, 1, \dots, n - 1$, then $\max_j \tau_j \leq h$ and the limiting case of equality leads to a consideration of a square grid $\omega_{\tau h}^0$ ($m = n$) which has the resolution to identify any macroscopic event relevant to system evolution in \tilde{G} when $n \rightarrow \infty$. This case implies $h \rightarrow 0$ (and as a consequence $\tau \rightarrow 0$) when the state space of the initial macroscopic event defined according to (7.10) degenerates into a ray that indicates the loss of connection between absolute DM-time and relative time of the dynamic system. We can circumvent this problem of uncontrolled propagation of initial uncertainty by a probabilistic description of macroscopic states which are subject to conservation of the Markov condition on the basis of an appropriately constructed Markov Chain associated with GDS evolution.

Definition 7.2 A set of macroscopic events defined by a mapping $\omega_{\tau h}^0 \rightarrow \Xi(i; j)$, where

$$\Xi(i; j) = \{(x_i, t_j), \quad i = \overline{k, 2n-k}, \\ j = k, \quad k = \overline{0, n}\}, \quad (7.10)$$

is called the cone of macroscopic events of system evolution.

Remark 7.2 The formula (7.10) in the definition 7.2 is given for a "one point target in absolute DM-time" and can be generalized for any target set including a set of isolated points in the DM-time scale ($t = T$). This may be of the great importance for some optimal control problems.

Our next step is an approximation of the macro-velocity function with respect to the micro-velocity using the definition 7.1. As a characteristic of the microscopic velocity function $\epsilon > 0$ we use a numerical index $\chi_n = o(\tau + h)$ defined in the macroscopic frame of reference by probability weights of the neighbourhood states of an associated Markov Chain.

Definition 7.3 A Markov Chain $\xi_n^{\tau h}$, $n < \infty$ is consistent with the Markov process $(h(\tau), \mu(t, x))$ defined by the mathematical model of GDS evolution (7.1), (7.2) if

$$E_n^{\tau h|(x_i, \mu^j)} \Delta \xi_j^{\tau h} = v_1(x_i, t^j, \mu^j) \tau + o(h + \tau) \quad (7.11)$$

and

$$\text{cov}_n^{\tau h|(x_i, \mu^j)} \Delta \xi_j^{\tau h} = o(h + \tau) \quad (7.12)$$

hold. We refer to the condition (7.11) as the condition of local consistency, whereas (7.12) is referred to as the global consistency condition.

Remark 7.3 The equalities (7.11), (7.12) imply the fact that the macroscopic properties of the system should not change dramatically in small (with respect to the whole evolution) DM-time-sets, although microscopic properties can vary significantly subject to the velocity function. Another way of putting it is that consistency conditions referring to the probabilistic microscopic level make explicit basic features of system evolution on the macroscopic level. The same role in physics is played by the second law of thermodynamics [43].

In general, even if in the reality of dynamic system evolution there exists

- a uniform movement of the microscopic frame of reference with respect to the macroscopic one with a velocity v_ϵ , and
- a linear dependency of the corresponding points (x, t) and (τ, h) ,

these facts can be established neither by mathematical modelling nor by a measuring experiment. However, the limiting case of our consideration (when $h \rightarrow 0$ and hence $\tau \rightarrow 0$) implies that

$$\text{cov}_n^{\tau h|(x_i, \mu^j)} \Delta \xi_j^{\tau h} \rightarrow 0 \text{ when } n \rightarrow \infty.$$

Of course, the infinite length Markov Chain is within the scope of the Markov theorem on the generalized law of big numbers.

Theorem 7.1 If a sequence of arbitrary random values $\Delta \xi_1, \Delta \xi_2, \dots, \Delta \xi_n, \dots$ satisfies the condition

$$\lim_{n \rightarrow \infty} \frac{1}{n^2} \sum_{i=1}^n \text{cov}[\Delta \xi_i] = 0,$$

then the limiting result

$$\lim_{n \rightarrow \infty} P\left\{ \left| \frac{1}{n} \sum_{i=1}^n (\Delta \xi_i - E[\Delta \xi_i]) \right| \geq \varepsilon \right\} = 0$$

holds for any arbitrary $\varepsilon > 0$.

Therefore, if we construct a Finite Length Perturbed Markov Chain (FLPMC) with the properties (7.11) and (7.12), we can guarantee convergence of such an approximation to ILPMC in the probabilistic sense of theorem 7.1 when the number of macroscopic states $n \rightarrow \infty$. The limit passages

$$\Delta \xi_n^\epsilon = \Delta \xi_n^{h\tau} \text{ (if } n \rightarrow \infty \text{ then)} \rightarrow$$

$$\Delta \xi_\infty^\epsilon \text{ (if } \epsilon \rightarrow 0 \text{ then)} \rightarrow \Delta \xi_\infty^{00}$$

illustrate schematically a connection between FLMC, ILPMC and ILUMC. An approximation error of FLMC with respect to ILUMC is defined by

$$E(\epsilon, n) = \delta(\epsilon) + \Delta_n,$$

which vanishes in the limit $\epsilon \rightarrow 0^+$ and $n \rightarrow \infty$. In this case¹⁰ the macro-velocity of the system coincides (see (7.7)) with the velocity of the associated ILUMC, and

$$(\xi_n^{\tau h}, \mu(x_i, t^j)) \rightarrow (h(\tau), \mu(x, t)).$$

Any other cases assume a probabilistic description of physical states (see [42]) that can be associated with an appropriately constructed Markov Chain. It makes it necessary to transform the continuous space-time of a macroscopic frame of reference into the discrete space-time of macroscopic events of system evolution, that is to construct the cone of macroscopic events. The base of this cone is subject to the implementation of the complementarity principle (6.7), which acknowledges the fact of the system existence at the initial moment of DM-time with the probability 1¹¹. We note that as an alternative approach there is the theoretical possibility to control possible changes of macro-velocity from the micro-level. In general, using an appropriate approximation (that is valid for the macroscopic level of system description), we can describe the event e_0 in the two complementary forms

- either position-and-DM formulation as $(x_0, 1)$,
- or time-and-macro-velocity formulation as $(t_0, 0)$.

Theoretically, we can combine both approaches by considering the problem in terms of macro-velocity and the DM-function that corresponds to the specification of the event e_0 as $(0, 1)$. Such a consideration is typical for

¹⁰when classical concept of continuous phase space trajectories can be formally applied

¹¹However, it does not give a way to specify the initial condition for the macro-model (7.1), (7.2)

mathematical models in optimal control theory, where the decision maker plays the role of the "error-nulling" optimizing device of a modeler type. This approach can be regarded as the *velocity-control formulation of evolutionary problems*. An alternative consideration of initial conditions as $(1, 0)$ seems to be intrinsic to the investigation of biological self-organizing dynamic systems. DM in such cases can be associated with the "observer", and this approach can be formally regarded as the *velocity-energy formulation of evolutionary problems*. To combine both possibilities in such a specification of the event e_0 , computational models of dynamic system evolution should be derived. The main difficulty that immediately arises stems from the necessity of an approximation of the limit of $n\epsilon(n)$ for any dynamic system which evolves in space-time ($n \rightarrow \infty$) under the possibility of vanishing perturbations ($\epsilon \rightarrow 0^+$). The method proposed in this paper is based on such a construction of computational event-models in the cone of macroscopic events that preserve the stability property of associated evolution. In general, such an approach permits the DM to switch from "observer" to "modeler" and vice versa whenever it is necessary.

To construct a stable approximation of the model (7.1), (7.2) the idea of the upwind discrete scheme with flux limiters [55] is used. Without loss of generality for the numerical procedure, we assume that $\bar{f}_0 = 0$, which reduces the equation (7.1) to (6.6(b)). First, let us introduce in the cone of macroscopic events (7.10) a floating grid:

$$\omega_{\tau h}^\Delta = \{(x_i, t_j^{\tau_{j-1}}), i = \overline{k, 2n-k}, j = k, k = \overline{0, n}\}, \quad (7.13)$$

where $t_j^{\tau_{j-1}} = t^{j-1} + \tau_{j-1}$ when $j > 1$, $t_j^{\tau_{j-1}} = t^0 + \tau$ when $j = 1$, and $t_j^{\tau_{j-1}} = t^0$ when $j = 0$. Provided all τ_{j-1} , $j = \overline{1, n}$, $\tau_0 = \tau$ are defined, the grid (7.13) generates a set of approximations to the macroscopic events defined by $\Xi(i; j)$. Since for a particular DM-time $t_j^{\tau_{j-1}}$ an associated event depends only on the macroscopic event that corresponds to the t^{j-1} -moment of DM-time, the value of τ_{j-1} is subject to stability conditions for the system. Such conditions depend on the velocity of the system, which is approximated using an evolution-associated Markov Chain. Now if we denote approximations to μ -function and v_1 on $\omega_{\tau h}^\Delta$ as d and v respectively, then the approximations

$$\frac{\partial \mu}{\partial x} \approx \frac{d_{i+\frac{1}{2}}^+ - d_{i-\frac{1}{2}}^+}{h} \text{ if } v_i < 0 \text{ and}$$

$$\frac{\partial \mu}{\partial x} \approx \frac{d_{i+\frac{1}{2}}^- - d_{i-\frac{1}{2}}^-}{h} \text{ if } v_i > 0$$

allow us to derive the discrete scheme

$$d_i^j = d_i^{j+1} + \tau \left\{ \frac{d_{i+\frac{1}{2}}^+ - d_{i-\frac{1}{2}}^+}{h} v^+ - \right.$$

$$\left. \frac{d_{i+\frac{1}{2}}^- - d_{i-\frac{1}{2}}^-}{h} v^- \right\}, \quad (7.14)$$

where $v^+ = \max[v_i, 0]$, $v^- = \max[-v_i, 0]$ and

$$d_{i+\frac{1}{2}}^- = d_i^j + \Delta d_i^j \gamma_1(r_i), \quad d_{i-\frac{1}{2}}^- = d_{i-1}^j + \nabla d_{i-1}^j \gamma_2(r_{i-1}),$$

$$d_{i+\frac{1}{2}}^+ = d_{i+1}^j + \Delta d_{i+1}^j \gamma_3(r_{i+1}), \quad d_{i-\frac{1}{2}}^+ = d_i^j + \nabla d_i^j \gamma_4(r_i).$$

Here $\gamma_i, i = \overline{1, 4}$ are flux limiters which are subject to definition with respect to the velocity function approximation. The other notations are the common

$$\nabla d_i^j = d_i^j - d_{i-1}^j, \quad \Delta d_i^j = d_{i+1}^j - d_i^j, \quad r_i = \frac{\nabla d_i}{\Delta d_i}.$$

Then the discrete scheme (7.14) can be rewritten in the form

$$\begin{aligned} d_i^{j+1} = d_i^j &\{1 - \frac{\tau}{h} [|v| + v^- \gamma_4 - v^+ \gamma_1]\} + \\ &\frac{\tau}{h} d_{i-1}^j \{[v^+(1 + \gamma_2) + v^- \gamma_4]\} + \\ &\frac{\tau}{h} d_{i+1}^j \{[v^-(1 - \gamma_3) - v^+ \gamma_1]\} + \\ &\frac{\tau}{h} d_{i-2}^j \{[-v^+ \gamma_2]\} + \frac{\tau}{h} d_{i+2}^j \{v^- \gamma_3\}. \end{aligned} \quad (7.15)$$

A verification of the sum of all coefficients near unknown function on the right hand side of (7.15) gives unity. Hence, provided nonnegativeness conditions are satisfied, we can associate these coefficients with transition probabilities of a Markov Chain. In fact, the conditions of nonnegativeness of probabilities are

$$1 - \frac{\tau}{h} (|v| + v^- \gamma_4 - v^+ \gamma_1) \geq 0, \quad \gamma_2 \leq 0, \quad \gamma_3 \geq 0, \quad (7.16)$$

$$v^+(1 + \gamma_2) + v^- \gamma_4 \geq 0, \quad v^-(\gamma_3 - 1) + v^+ \gamma_1 \leq 0. \quad (7.17)$$

The partial cases of $v^- = 0$ ($v^+ \neq 0$) and $v^+ = 0$ ($v^- \neq 0$) give the results

$$1 - \frac{\tau}{h} v^+(1 - \gamma_1) \geq 0, \quad \gamma_1 \leq 0, \quad -1 \leq \gamma_2 \leq 0,$$

and

$$1 - \frac{\tau}{h} v^-(1 + \gamma_4) \geq 0, \quad \gamma_4 \geq 0, \quad 0 \leq \gamma_3 \leq 1$$

respectively.

Lemma 7.1 Under the conditions (7.16), (7.17) the Markov Chain defined by time-transitions of the discrete scheme (7.15) is locally consistent with the process $(h(\tau), \mu(t, x))$ defined by the model (7.1), (7.2) if the equality for flux limiters

$$\tau[v^-(1 - \gamma_4 + \gamma_3) - v^+(1 + \gamma_1 - \gamma_2) - v] = o(\tau + h) \quad (7.18)$$

holds.

Proof. If a previous state of the Markov Chain was $\xi_j^{ht} = x$ subject to control d^j , then the assumption of lemma 7.1 we have the following table of transition probabilities for a new state ξ_{j+1}^{ht} :

New state	Probability of transition
$x - h$	$\frac{\tau}{h}[v^+(1 + \gamma_2) + v^- \gamma_4]$
$x + h$	$\frac{\tau}{h}[v^-(1 - \gamma_3) - v^+ \gamma_1]$
x	$1 - \frac{\tau}{h}[v + v^- \gamma_4 - v^+ \gamma_1]$
$x - 2h$	$\frac{\tau}{h}[-v^+ \gamma_2]$
$x + 2h$	$\frac{\tau}{h}[v^- \gamma_3]$

Therefore it can be verified that

$$\begin{aligned} E_j^{\tau h|(x, d^j)} \Delta \xi_j^{\tau h} &= -\tau[v^+(1 + \gamma_2) + v^- \gamma_4] + \\ &\quad \tau[v^-(1 - \gamma_3) - v^+ \gamma_1] + \\ 0[1 - \frac{\tau}{h}(|v| + v^- \gamma_4 - v^+ \gamma_1)] - 2\tau[-v^+ \gamma_2] + 2\tau[v^- \gamma_3] &= \\ &\quad \tau[v^-(1 - \gamma_4 + \gamma_3) - v^+(1 + \gamma_1 - \gamma_2)]. \end{aligned}$$

This equality together with definition (7.11) completes the proof. ■

Remark 7.4 The Markov Chain velocity $v_{MC} = [v^-(1 - \gamma_4 + \gamma_3) - v^+(1 + \gamma_1 - \gamma_2)]$ between two successive macroscopic events coincides with the velocity of the process when $n \rightarrow \infty$. For any finite value of n we have $v_{MC} \geq v$ which corresponds to the nonnegativeness of the covariance of the Markov Chain jump between these macroscopic events.

Lemma 7.2 Under the conditions (7.16), (7.17) the Markov Chain defined by time-transitions of the discrete scheme (7.15) is globally consistent with the process $(h(\tau), \mu(t, x))$ if the equality for flux limiters

$$\tau\{h[v^+(1 - \gamma_1 - 3\gamma_2) + v^-(1 + \gamma_4 + 3\gamma_3)] - \tau v_{MC}^2\} = o(\tau + h) \quad (7.19)$$

holds.

Proof. In a similar way to what was done in the proof of lemma 7.1, we construct the following table of transition probabilities:

New state	Value of $[\Delta \xi - E \Delta \xi]^2$
$x - h$	$(-h - \tau v_{MC})^2$
$x + h$	$(h - \tau v_{MC})^2$
x	$(0 - \tau v_{MC})^2$
$x - 2h$	$(-2h - \tau v_{MC})^2$
$x + 2h$	$(2h - \tau v_{MC})^2$

We notice that the probabilities of transitions correspond to those from the transition probability table in lemma 7.1. Therefore the computation of covariance

$$\begin{aligned} cov_j^{\tau h|(x, d^j)} \Delta \xi_j^{\tau h} &= E[\Delta \xi - E \Delta \xi]^2 = \\ \frac{\tau}{h} &[h^2 + 2h\tau v_{MC} + \tau^2 v_{MC}^2][v^+(1 + \gamma_2) + v^- \gamma_4] + \\ [h^2 - 2h\tau v_{MC} + \tau^2 v_{MC}^2][v^-(1 - \gamma_3) - v^+ \gamma_1] - \\ \tau^2 v_{MC}^2 &\{1 - \frac{\tau}{h}[|v| + v^- \gamma_4 - v^+ \gamma_1]\} + \\ [4h^2 + 4h\tau v_{MC} + \tau^2 v_{MC}^2] &[-v^+ \gamma_2] \\ [4h^2 - 4h\tau v_{MC} + \tau^2 v_{MC}^2][v^- \gamma_3] &+ \tau^2 v_{MC}^2 = \\ \tau h[v^+(1 - \gamma_1 - 3\gamma_2) + v^-(1 + \gamma_4 + 3\gamma_3)] - \tau v_{MC}^2 & \end{aligned}$$

gives the required equality (7.19), if we take into account (7.12). ■

Remark 7.5 For each cell $c_{ij} \subset \omega_{\tau h}$ a probabilistic analogue of the characteristics of equation (7.1) can be defined by the equality

$$cov_j^{\tau h|x, d^n} \Delta \xi_j^{\tau h} \pm \tau v_{MC} = const. \quad (7.20)$$

To estimate the value of *const* in (7.20) we can eliminate the term $o(\tau + h)$ in our approximation using (7.11) and (7.12):

$$cov_j^{\tau h|x, d^j} \Delta \xi_j^{\tau h} - \tau v_{MC} = -\tau v. \quad (7.21)$$

Using lemma 7.2 the equality (7.21) can be rewritten as

$$\begin{aligned} h[v^+(1 - \gamma_1 - 3\gamma_2) + v^-(1 + \gamma_4 + 3\gamma_3)] \\ - \tau v_{MC}^2 = \tau(v_{MC} - v). \end{aligned} \quad (7.22)$$

Therefore nonnegativeness of covariance is equivalent to the stability

$$\frac{\tau}{h} \leq \frac{v^+(1 - \gamma_1 - 3\gamma_2) + v^-(1 + \gamma_4 + 3\gamma_3)}{[v^-(1 + \gamma_3 - \gamma_4) - v^+(1 + \gamma_1 - \gamma_2)]^2}, \quad (7.23)$$

which follows directly from (7.22). Provided flux limiters are chosen in such a way that the equality

$$\begin{aligned} v^+(1 - \gamma_1 - 3\gamma_2) + v^-(1 + \gamma_4 + 3\gamma_3) = \\ [v^-(1 + \gamma_3 - \gamma_4) - v^+(1 + \gamma_1 - \gamma_2)]^2 \end{aligned} \quad (7.24)$$

holds, the stability condition (7.23) is satisfied.

Example 7.1 Examples of the choices of flux limiters are given below for two partial cases.

- If $v^- = 0$ and $\gamma_2 = 0$ ($i = j$) then the value of the flux limiter γ_1 can be found from (4.16) in the form

$$\gamma_1 = -1 - \frac{\sqrt{8v^+ + 1} + 1}{2v^+}.$$

- If $v^+ = 0$ and $\gamma_3 = 0$ ($i = N - j$) then the value of the flux limiter γ_4 is defined as

$$\gamma_4 = 1 + \frac{\sqrt{8v^- + 1} + 1}{2v^-}.$$

The identification of flux limiters completes the construction of the discrete scheme which defines the Markov Chain with the corresponding interpolation interval τ (subject to stability conditions) and transition probabilities. We state the result in the form of the theorem on the Markov-Chain-approximation stability in discrete space-time of events.

Theorem 7.2 If transition probabilities of a Markov Chain (ξ_n^{rh} , $n < \infty$) are defined by the formula

$$p^{rh}[x_k^j, x_i^{j+1} | d(x_k^j, t^j)] = \begin{cases} 1 - \frac{\tau}{h} [v| + v^- \gamma_4 - v^+ \gamma_1], & k = i, \\ \frac{\tau}{h} [v^+(1 + \gamma_2) + v^- \gamma_4], & k = i - 1, \\ \frac{\tau}{h} [v^- (1 - \gamma_3) - v^+ \gamma_1], & k = i + 1, \\ -\frac{\tau}{h} (v^+ \gamma_2), & k = i - 2, \\ \frac{\tau}{h} (v^- \gamma_3), & k = i + 2, \\ 0, & \text{otherwise,} \end{cases}$$

$\forall j = \overline{0, n-1}$ and $i = \overline{j, N-j}$ ($\gamma_2 = 0$ for $i = j$ and $\gamma_3 = 0$ for $i = N - j$), whereas the interpolation interval τ satisfies the conditions (7.16), (7.17), (7.23), then the Markov Chain approximation of the process $(h(\tau), \mu(x, t))$ is stable, and discrete values of the DM-function can be found from the formula

$$d(x_i^{j+1}, t^{j+1}) = \sum_k p^{rh}[x_k^j, x_i^{j+1} | d(x_k^j, t^j)] d(x_k^j, t^j). \quad (7.25)$$

Remark 7.6 (on convergence). When $n \rightarrow \infty$ the velocity of the Markov Chain converges to the velocity of the process in the sense of theorem 7.1. If we consider, for example, a formulation of the problem in terms of velocity-control, then due to the complementarity principle the discrete function (7.25) converges to the decision maker function of the system.

Remark 7.7 (on numerical procedures). A numerical method proposed in this section is an explicit (evolution forward) stabilization procedure where the DM-function is a stabilizing factor subject to the velocity of the system.

Remark 7.8 (on backward evolution operators and continuity of phase space trajectories). A probabilistic description of event e_{n_0} precludes the situation where terminating data for backward evolution procedures can be specified in a "deterministic" way.

Moreover, states $x(t_0)$ and $x(T)$ of the system in DM-absolute-time scale can be characterised by different probability weights, which makes the continuity assumption for the connecting trajectory inapplicable in general.

8 Computational Aspects of Discrete Markov Decision Processes.

In a vicinity of any event e_0 which we might conditionally associate with the present of GDS evolution, there are infinitely many events relevant to the GDS evolution which might be called *past* and *future* events of evolution. As a result, an event itself can be formalized mathematically, neither with a deterministic certainty, nor with a precise probability. This implies difficulty in justifying the separability of topological spaces when the evolution of UGDS and PGDS is investigated.

Let us denote a probabilistic error of the inevitable approximation of such an event in the initial conditions of a mathematical model as $\varrho_0^\epsilon \in (0, 1]$, $\epsilon > 0$. Then the principal mathematical assumption which allows us to develop analytical theory of dynamic system in continuous (space)time is a *possibilistic assumption of vanishing error*

$$\lim_{\epsilon \rightarrow 0^+} \varrho_0^\epsilon = 0. \quad (8.1)$$

Moreover, a concept of absolute or "external" to the system DM-time [43,38] leads to the theoretical possibility of predicting a future event e_{n_0} which is associated with the DM-time $t = T$ (possibly $T = \infty$) with the probability 1. This means that

$$\lim_{\epsilon \rightarrow 0^+} \varrho_{n_0}^\epsilon = 0, \quad (8.2)$$

where $\varrho_{n_0}^\epsilon$ is a probabilistic error in the definition of this event. This approach (which is deterministic in its essence) usually visualizes evolution as a continuous trajectory $x(t)$ between present e_0 (time $t = t_0$) and future e_{n_0} (time $t = T$) events along which positions of the system can be determined at least in principle with the probability 1. Assuming that (8.2) holds, let us try to go backward in continuous DM-time. If evolution of the system in continuous space-time has taken place at all, we can select between events e_{n_0} and e_0 at least $(n_0 - 1)$ events relevant to system evolution, which we will refer to as macroscopic. Further, we can extract between macroscopic events e_1 and e_0 at least $(n_1 - 1)$ events relevant to system evolution which we will call microscopic, and will denote as $e_1^{01}, e_2^{01}, \dots, e_{n_1-1}^{01}$. In the same way, we can find $(n_2 - 1)$ sub-microscopic events $e_1^{011}, e_2^{011}, \dots, e_{n_2-1}^{011}$ etc. As a result, we obtain a functional of the event-transition-error in the form

$$F(x, t) = \sum_{i=1}^{n_0-1} \varrho_{i+1}(e_i) + \sum_{i=1}^{n_1-1} \varrho_{i+1,0}(e_{i,0}) +$$

$$\begin{aligned} & \sum_{i=1}^{n_2-1} \varrho_{i+1,00}(\varrho_{i,00}) + \dots \\ & + \sum_{i=1}^{n_k-1} \underbrace{\varrho_{i+1,00}}_{\star}(\underbrace{\varrho_{i,00}}_{\star} \dots) + \dots \end{aligned} \quad (8.3)$$

where, for example, a probabilistic error in a transition between events e_i^{011} and e_{i+1}^{011} ($i = 1, \dots, n_2 - 1$) is defined by $\varrho_{i+1,00}(\varrho_{i,00})$. To guarantee convergence of the series in the right hand side of (8.3) we should require

$$\lim_{k \rightarrow \infty} \underbrace{\varrho_{2,00}}_{\star} \dots \underbrace{(\varrho_{1,00}}_{\star} \dots) = 0,$$

where, assuming that (8.1) holds, we also have

$$\lim_{k \rightarrow \infty} \underbrace{\varrho_{1,00}}_{\star} \dots = 0.$$

Applying the same arguments in the forward DM-time we can draw the conclusion that for any “middle” macroscopic event $e_m \in (e_0, e_{n_0})$ (DM-time $t_m \in (t_0, T)$) both events e_0 (DM-time $t = t_0$) and e_{n_0} (DM-time $t = T$) are *infinitely far from it in the continuous (space)time of events*. However, in the macroscopic frame of reference, the distance between the events e_m and e_0 as well as a distance between e_m and e_{n_0} are well-defined in terms of absolute DM-time by the intervals $\Delta_{0,m} = t_m - t_0$ and $\Delta_{m,n_0} = T - t_m$ respectively. In other words, provided that both assumptions (8.1) and (8.2) can be justified, any event e_m of GDS evolution has two time-characteristics: (absolute macroscopic) DM-time $t_m \in (t_0, T)$ and (relative microscopic) system-time $\tau_m \in (-\infty, +\infty)$. The mathematical formalism, that allows us to circumvent the arising difficulty of time scaling, is based on the Cauchy-type models, and requires an exact specification of initial (or terminal) conditions for the position-vector or the density function in a separable topological space. Eventually, mathematically rigorous justification of such models requires simultaneous application of the concept of a time-infinity (either in the form of ergodic-type hypotheses or infinite-step algorithm) and the possibility of vanishing perturbations when time goes by. Another way of putting it is that infinite time is a necessary condition for the justification of unperturbed mathematical models. However, *sufficiency* of this condition is subject to possibility theory [15,47] rather than the theory of probabilities. From the physical point of view the analysis of the described problem requires the concept of relative time. The mathematical idealization which reconciles the concepts of absolute and relative time of dynamic system evolution is ILUMC in the continuous space-time of events, for which the claim of $(t, \tau) \in (-\infty, +\infty)$ is natural. The very next step in the modelling of dynamic system evolution is ILPMC. Such models imply an *approximation* of an event e_0 that formally gives two rays in relative-time directions

$((-\infty, \tau_0)$ and $(\tau_0, +\infty)$). Our knowledge of the relative time τ_0 is based on its intermediate influence on the quality of approximations of objects of mathematical modelling with respect to the moment t_0 of absolute time. A selection of one of the two rays in relative-time directions corresponds to the choice of a Markov semigroup [42] associated either with a covariance-non-negative (for future) or a covariance-non-positive (for past) Markov Chain. Whatever model is chosen, the Markovian property for the evolution should be preserved by an appropriate algorithm. It requires consideration of perturbed mathematical models with the specified level of error.

An approximation of event e_0 implies a truncation of the series $F(x, t)$ in (8.3). Let us denote a probabilistic error induced by such a truncation as

$$\varrho_0^\epsilon = \underbrace{\varrho_{1,00}}_{\star} \dots > 0.$$

Let us also assume that the limit of vanishing error,

$$\lim_{t \rightarrow \infty} \underbrace{\varrho_{1,00}}_{\star} \dots = \lim_{t \rightarrow 0} \varrho_0^\epsilon = 0,$$

holds. Then, in general, the quality of prediction of GDS evolution by means of mathematical modelling is defined by the quality of a solution of the optimization problem

$$\sum_{i=0}^{\infty} \varrho_{i+1}^\epsilon(\varrho_i^\epsilon) \rightarrow \min. \quad (8.4)$$

Since the difference between an unperturbed trajectory x_t of ILUMC and a perturbed trajectory x_t^ϵ of ILPMC at a certain moment $t = t_m$ of DM-time can be arbitrary big, the necessary condition for convergence of series (8.4),

$$\lim_{n \rightarrow \infty} \varrho_{n+1}^\epsilon(\varrho_n^\epsilon) = 0,$$

cannot be guaranteed in general, no matter how small $\epsilon > 0$ is assumed. This is not a surprising fact since in general the optimizing function is a function of an infinite degree of recursion of the density function. The intrinsic idea in mathematical modelling and computational experiments is to reduce the degree of recursion to a finite number. In doing so we arrive at the problem

$$\sum_{i=0}^{n_0} \tilde{\varrho}_{i+1}^\epsilon(\tilde{\varrho}_i^\epsilon) \rightarrow \min,$$

which implies the construction of FLPMC. Though the difference between two macroscopic states x_k^ϵ and x_{k+1}^ϵ in DM-time scale might still be arbitrary big in general (between two macroscopic events e_k and e_{k+1} there might be an infinite number of microscopic events relevant to system evolution), we are now able to estimate a probability of corresponding transition using the values of $\tilde{\varrho}_{i+1}^\epsilon(\tilde{\varrho}_i^\epsilon)$. By means of FLPMC we preserve the

stability of the macroscopic system (the object of mathematical modelling) with respect to its microscopic dynamics. Although stability of the microscopic dynamics with respect to a macroscopic system will follow in the limit of our construction, any finite time computational procedure is not necessarily a reflection (even qualitatively) of the latter. To put it differently, using tools of mathematical modelling, results generated by ILUMC or ILPMC (i.e. a complete description of GDS evolution) cannot be guaranteed with the probability 1. If it is granted that mathematical modelling can give a way to describe the real processes, systems, and phenomena, then a conceptually necessary passage from continuous trajectories ($x(t)$ or $x'(t)$) in absolute ("external" to the system) DM-time to a probabilistic description of physical states should be undertaken. A convenient framework for a probabilistic description of system evolution from one macroscopic event to another provides the concept of Discrete Markov Decision Processes (DMDP) [26]. Since DMDP is considered in the macroscopic frame of reference, both

- a number of observed macroscopic events (which is finite $n_0 = n_0(T, t_0, h(\tau))$), and
- a topology of the state space,¹²

depends on an approximation of initial e_0 and terminating e_{n_0} events. In the macroscopic frame of reference the state space gives rise to the cone of macroscopic events

$$\Xi(i; j) = \{x_i^j, \quad i = \overline{n_j, N_j}\},$$

where for $j = 0$ we have $n_j = 0$ and $N_j = N$. In the case of 1-D approximation (one-point-target), the cone of macroscopic events was defined by (7.10). At a certain moment t^j of absolute DM-time, the system can be in one of the states x_i^j to which we assign different probabilistic weights p_i^j . With the same probability weights we associate the corresponding action set defined by

$$\mathcal{A}(i; j) = \{\mu(t^j, x_i^j), \quad i = \overline{n_j, N_j}\}.$$

If we now define an allowable decision set for each macroscopic event e_j , $j = \overline{0, n_0}$ as

$$\mathcal{D}(i; j) = \Xi(i; j) \otimes \mathcal{A}(i; j),$$

then the construction of a probabilistic model for each macroscopic event of system evolution

$$e_j \equiv \{\mathcal{D}(i; j); p_i^j\}$$

has been completed. If $\mu(t^j, x_i^j) \forall i = \overline{n_j, N_j}$ is known, a description of the macroscopic event at DM-time t^j becomes totally deterministic. A reward set is defined

¹²which can change in general with respect to absolute DM-time due to fluctuating system-environment boundaries

by the probability distribution of the next macroscopic event:

$$R(e_j \rightarrow e_{j+1}) = (\{(x_i^{j+1}, \mu(t^{j+1}, x_i^{j+1}); p_i^{j+1}), \\ i = \overline{n_{j+1}, N_{j+1}}\}).$$

The following stabilization procedure (which is described with respect to the approximations used in section 7) can be applied as an implementation of theorem 7.1.

Algorithm 8.1 • Initialization of initial event.

Find initial values of the DM-function at $t = t^0$. That is, define event e_0 by triples $(x_i^0, \mu(x_i^0, t^0); p_i^0)$. Then set complementary description of the initial event as $(t^0, v_i^0; p_i^0)$.

• Prediction step for an event-model.

Given values of $\mu(x_i^0, t^0)$, define an approximation to the velocity function of the process $v(t^0 + h, x_i + h, \mu(x_i^0, t^0)) = v_i^1$, $i = \overline{2, 2n}$ at the next DM-time-layer $t^1 = t^0 + h$. Set $(t^1, v_i^1; p_i^1)$ as an approximate description of the next macroscopic event.

• Correction step for the event-model.

Using the approximate description of the event in terms of time-velocity, find flux limiters and the time-step of stability τ , for which define an event-model as $(t^{1,\tau}, \mu(x_i^{1,\tau}, t^{1,\tau}); p_i^1)$, where $i = \overline{2, 2n}$ and $t^{1,\tau} = t^0 + \tau$.

• Event definition.

The definition of DM-function by $\mu(x_i^1, t^1) = \mu(x_i^{1,\tau}, t^{1,\tau})$, $i = \overline{2, 2n}$ gives the new macroscopic event in the form of the set of triples $(x_i^1, \mu(x_i^1, t^1); p_i^1)$ $i = \overline{2, 2n}$.

• Complementary description of the event.

Define complementary description of the event as $(t^1, v_i^1; p_i^1)$ and repeat the procedure for the next DM-time-layer, etc. ■

Since in the DM-time scale (where the stabilization procedure has to be employed) a real event always follows after its event-model counterpart, this implies that an error at each step of the procedure is defined by a time-discrepancy between the event and its event-model (for example, the first step of 1-D approximation gives $\Delta t = t^1 - t^{1,\tau} = h - \tau$). To minimize this error we should find a Markov strategy which at each DM-time level chooses the highest probability of a transition. In general we have the whole family of DMDP defined as

$$\begin{aligned} M(e_0, e_1, \dots, e_{n_0}) &= \{\mathcal{D}(i; j), R(e_j \rightarrow e_{j+1}), \\ &p_j^{j+1}(k; i), i = \overline{n_j, N_j}, j = \overline{0, n_0}\}, \quad (8.5) \end{aligned}$$

where $p_j^{j+1}(k; i) = p^{h\tau}[x_k^j, x_i^{j+1} | \mu(x_k^j, t^j)]$. Each of such a DMDP constructs a probabilistic trajectory of system evolution

$$T(k_0, \dots, k_{n_0}) = \{p_0^0(k_0), p_0^1(k_0; k_1), p_1^2(k_1; k_2), \dots,$$

$$p_{n_0-1}^{n_0}(k_{n_0-1}; k_{n_0})\},$$

($k_l \in [n_l, N_l]$, $l = \overline{0, n_0}$), in the cone of macroscopic events. In general the equality $p_{n_0-1}^{n_0}(k_{n_0-1}; k_{n_0}) = 1$ cannot be guaranteed, and closeness of this probability to 1 depends on values of $p_0(k_0)$ and the structure of the cone of macroscopic events (i.e. on the approximation of e_0 and e_{n_0}). To single out amongst all probabilistic trajectories defined by DMDP (5.5) an optimal one we define the probabilities of successful prediction as

$$\bar{p}_0^0(k_0^*) = \max_i p_0^0(i; i), \quad i = \overline{n_0, N_0},$$

and then

$$p_0^1(k_0^*; i) = \max_k p_0^1(k; i),$$

$$\bar{p}_0^1(k_0^*; k_1^*) = \max_i p_0^1(k_0^*; i), \quad i = \overline{n_1, N_1}.$$

In general we have

$$p^{j+1}(k_j^*; i) = \max_k p_j^{j+1}(k; i), \quad \bar{p}_j^{j+1}(k_j^*; k_{j+1}^*) = \max_i p_j^{j+1}(k_j^*; i),$$

where $i = \overline{n_j, N_j}$, $j = \overline{1, n_0 - 1}$. Then the policy

$$\pi = \{(x_{k_l^*}, \mu(x_{k_l^*}, t^l)), l = \overline{0, n_0}\},$$

induced by the optimal probabilistic trajectory

$$T^*(k_0^*, \dots, k_{n_0}^*) = \{\bar{p}_0^0(k_0^*), \bar{p}_0^1(k_0^*; k_1^*), \bar{p}_1^2(k_1^*; k_2^*), \dots, \\ p_{n_0-1}^{n_0}(k_{n_0-1}^*; k_{n_0}^*)\},$$

we call the **optimal Markov policy**, which gives (in the DM-time scale) the Markov Chain approximation to GDS evolution $(h(\tau), \mu(t, x))$. The DMDP, which is associated with this policy, corresponds to the construction of such a Markov Chain which evolves to the most probable state of the system preserving the property of strong causality of macroscopic events.

9 Conclusion.

In this paper mathematical modelling of dynamic system evolution has been studied as a problem in information theory. Computational models for evolution based on the ideas of evolution-associated Markov Chain approximations have been developed. Since the velocity function of the system is coupled to perturbations of its environment, stability conditions for the system have been derived in an explicit form.

Mathematical models for the evolution of dynamic systems are closely connected with discrete optimization problems through the definition of information and the associated notion of entropy for thermodynamic systems. Information uncertainty in knowledge bases influences the construction of mathematical models, and should be taken into account. This implies a certain heuristic nature in such a construction. Such heuristic approaches are an important part of studying dynamic system evolution, and will remain as such in the foreseeable future, supplementing achievements obtained with the increasing computational power of modern computers and improved methods of data collection and analysis. Moreover, hybrid procedures combining the features of constructive, sequential, and evolutionary algorithms of discrete optimization give a general framework that could challenge well-established techniques in optimization theory.

Many important breakthroughs in optimization theory are intrinsically connected with the application of algorithms of sequential analysis that are based on the Markovian-type schemes. Such schemes are typical in computational models where minimax concepts of optimality are used. A mathematical formalization of the problem is quite natural, and is computationally consistent. The problem is viewed as attempts by the decision maker to obtain the best guaranteed result with respect to available information about the problem. The same formalization is a starting point for constructing mathematical models where other (such as probabilistic) concepts of optimality are used. In applications of such decision-maker schemes there is a natural contradiction between a desire for informational completeness in the model that is being constructed and a desire to choose functional classes for which effective computational algorithm exists. In a search for a compromise between these two extremes induced by the "energetic" (combinatorial) and informational complexity of the underlying algorithm [54,56] it is reasonable to include the decision maker as an intrinsic part of the constructed model using some learning rules. As a result, mathematical models become coupled to their computational associate. This allows us to look for the optimal algorithms as those that at each step of their performance in the best way use the information, which is accumulated by this step. The number of steps and quality of performance can be mathematically defined by the degree of recursion of an approximation to the system Hamiltonian (with respect to the density function) and the parameter of perturbations. In studying dynamic system evolution it is expected that a compromise between the two mentioned types of complexity can be achieved by the requirement of system stability. This cannot be guaranteed in general unless the underlying model is defined by *hyperbolic* rather than *parabolic* dynamic rules. Examples of this type have been derived, and the limiting cases

of vanishing perturbations and infinite recursion rule have been discussed. The results on the derivation of hyperbolic equations of the Hamilton-Jacobi-Bellman type for non-smooth and stochastic optimal control will be published separately [39]. Their connection with the principles of extended irreversible thermodynamics [44,30] as well as computational algorithms shall be also discussed elsewhere.

Acknowledgement

I wish to thank Dr S. Lucas for his careful reading and suggestions of improvement.

References

- [1] M. Abbad and J. Filar, "Perturbation and stability theory for Markov control problems", *IEEE Trans. Automat. Contr.*, vol.37, no.9, September, pp. 1415-1420, 1992.
- [2] J. P. Aubin and H. Frankowska, *Set-Valued Analysis*, Boston: Birkhauser, 1990.
- [3] J. A. Bather and H. Chernoff, "Sequential decision in the control of a spaceship", *Fifth Berkely Symposium on Mathematical Statistics and Probability*, 3, pp.181-207, 1967.
- [4] J. A. Bather and H. Chernoff, "Sequential decision in the control of a spaceship (finite fuel)", *J. Appl. Prob.*, 49, pp.484-604, 1967.
- [5] R. Bellman, "A Markovian decision problem", *J. Math. Mech.*, 6, pp.679-684, 1957.
- [6] R. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, NJ, 1957.
- [7] L. Bombelli, J. Lee, D. Meyer, R. Sorkin, "Space-Time as a Causal Set", *Phys. Rev. Lett.*, 59(5), pp.521-524, 1987.
- [8] B. Bouchon-Meunier, R.R. Yager and L.A. Zadeh (Eds), *Uncertainty in Knowledge Bases*. Springer-Verlag, 1991.
- [9] G. Brightwell and R. Gregory, "Structure of Random Discrete Spacetime", *Phys. Rev. Lett.*, 66(3), pp.260-263, 1991.
- [10] D. Costa, A. Hertz and O. M. Dubuis, "Embedding a Sequential Procedure Within an Evolutionary Algorithm for Coloring Problems in Graphs", *Journal of Heuristics*, 1, pp.105-128, 1995.
- [11] M. Crandall and P. Lions, "Viscosity solutions of Hamilton-Jacobi equations", *Trans. of the American Math. Society*, vol. 277, no.1, pp.1-42, 1983.
- [12] R. Courant, K. Friedrichs and H. Lewy, "On the Partial Differential Equations of Mathematical Physics", New York University, Courant Institute of Mathematical Sciences, Report NYO-7689, 1956.
- [13] M. Davis, "Piecewise-deterministic Markov processes: a general class of non-diffusion stochastic mod- els", *J.Royal Statistical Soc., Ser B*, 46, pp.353-388, 1984.
- [14] P. Diamond, P. Kloeden and A. Pokrovskii, "Degree theory on finite lattices: discretizations of dynamical systems", *Numer. Funct. Anal. and Optimiz.*, vol.16, no. 1&2, pp.43-52, 1995.
- [15] D. Dubois and H. Prade, *Possibility Theory*, Plenum Press, New York, 1988.
- [16] J. Filar and D. Krass, "Hamiltonian Cycles and Markov Chains", *Mathematics of Operations Research*, vol.19, pp.223-237, 1994.
- [17] J. Filar and O. Vrieze, *Competitive Markov Decision Processes*, Springer-Verlag, 1996.
- [18] W. Fleming, Optimal continuous-parameter stochastic control, *SIAM Review*, vol.11, no.4, pp.470-509, 1969.
- [19] W. Fleming and H. Soner, *Controlled Markov Processes and Viscosity Solutions*. Springer-Verlag, 1993.
- [20] H. Frankowska, "Lower semicontinuous solutions of Hamilton-Jacobi-Bellman equations," *SIAM J. Control and Optimiz.*, vol.31, no.1, pp.257-272, 1993.
- [21] S. Geman, and D. Geman Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, pp.721-741, 1984.
- [22] J. W. Gibbs, *Elementary Principles in Statistical Mechanics*, Dover, New York, 1960.
- [23] S. Godunov et al, "Guaranteed Accuracy in Numerical Linear Algebra", Kluwer Academic Publishers, 1992.
- [24] S. Goldstein, "Sufficient Conditions to Single out the Gibbs Measure from other Time-Invariant Measures", in *Long-Time Prediction in Dynamics* Ed. by C.W.Horton, Jr., L.E.Reichl, and V.G.Szebehely, John Wiley & Sons, pp.71-78, 1983.
- [25] O. Hernandez-Lerma, "Adaptive Markov control processes," in *Applied Mathematical Sciences*, vol.79, New York: Springer-Verlag, 1989.
- [26] D. P. Heyman and M.J. Sobel (Ed.) *Stochastic Models*, Elsevier Science, 1990.
- [27] R. W. Hockney and J. W. Eastwood, *Computer Simulation Using Particles*, McGraw-Hill Inc., 1981.
- [28] H. Ishii, "Viscosity solutions of nonlinear second-order partial differential equations in Hilbert spaces", Preprint no.22, Chuo University, Department of Mathematics, 1991.
- [29] A. Jadczyk, "On quantum jumps, events, and spontaneous localization models", *Foundations of Physics*, vol.25, no.5, pp.743-763, 1995.
- [30] D. Jou, J. Casas-Vazquez and G. Lebon, *Extended Irreversible Thermodynamics*, Springer-Verlag, 1993.
- [31] S. Kirkpatrick, C.D. Gelatt, Jr., and M.P. Vecchi, "Optimization by simulated annealing", *Science*, 220, pp.671-681, 1983.

- [32] A. Kolmogorov and S. Fomin, *Fundamentals of Theory of Functions and Functional Analysis*, Moscow: Nauka, 1989 (or by Dover Publications, New York, 1975).
- [33] V. Korotkikh, "Multicriteria analysis in problem solving and structural complexity", in *Advances in Multicriteria Analysis*, P.M. Pardalos et al. Eds. Kluwer Academic Publishers, pp.81-90, 1989.
- [34] P. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification and Adaptive Control*, Englewood Cliffs, NJ: Prentice-Hall, 1986.
- [35] H. Kushner and P. Dupuis, *newblock Numerical Methods for Stochastic Control Problems in Continuous Time*. New York: Springer-Verlag, 1992.
- [36] L. D. Landau and E. M. Lifshitz, *Statistical Physics*. Pergamon Press, 1959.
- [37] M. C. Mackey, *Time's Arrow: The Origins of Thermodynamic Behaviour*. Springer-Verlag, 1992.
- [38] V. N. Melnik, "Nonlinear Dynamical Systems: Coupling Information and Energy in Mathematical Models", *40th Conference of the Australian Mathematical Society, Adelaide, July, 1996*.
- [39] V. N. Melnik, "Nonconservation Law Equation In Mathematical Modelling: Aspects of Approximation", *Proceedings of the International Conference AEMC'96, Sydney*, pp.423-430, 1996.
- [40] N. Merhav and M. Feder, "Universal schemes for sequential decision from individual data sequences", *IEEE Trans. Inform. Theory*, vol.39, no.4, July, pp.1280-1292, 1993.
- [41] M. Metropolis et al. "Equation of state calculations by fast computing machines", *Journal of Chemical Physics*, 21, pp.1087-1092, 1953.
- [42] B. Misra, I. Prigogine and M. Courbage "From Deterministic Dynamics to Probabilistic Descriptions", *Physica 98A*, pp.1-26, 1979.
- [43] B. Misra, and I. Prigogin, "Time, Probability, and Dynamics", in *Long-Time Prediction in Dynamics*, Ed. by C.W.Horton, Jr., L.E.Reichl, and V.G.Szebehely, John Wiley & Sons, pp.21-43, 1983.
- [44] I. Muller and T. Ruggeri, *Extended Thermodynamics*, Springer-Verlag, 1993.
- [45] M. Nisio, "Optimal control for stochastic partial differential equations and viscosity solutions of Bellman equations", *Nagoya Math. J.*, vol.123, pp.13-37, 1991.
- [46] D. Potter, *Computational Physics*, John Wiley & Sons, 1973.
- [47] A. Ramer, "Information theory based on fuzzy (possibilistic) rules", *Proceedings of the 3rd Intern. Conf. on IPMU, Paris, 1990*, pp.317-326, 1991.
- [48] C.R. Reeves (Ed.) *Modern Heuristic Techniques for Combinatorial Optimization*. Oxford: Blackwell Scientific, 1993.
- [49] W. Rudin, *Functional Analysis*, McGraw-Hill, 1973.
- [50] P. J. Schweitzer, "Perturbation theory and finite Markov chains", *J. Appl. Probability*, vol.5, pp.401-413, 1968.
- [51] C. Shannon, "A mathematical theory of communication", *Bell. Sys. Tech. J.*, vol.27, pt.1, pp.379-423; pt.II, pp.623-656, 1948.
- [52] H. M. Soner and S. E. Shreve, "Regularity of the value function for a two-dimensional singular stochastic control problem", *SIAM J. Control and Optimiz.*, vol.27, no.4, pp.876-907, 1989.
- [53] M. Struwe, *Variational Methods*, Springer-Verlag, 1990.
- [54] A. Sucharev, *Minimax Models in the Theory of Numerical Methods*, Kluwer Academic Publishers, 1992.
- [55] P. Sweby, "High resolution schemes using flux limiters for hyperbolic conservation laws", *SIAM J. Numer. Anal.* 21(5), pp.995-1011, 1984.
- [56] J. F. Traub and H. Wozniakowski, *A General Theory of Optimal Algorithms*, New York: Academic Press, 1980.
- [57] A. Vasilieva, V. Butuzov and L. Kalachev, *The Boundary Function Method for Singular Perturbation Problems*, SIAM Studies in Applied Mathematics, 14, Philadelphia, PA, 1995.
- [58] N. Wiener, *Cybernetics*. New York: John Wiley & Sons, 1948.

UNIVERSITY OF SOUTH AUSTRALIA



TECHNICAL REPORT SERIES

**Nonnegativity of Energy and
Stability Requirements in a Dynamic
Problem of Coupled Field Theory**

by

(R.) V Nick Melnik

Report No. 1996/12

CENTRE FOR INDUSTRIAL
AND APPLIED MATHEMATICS
SCHOOL OF MATHEMATICS

Faculty of Information Technology

The Levels, South Australia 5095, Telephone (08) 8302 3343 Facsimile (08) 8302 5785

TECHNICAL REPORT SERIES

**Nonnegativity of Energy and
Stability Requirements in a Dynamic
Problem of Coupled Field Theory**

by

(R.) V Nick Melnik

Report No. 1996/12

NONNEGATIVITY OF ENERGY AND STABILITY REQUIREMENTS IN A DYNAMIC PROBLEM OF COUPLED FIELD THEORY

V. Nick Melnik ¹

Abstract

In this paper a coupled problem of dynamic electroelasticity has been investigated using the variational approach and the concept of generalized solutions. We have derived a numerical procedure directly from the definition of the generalized solution of the problem. We prove the convergence of the numerical scheme (with the second order in space-time) to the solution of the original problem from a class of generalized solutions. The stability condition has been obtained from an energy estimate. It was shown that such a condition is the Courant-Friederichs-Lowy-type stability condition being depend on the velocity of *mixed electro-elastic waves*. Coupling effects are discussed with a numerical example.

Key words: generalized solutions, coupling, mixed electroelastic waves, convergence and stability.

1. INTRODUCTION.

Modern applications of the coupled theory of dynamic electroelasticity include situations where solutions of the underlying problems do not have to be "smooth" in a classical sense. In many practicaly important cases we are dealing with steep gradients of solutions as well as with different kinds of wave discontinuities. Of course, the notion of classical solutions is not appropriate in such situations, and the original problem should be reformulated using variational principles. Such a reformulation is especially important for coupled non-stationary problems of electroelasticity. However, an approach coupling mechanical and electric fields which is essential to obtain a plausible picture of described phenomena, involves some difficulties. They are induced primarily by the necessity to deal with the coupling phenomenon from the very beginning of the justification of the underlying mathematical model. The use of numerical procedures does not remove these difficulties. Moreover, we should be able to justify corresponding numerical procedures not only under classical smoothness

¹The main results of this paper are submitted to *Mathematics & Mechanics of Solids*, Sage Science Press

assumptions,² but also show their robustness on the classes of generalized solutions as well.

Studies of electromechanical interactions are important in classical areas of mechanics of solids, as well as in many new areas of applications including engineering and biophysics. An increasing range of applications of piezoelectrics in semiconductors and intelligent structures has stimulated a greater interest in coupling effects between mechanical and electric fields [2], [3], [6]. For many of the arising problems, solutions with steep gradients or even discontinuities are typical features of underlying physical processes.

Historically, the most well developed area of research in electroelasticity abuts to the steady-state case of harmonic oscillations. The practical importance of this case is obvious: many technical devices work in the regime of steady-state harmonic oscillations. Nevertheless, many applications show the necessity of investigation of coupled electromechanical fields that have a nonstationary rather than steady-state character. Such problems are typical in the analysis of transient processes in various technical devices.

Certainly, in many applications practical considerations allow us to use some simplifying assumptions in coupled electroelasticity theory. Many methods for the solution of dynamic problems in electroelasticity are based on thickness averaging³ and the use of the Kirchhoff-type hypothesizes. In general, such simplifications may not be appropriate for thin structures, which are important in many applications. One of the typical examples of this type is provided by thin hollow piezoceramic cylinders which are used as active elements in many technical devices. Furthermore, thin hollow cylinders may provide a basis for investigation of electromechanical processes in bones and other biological tissues.

In all such cases, as well as in many other applications, consistent solutions of the coupled nonstationary problem of electroelasticity are required. In its generality, such applications imply that numerical methods become a natural and efficient way to find such consistent solutions. A theoretical framework for the derivation of underlying numerical procedures and their justifications is provided by the concept of generalized solutions.

This paper is organized as follows:

- Section 2 concerns notation and basic preliminaries for the problem.
- In Section 3, we formulate the mathematical model and address the issues related to generalized solutions of the underlying differential problem.
- In Section 4, we derive a computational procedure using the variational approach and the definition of generalized solutions for the problem.
- The main result of Section 5 is an a-priori estimate for the energy integral of the original problem.
- In Section 6, we investigate the stability of the computational model. It is shown that the stability condition can be derived from the discrete analogue of the a-priori estimate obtained in Section 5.

²a-priori imposed on the solution

³usually, for mechanical components of electroelastic fields

- Section 7 provides the proof of convergence of the discrete scheme in the class of generalized solutions $W_2^4(Q_T)$ with second order with respect to space-time discretization.

- Some numerical results are presented in Section 8. Conclusions and future directions are also discussed.

2. PRELIMINARIES AND NOTATION.

In the subsequent sections we deal with a mathematical model of electroelasticity where coupled investigation of electrical and elastic fields under nonstationary conditions is essential to obtain a plausible picture of physical phenomena in a piezoceramic solid. We are interested in the process of coupled electroelastic nonstationary oscillations of a piezoceramic cylinder under radial preliminary polarization [1]. Our main results concern the adequate modelling of such processes for thin hollow cylinders.

The following notation is used throughout the paper.

Mechanical notations:

- u denotes radial displacements of the cylinder;
- ϕ denotes electric field potential;
- ϵ_r and ϵ_θ are components of the field of deformations;
- σ_r and σ_θ are components of the field of stresses;
- E_r denotes the radial component of electric field strength;
- D_r denotes the radial component of electric induction;
- $f_1(r, t)$ denotes a given function for the density of mass forces of solid;
- $f_2(r, t)$ denotes a given function for electric charge density within the solid;
- e_{ij} denote given constants of piezomoduli;
- c_{kl} denote given constants of elastic moduli;
- ϵ_{11} denotes given dielectric permittivity;
- ρ denotes given density of the piezoceramic material;
- $V(t)$ denotes absolute value of given potential functions on the internal and external surfaces of the cylinder ⁴;
- p_0 and p_1 denote given functions of stresses on the internal and external surfaces respectively;
- $u_0(r)$ and $u_1(r)$ denote given functions of displacements and the velocity of their propagation at the initial moment of time respectively.

Mathematical notations:

- $G = (R_0, R_1)$ and $I = (0, T)$ define the range of spacial and temporal variables respectively where T , R_0 , R_1 are assumed to be given (R_0 and R_1 are internal and external cylinder radii); $\bar{I} = [0, T]$;
- $Q_T = I \times G$ defines the space-time region of interest; $Q_{t_1} = \{(r, t) : R_0 < r < R_1, 0 < t < t_1\}$;

⁴it is assumed that electrodes cover the surface of the cylinder and active electric load is given by the potential difference $2V$

- $L^2(Q_T)$ denotes the space of functions which are square integrable in Q_T ; $W_2^k(Q_T)$ denote Sobolev's spaces with an appropriate integer k ($W_2^k(Q_T)$ denotes the Sobolev class of functions with homogeneous boundary conditions);
- M_i denotes appropriate constants in the derivation of a-priori estimates (i is an integer number);
- $\check{\gamma} \equiv \gamma(t - \tau)$ denotes the value of the discrete function γ on the "lower" (with respect to t) time level (τ denotes the time-step of the discretization);
- $\hat{\gamma} \equiv \gamma(t + \tau)$ denotes the value of the discrete function γ on the "upper" time level;
- $\gamma_t \equiv (\gamma(t + \tau) - \gamma(t - \tau))/(2\tau)$ denotes the first central difference of the function γ ;
- $\gamma_{tt} \equiv (\gamma(t + \tau) - 2\gamma(t) + \gamma(t - \tau))/\tau^2$ denotes the second central difference of the function γ ;
- $\gamma_{\bar{t}} \equiv (\gamma(t) - \gamma(t - \tau))/\tau$ denotes the first forward difference of the function γ ;
- $\gamma_{\bar{t}\bar{t}} \equiv (\gamma(t + \tau) - \gamma(t - \tau))/\tau^2$ denotes the first backward difference of the function γ ;
- $\gamma_{\bar{t}\bar{t}} \equiv (\gamma(t + \tau) - 2\gamma(t) + \gamma(t - \tau))/\tau^2$ denotes the second central difference of the function γ .

Other notation is explained in the text.

3. MATHEMATICAL MODEL AND ITS GENERALIZED SOLUTION.

The main results of this paper are obtained for a mathematical model of dynamic electroelasticity for which coupling effect between electrical and elastic fields is important for the adequate (both quantitative and qualitative) description of physical phenomena in a piezoceramic solid.

The process of coupled electroelastic nonstationary oscillations of a piezoceramic cylinder is modelled by a system of partial differential equations in the time-space region Q_T . The system includes the equation of motion of a continuous medium in stress and the Maxwell equation for piezoelectrics:⁶

$$(a) \rho \frac{\partial^2 u}{\partial t^2} = \frac{1}{r} \frac{\partial}{\partial r}(r \sigma_r) - \frac{\sigma_\theta}{r} + f_1(r, t), \quad (b) \frac{1}{r} \frac{\partial}{\partial r}(r D_r) = f_2(r, t). \quad (3.1)$$

The system is supplemented by initial conditions

$$u(r, 0) = u_0(r), \quad \frac{\partial u(r, 0)}{\partial t} = u_1(r), \quad (3.2)$$

and boundary conditions

$$\sigma_r = p_1(t), \varphi = V(t) \text{ for } r = R_0, \text{ and } \sigma_r = p_2(t), \varphi = -V(t) \text{ for } R_1. \quad (3.3)$$

For the radial preliminary polarization (strongly coupled) case the connection between electric and elastic fields are given by the state equations

$$\begin{cases} \sigma_r = c_{11}\epsilon_r + c_{12}\epsilon_\theta - e_{11}E_r, \\ \sigma_\theta = c_{12}\epsilon_r + c_{22}\epsilon_\theta - e_{12}E_r, \\ D_r = e_{11}E_r + e_{12}\epsilon_\theta + e_{11}\epsilon_r, \end{cases} \quad (3.4)$$

⁵analogous notations are used with respect to the spacial variable r and with a spacial-step of discretization h

⁶in the acoustic range of frequencies, it is the forced electrostatic equation of dielectrics

where the relationship between deformations and displacements are given in the Cauchy form, and potential is introduced in its electrostatic form as

$$\epsilon_r = \frac{\partial u}{\partial r}, \quad \epsilon_\theta = \frac{u}{r}, \quad E_r = -\frac{\partial \varphi}{\partial r}. \quad (3.5)$$

We assume non-negativeness for potential energy of deformation, i.e. $\exists \delta > 0$ such that $\forall \xi_1, \xi_2$,

$$\delta(\xi_1^2 + \xi_2^2) \leq c_{11}\xi_1^2 + 2c_{12}\xi_1\xi_2 + c_{22}\xi_2^2. \quad (3.6)$$

Solutions of such non-stationary problems of coupled electroelasticity are of great importance for reliable modelling of many technical devices such, as piezovibrators, several different types of transmitters, generators etc [1], [12], [16]. Other applications arise in different areas of engineering such as hydrodynamics, as well as in biophysics.

Rigorous mathematical investigation of the model involves difficulties caused by the coupling effect. When we rewrite (3.1) using (3.4), (3.5), we have not only a strongly coupled system of partial differential equations ⁷ that consists of two types of operators (hyperbolic and elliptic)

$$\begin{cases} \rho \frac{\partial^2 u}{\partial t^2} = \frac{1}{r} c_{11} \frac{\partial}{\partial r} (r \frac{\partial u}{\partial r}) - c_{22} \frac{u}{r^2} + \frac{1}{r} c_{11} \frac{\partial^2 \varphi}{\partial r^2} - \frac{1}{r} c_{12} \frac{\partial \varphi}{\partial r} + f_1(r, t), \\ -\epsilon_{11} \frac{1}{r} \frac{\partial}{\partial r} (r \frac{\partial \varphi}{\partial r}) + \epsilon_{11} \frac{1}{r} \frac{\partial}{\partial r} (r \frac{\partial u}{\partial r}) + \epsilon_{12} \frac{1}{r} \frac{\partial u}{\partial r} = f_2(r, t), \end{cases}$$

but also a strong coupling effect on the boundary through the boundary conditions for stresses in (3.3). Under simultaneous solution of the electroelasticity system both elastic and electroelastic non-linearities in piezoelectrics are small, ⁸ thus the model (3.1)-(3.5) provides an adequate description of the underlying physical processes. Earlier the correctness of this model was investigated in [10],[12].

In what follows, we develop a technique that allows us to derive efficient computational procedures for investigation of coupling effects in dynamic electroelasticity problems. We are especially interested in cases where coupling effects manifest themselves significantly, having decisive influence decisively on the output characteristics of the designed devices and feedback mechanisms. We shall give a rigorous basis for the derived computational procedure, which can give practically acceptable results even if the solution is not required to be smooth. A general framework for our approach gives the concept of generalized solutions.

According to the general approach to the problems of mathematical physics [7] we introduce the following definition.

Definition 3.1. We call a pair of functions

$$(u(r, t), \phi(r, t)) \in W_2^1(Q_T) \times L^2(I, \overset{0}{W}_2^1(G))$$

⁷it is easy to notice that the second derivative of φ with respect to r is present in the first equation which is primarily responsible for the elastic field; also the second derivative of u with respect to r is present in the second equation which is primarily responsible for the electric field

⁸in particular, such a situation holds for the range of acoustic frequencies [1]

$(u(r,t)$ equals to $u_0(r)$ for $t = 0$) a generalized solution of the coupled problem of dynamic electroelasticity (3.1)-(3.5) if it satisfies the following integral identities:

$$\int_{Q_T} r(-\rho \frac{\partial u}{\partial t} \frac{\partial \eta}{\partial t} + \sigma_r \frac{\partial \eta}{\partial r} + \frac{\sigma_\theta}{r} \eta) dr dt - \int_{R_0}^{R_1} r \rho u_1(r) \eta(r,0) dr = \int_{Q_T} r f_1 \eta dr dt \quad \forall \eta \in \hat{W}_2^1(Q_T), \quad (3.7)$$

$$\int_{R_0}^{R_1} (\epsilon_{11} r \frac{\partial \phi}{\partial r} \frac{\partial \zeta}{\partial r} + \epsilon_{11} r \epsilon_r \frac{\partial \zeta}{\partial r} + \epsilon_{12} r \epsilon_\theta \frac{\partial \zeta}{\partial r}) dr = \int_{R_0}^{R_1} r f_2 \zeta dr \quad \forall \zeta \in W_2^1(G) \quad (3.8)$$

almost everywhere in I . Here $\hat{W}_2^1(Q_T)$ stands for the subspace of $W_2^1(Q_T)$ that consists of all elements of $W_2^1(Q_T)$ which equal zero when $t = T$ (for simplicity we set $p_i(t) = V(t) = 0$, $i = 1, 2$ in the boundary conditions (3.3)).

The differential equations of electroelasticity are a partial case of a more general variational formulation [13], [19] induced by the concept of generalized solutions and physical principles. Moreover, the set of differential equations (3.1)-(3.5) may be obtained from such formulations only under certain (usually excessive) smoothness assumptions. One way to do this is to equate the first variation of the Lagrangean of the electromechanical system to zero or, alternatively, to use different forms of conservation laws to obtain differential forms of the equations. With this reasonings an assumed *a-priori* smoothness for the solution is often questionable in practical applications [18].

Therefore, to have robust numerical procedures it is reasonable to derive computational models directly from the definition 3.1 with generalized solutions. If a discrete model is obtained by an appropriate approximation of the variational functional that involves the energy of the system, the corresponding discrete model will ensure robustness for non-smooth solutions of the problem ⁹.

4. NUMERICAL PROCEDURE FOR FINDING GENERALIZED SOLUTIONS.

Let us assume that generalized second derivatives of the solution are square integrable functions from L^2 (existence and uniqueness of such generalized solutions was proved in [10]). Then the solution $(u(r,t), \phi(r,t))$ satisfies the initial system (3.1)-(3.5) in the sense of the integral identities (3.7), (3.8), as well as the following integral identity

$$\int_{Q_T} r(\rho \frac{\partial^2 u}{\partial t^2} \eta + \sigma_r \frac{\partial \eta}{\partial r} + \frac{\sigma_\theta}{r} \eta) dr dt = \int_{Q_T} r f_1 \eta dr dt, \quad (4.1)$$

where η is an arbitrary element from $W_2^{1,0}(Q_T)$ ¹⁰. Choosing in (4.1) the function

⁹ note also that conservative properties for such discrete models follow from the Noëter theorem [17]

¹⁰this is a Hilbert space that consists of elements $u(r,t) \in L^2(Q_T)$ which have square summable generalized derivatives $\partial u / \partial r$

$\eta(r, t)$ in the form

$$\eta(r, t) \equiv \begin{cases} 0 & \text{for } t \in [t_1, T], \\ \frac{\partial u}{\partial t} & \text{for } t \in [0, t_1), \end{cases}$$

and taking into consideration that ¹¹

$$\int_{Q_T} (\sigma_r \frac{\partial \epsilon_r}{\partial t} + \sigma_\theta \frac{\partial \epsilon_\theta}{\partial t}) dr dt = \int_{Q_T} [c_{11} \epsilon_r \frac{\partial \epsilon_r}{\partial t} + c_{12} (\frac{\partial \epsilon_r}{\partial t} \epsilon_\theta + \frac{\partial \epsilon_\theta}{\partial t} \epsilon_r) + c_{22} \epsilon_\theta \frac{\partial \epsilon_\theta}{\partial t} + \epsilon_{11} \frac{\partial E_r}{\partial t} E_r - \frac{\partial D_r}{\partial t} E_r] dr dt,$$

we obtain the following integral equality to characterize energy (\mathcal{E}) change in the electromechanical system:

$$\begin{aligned} \int_0^{t_1} \frac{d\mathcal{E}}{dt} dt &= \int_{Q_{t_1}} r f_1 \frac{\partial u}{\partial t} dr dt + \int_{Q_{t_1}} r \frac{\partial D_r}{\partial t} E_r dr dt + \\ &\quad \int_0^{t_1} \left[R_1 p_1 \frac{\partial u(R_1, t)}{\partial t} - R_0 p_0 \frac{\partial u(R_0, t)}{\partial t} \right] dt \end{aligned} \quad (4.2)$$

The total energy of the electromechanical system \mathcal{E} can be written as a sum $\mathcal{E} = K + W + P$, where

$$K = \frac{\rho}{2} \int_{R_0}^{R_1} r \left(\frac{\partial u}{\partial t} \right)^2 dr$$

is the kinetic energy,

$$W = \frac{1}{2} \int_{R_0}^{R_1} r [c_{11} \epsilon_r^2 + 2c_{12} \epsilon_r \epsilon_\theta + c_{22} \epsilon_\theta^2] dr$$

is the energy of elastic deformation, and

$$P = \frac{\epsilon_{11}}{2} \int_{R_0}^{R_1} r E_r^2 dr$$

is the energy of electric field of the system.

To find the integral $\int_{Q_{t_1}} r \frac{\partial D_r}{\partial t} E_r dr dt$ we use identity (3.8) integrated in t from 0 to t_1 , setting in it

$$\zeta(r, t) \equiv \begin{cases} 0 & \text{for } t \in [t_1, T], \\ \frac{\partial \phi}{\partial t} & \text{for } t \in [0, t_1). \end{cases}$$

After a simple transformations we have ¹²:

$$-\int_{Q_{t_1}} r \frac{\partial D_r}{\partial t} E_r dr dt = -\int_{Q_{t_1}} r \frac{\partial f_2}{\partial t} \phi dr dt + \int_0^{t_1} r \frac{\partial D_r}{\partial t} \phi|_{R_0}^{R_1} dt \quad (4.3)$$

Taking into consideration the fact that identities (4.2) and (4.3) are satisfied for any $t_1 \in \bar{I}$, we obtain the energy balance identity for a piezoelectric solid:

$$\frac{d\mathcal{E}}{dt} = [R_1 p_1 \frac{\partial u(R_1, t)}{\partial t} - R_0 p_0 \frac{\partial u(R_0, t)}{\partial t}] + \int_{R_0}^{R_1} r f_1 \frac{\partial u}{\partial t} dr +$$

¹¹this equality is a consequence of the state equations (3.4)

¹²using (3.8) with $\zeta = \phi$, and $t = 0, t_1$

$$\int_{R_0}^{R_1} r \phi \frac{\partial f_2}{\partial t} dr + V(t) \left[\frac{\partial D_r(R_1, t)}{\partial t} R_1 + \frac{\partial D_r(R_0, t)}{\partial t} R_0 \right]. \quad (4.4)$$

The right hand side of (4.4) contains those sources that causes dynamic behaviour, i.e. loads on the surface of the body, mass forces and surface charges. It is easy to see that from the relationship (4.4) we can obtain the equation of motion (3.1,a), the Maxwell equation (3.1,b) and the *natural* boundary conditions of the problem (3.1)-(3.5). Below we derive a discrete version of (4.4).

Now let us introduce a difference grid covering the region Q_T :

$$\bar{\omega}_{h\tau} = \bar{\omega}_h \times \bar{\omega}_\tau$$

where

$$\begin{aligned} \bar{\omega}_h &= \{r_i = R_0 + ih, h = \frac{R_1 - R_0}{N}, i = 0, 1, \dots, N\}, \\ \bar{\omega}_\tau &= \{t_j = j\tau, \tau = T/L, j = 0, 1, \dots, L\}. \end{aligned}$$

Let y and μ be the functions of two discrete variables defined on this grid which approximate displacement $u(r, t)$ and electrostatic potential $\varphi(r, t)$ respectively. For each $t \in \bar{\omega}_\tau$ these functions are elements of the Hilbert spaces

$$H_1 = \{y(r) : r \in \bar{\omega}_h\}, \quad H_2^0 = \{\mu(r) : r \in \bar{\omega}_h; \mu = 0, r = R_0, R_1\}$$

with the scalar product $(y, v) = \sum_{\bar{\omega}_h} \hbar y v$, where $\hbar = \frac{h}{2}$ for $i = 0, N$ and $\hbar = h$ for $i = 1, \dots, N-1$. Also let

$$\omega_h^+ = \{r_i = R_0 + ih, i = \overline{1, N}\}, \quad \omega_h^- = \{r_i = R_0 + ih, i = 0, 1, \dots, N-1\}.$$

We shall derive our computational model in two stages.

First we approximate the integral of kinetic energy by the composite trapezoidal rule in space variable r , that is ¹³

$$K^h = \frac{\rho}{2} \sum_{\bar{\omega}_h} \hbar r \left(\frac{d\tilde{u}}{dt} \right)^2, \text{ where } K = K^h + \mathcal{O}(h^2),$$

whereas the integrals of elastic deformation and electric field are approximated by the composite rectangular rule:

$$W^h + P^h = \frac{1}{2} \sum_{\omega_h^+} \hbar \bar{r} [c_{11} \tilde{\epsilon}_r^2 + 2c_{12} \tilde{\epsilon}_r \tilde{\epsilon}_\theta + c_{22} \tilde{\epsilon}_\theta^2 + \epsilon_{11} \tilde{E}_r^2], \text{ where } W + P = W^h + P^h + \mathcal{O}(h^2).$$

where

$$\tilde{\epsilon}_r = \tilde{u}_\tau, \quad \tilde{\epsilon}_\theta = (\tilde{u} + \tilde{u}^{(-1)})/(2\bar{r}), \quad \tilde{E}_r = -\tilde{\phi}_\tau, \quad \tilde{u}^{(\pm 1)} = \tilde{u}(r \pm h, t), \quad \bar{r} = r - h/2, \quad r \in \omega_h.$$

¹³in what follows we denote functions of discrete variable $r \in \bar{\omega}_h$ and continuous $t \in \bar{I}$ by an upper tilde

Then, we approximate the left hand side of (4.4) as follows:

$$\frac{d\tilde{\mathcal{E}}}{dt} = \rho \sum_{\omega_h} \hbar r \tilde{v} \frac{d\tilde{v}}{dt} + \sum_{\omega_h^+} \hbar r [\frac{\partial \tilde{\epsilon}_r}{\partial t} \tilde{\sigma}_r + \frac{\partial \tilde{\epsilon}_\theta}{\partial t} \tilde{\sigma}_\theta] + \sum_{\omega_h^+} \hbar \tilde{r} \tilde{E}_r \frac{\partial \tilde{D}_r}{\partial t},$$

where $\tilde{\mathcal{E}}$ is a differential-difference analog of the total energy of the electromechanical system, $\tilde{v} = d\tilde{u}/dt$, $\tilde{\sigma}_r = c_{11}\tilde{\epsilon}_r + c_{12}\tilde{\epsilon}_\theta - e_{11}\tilde{E}_r$, $\tilde{\sigma}_\theta = c_{12}\tilde{\epsilon}_r + c_{22}\tilde{\epsilon}_\theta - e_{12}\tilde{E}_r$. Now after simple transformations ¹⁴ we obtain a differential-difference analog of the energy identity (4.4):

$$\begin{aligned} & \rho \sum_{\omega_h} \hbar r \tilde{v} \frac{d\tilde{v}}{dt} - \sum_{\omega_h} r \tilde{v} h \frac{1}{r} (\tilde{r} \tilde{\sigma}_r)_r + \sum_{\omega_h^+} r h \tilde{v} \frac{\tilde{\sigma}_\theta}{2r} + \sum_{\omega_h^-} r h \tilde{v} \frac{\tilde{\sigma}_\theta^{(+1)}}{2r} + \sum_{\omega_h^+} \tilde{r} h \tilde{E}_r \frac{\partial \tilde{D}_r}{\partial t} = \\ & [R_1 p_1 \tilde{v}(R_1, t) - R_0 p_0 \tilde{v}(R_0, t)] - \tilde{v}_N \tilde{r}_N (\tilde{\sigma}_r)_N + \tilde{v}_0 \tilde{r}_1 (\tilde{\sigma}_r)_1 + \sum_{\omega_h} \hbar r \tilde{v} f_1 + \sum_{\omega_h} r h \tilde{\phi} \frac{\partial f_2}{\partial t} + \\ & V(t) \left[\tilde{R}_1 \frac{\partial \tilde{D}_r(\tilde{R}_1, t)}{\partial t} + \tilde{R}_0^{(+1)} \frac{\partial \tilde{D}_r(\tilde{R}_0^{(+1)}, t)}{\partial t} \right]. \end{aligned} \quad (4.5)$$

- Assuming \tilde{v} is not identical zero when $\partial \tilde{D}_r / \partial t = \partial f_2 / \partial t = 0$ from (4.5), we can derive a differential-difference analogue of the equation for continuous medium motion, as well as boundary conditions for stresses in (3.3) (since the latter are natural boundary conditions).
- On the other hand, assuming that $\partial \tilde{D}_r / \partial t$ is not identical zero when $\tilde{v} = f_1 = 0$, we can obtain a differential-difference analogue of the Maxwell equation for piezoelectrics:

$$\sum_{\omega_h^+} \tilde{r} h \tilde{E}_r \frac{\partial \tilde{D}_r}{\partial t} = \sum_{\omega_h} r h \tilde{\phi} \frac{\partial f_2}{\partial t} + V(t) [\tilde{R}_1 \frac{\partial \tilde{D}_r(\tilde{R}_1, t)}{\partial t} + \tilde{R}_0^{(+1)} \frac{\partial \tilde{D}_r(\tilde{R}_0^{(+1)}, t)}{\partial t}].$$

If we take into consideration that $\tilde{E}_r = -\tilde{\phi}_r$ and set ¹⁵

$$\tilde{\phi} = V(t) \text{ when } r = R_0, \text{ and } \tilde{\phi} = -V(t) \text{ when } r = R_1,$$

then from the differential-difference analogue of the Maxwell equation can be rewritten as follows:

$$(\tilde{\phi}, (\tilde{r} \frac{\partial \tilde{D}_r}{\partial t})_r) = (\tilde{\phi}, r \frac{\partial f_2}{\partial t}).$$

The second stage of our derivation consists of time-discretisation of the differential-difference scheme obtained with the first stage. Finally, we obtain the discrete space-time scheme for the solution of the problem (3.1)-(3.5) that consists of

¹⁴using grid formulas for summation by parts

¹⁵since boundary conditions for potential are main, they do not follow directly from (3.7)

- the approximation of the equation of motion and boundary conditions for stresses for $t \in \bar{\omega}_r$:

$$\rho y_{tt} = \begin{cases} \frac{1}{r}(\bar{r}\bar{\sigma}_r)_r - \frac{\bar{\sigma}_\theta^{(+1)} + \bar{\sigma}_\theta}{2r} + f_1 & \text{for } r \in \omega_h, \\ \frac{2}{h} \frac{1}{r} \bar{r}^{(+1)} \bar{\sigma}_r^{(+1)} - \frac{\bar{\sigma}_\theta^{(+1)}}{r} + f_1 - \frac{2}{h} p_0 & \text{for } r = R_0, \\ -\frac{2}{h} \frac{1}{r} \bar{r} \bar{\sigma}_r - \frac{\bar{\sigma}_\theta}{r} + f_1 + \frac{2}{h} p_1 & \text{for } r = R_1; \end{cases} \quad (4.6)$$

- the approximation of the Maxwell equation for piezoelectrics and the relationship for electric potential:

$$\frac{1}{r}(\bar{r}\bar{D}_r)_r = f_2, \quad \bar{E}_r = -\mu_r, \quad (4.7)$$

- approximation of the state equations:

$$\begin{cases} \bar{\sigma}_r = c_{11}\bar{\epsilon}_r + c_{12}\bar{\epsilon}_\theta - e_{11}\bar{E}_r, \\ \bar{\sigma}_\theta = c_{12}\bar{\epsilon}_r + c_{22}\bar{\epsilon}_\theta - e_{12}\bar{E}_r, \\ \bar{D}_r = e_{11}\bar{E}_r + e_{12}\bar{\epsilon}_\theta + e_{11}\bar{\epsilon}_r, \end{cases} \quad (4.8)$$

where we approximate the Cauchy relations as follows

$$\bar{\epsilon}_r = (y - y^{(-1)})/h, \quad \bar{\epsilon}_\theta = (y + y^{(-1)})/(2r); \quad (4.9)$$

- the exact boundary conditions for potential function

$$\mu = V(t) \text{ for } r = R_0, \text{ and } \mu = -V(t) \text{ for } r = R_1; \quad (4.10)$$

- the first initial condition:

$$y(r, 0) = u_0(r), \quad (4.11)$$

- and the second initial condition is approximated by the central difference derivative with subsequent elimination of the fictitious time layer for $t = 0$, i.e.

$$\rho y_t = \rho u_1(r) + \frac{\tau}{2} \begin{cases} \frac{1}{r}(\bar{r}\bar{\sigma}_r)_r - \frac{\bar{\sigma}_\theta^{(+1)} + \bar{\sigma}_\theta}{2r} + f_1 & \text{for } r \in \omega_h, \\ \frac{2}{h} \frac{1}{r} \bar{r}^{(+1)} \bar{\sigma}_r^{(+1)} - \frac{\bar{\sigma}_\theta^{(+1)}}{r} + f_1 - \frac{2}{h} p_0 & \text{for } r = R_0, \\ -\frac{2}{h} \frac{1}{r} \bar{r} \bar{\sigma}_r - \frac{\bar{\sigma}_\theta}{r} + f_1 + \frac{2}{h} p_1 & \text{for } r = R_1. \end{cases} \quad (4.12)$$

Investigation of the stability of this discretized scheme can be performed using a difference analogue of the energy identity (4.4). The stability condition for the scheme is provided by the requirement of non-negativeness of the difference analogue of energy integral. Of course, the rate of convergence of such a scheme is virtually defined by the order of approximation, which depends on *a-priori* assumptions for the solution smoothness. Since in many practical applications the solution does not possess continuous derivatives, in order to justify the computational model we should investigate its convergence in classes of generalized solutions.

5. AN A-PRIORI ESTIMATE FOR THE ENERGY INTEGRAL.

Mathematically speaking, under the condition (3.6) we can obtain an *a-priori* estimate for the solution of the problem (3.1)-(3.5) from the energy balance identity (4.4). It should be noted, however, that the presence of time-derivatives of electric induction in (4.4) indicates that in general the equation of forced electrostatics (3.1(b)) should be supplemented by the equation:

$$\nabla \times H = \frac{1}{c_l} \frac{\partial D_r}{\partial t}$$

where H is strength of magnetic field and c_l is the velocity of light [8]. We assume, however, that ¹⁶

$$\frac{\partial D_r}{\partial t} = 0. \quad (5.1)$$

Then the energy balance equation (4.4) can be rewritten in the form

$$\frac{d\mathcal{E}}{dt} = [R_1 p_1 \frac{\partial u(R_1, t)}{\partial t} - R_0 p_0 \frac{\partial u(R_0, t)}{\partial t}] + \int_{R_0}^{R_1} r f_1 \frac{\partial u}{\partial t} dr. \quad (5.2)$$

Integrating (5.2) in t from zero to a certain t_1 ($0 \leq t_1 \leq T$) we obtain

$$\mathcal{E}(t_1) = \mathcal{E}(0) + \int_0^{t_1} [R_1 p_1 v_N - R_0 p_0 v_0] dt + \int_0^{t_1} \int_{R_0}^{R_1} r f_1 \frac{\partial u}{\partial t} dr dt. \quad (5.3)$$

Performing estimates of terms of the right hand part of (5.3) using the Cauchy-Schwarz inequality and embedding theorems ¹⁷, after simple but cumbersome calculations we obtain:

$$\begin{aligned} \mathcal{E}(t_1) &\leq M_1 \mathcal{E}(0) + M_2 \sum_{i=0}^1 [|p_i(t_1)|^2 + |p_i(0)|^2] + M_3 \int_0^{t_1} [(\frac{\partial p_0}{\partial t})^2 + (\frac{\partial p_1}{\partial t})^2] dt + \\ &M_4 \int_0^{t_1} \mathcal{E}(t) dt + M_5 \int_0^{t_1} \int_{R_0}^{R_1} r f_1^2 dr dt. \end{aligned}$$

Using the lemma on integral inequality from the last expression we get:

$$\begin{aligned} \mathcal{E}(t_1) &\leq M_6 \mathcal{E}(0) + M_2 \sum_{i=0}^1 [|p_i(t_1)|^2 + |p_i(0)|^2] + \\ &M_3 \int_0^{t_1} [(\frac{\partial p_0}{\partial t})^2 + (\frac{\partial p_1}{\partial t})^2] dt + M_5 \int_0^{t_1} \int_{R_0}^{R_1} r f_1^2 dr dt. \quad (5.4) \end{aligned}$$

¹⁶it can be assumed, for example, in the acoustic range of frequencies, piezoelectric oscillations are accompanied only by negligible magnetic effects

¹⁷we also need two obvious results: $\frac{\partial u}{\partial t} v = \frac{\partial}{\partial t}(uv) - u \frac{\partial v}{\partial t}$ and $|(u, v)| \leq \|u\| \|v\| \leq \epsilon \|u\|^2 + \frac{1}{4\epsilon} \|v\|^2 \forall \epsilon > 0$; the first one is a consequence of the Newton-Leibniz formula, whereas the second is known as the ϵ -inequality

Estimation of the total energy of the electromechanical system at the initial moment of time $\mathcal{E}(0)$ is made by taking into consideration the initial conditions of the problem. The main difficulty is estimating the functional $\int_{R_0}^{R_1} r E_r^2 |_{t=0} dr$. This functional can be estimated using the Maxwell equation written formally for $t = 0$, multiplied by $r\phi$ and integrated in r from R_0 to R_1 . As a result, taking into consideration (3.6) and (5.4) we have the a-priori estimate

$$\begin{aligned} \mathcal{E}(t_1) \leq & M \left\{ \rho \int_{R_0}^{R_1} r u_1^2 dr + \int_{R_0}^{R_1} r (c_{11}\epsilon_r^2 + 2c_{12}\epsilon_r\epsilon_\theta + c_{22}\epsilon_\theta^2) |_{t=0} dr \right. \\ & + |V(0)|^2 + \sum_{i=0}^1 [|p_i(t_1)|^2 + |p_i(0)|^2] + \int_0^{t_1} \left[\left(\frac{\partial p_0}{\partial t} \right)^2 + \left(\frac{\partial p_1}{\partial t} \right)^2 \right] dt + \\ & \left. \int_{R_0}^{R_1} r \lambda^2 |_{t=0} dr + \int_0^{t_1} \int_{R_0}^{R_1} r f_1^2 dr dt \right\}, \end{aligned} \quad (5.5)$$

where λ is defined from $rf_2 = \partial(r\lambda)/\partial r$, $\lambda|_{r=R_0} = 0$. We formulate the final result as follows.

Lemma 5.1. *If the condition (3.6) holds then the solution of the coupled problem of dynamic electroelasticity (3.1)-(3.5) satisfies the energy inequality (5.5) for all $t_1 \in (0, T]$ and a certain constant $M > 0$. The energy integral in (5.5) is defined as follows:*

$$\mathcal{E}(t) = \frac{\rho}{2} \int_{R_0}^{R_1} r \left(\frac{\partial u}{\partial t} \right)^2 dr + \frac{1}{2} \int_{R_0}^{R_1} r [c_{11}\epsilon_r^2 + 2c_{12}\epsilon_r\epsilon_\theta + c_{22}\epsilon_\theta^2] dr + \frac{\epsilon_{11}}{2} \int_{R_0}^{R_1} r E_r^2$$

The lemma 5.1 is a key factor in establishing a stability condition for the discretized problem (4.6)-(4.12).

6. ANALYSIS OF STABILITY OF THE DISCRETIZED SCHEME.

The main result to be established in this section is a discrete analogue of (5.5). To obtain this result we should first derive an analogue of the energy balance equation (5.2) for the discrete problem (4.6)-(4.12).

Let us take a scalar product between the equation (4.6) ($r \in \omega_h$) and the discrete function $2\tau r y_t$. We sum the result ($\forall r \in (R_0, R_1)$) in i from 1 to $N - 1$, and use (4.6) written for $r = R_0, R_1$, which give approximations to boundary conditions for stresses

$$\begin{aligned} \frac{h}{2} (r 2\tau v) \rho y_{it} &= (2\tau r v) \frac{1}{r} (\bar{r} \bar{\sigma}_r)^{(+)1)} - \frac{h}{2} (2\tau r v) \frac{\bar{\sigma}_\theta^{(+1)}}{r} + \frac{h}{2} (2\tau r v) f_1 - (2\tau r v) p_0, \quad r = R_0 \\ \frac{h}{2} (r 2\tau v) \rho y_{it} &= (2\tau r v) \frac{1}{r} (\bar{r} \bar{\sigma}_r) - \frac{h}{2} (2\tau r v) \frac{\bar{\sigma}_\theta}{r} + \frac{h}{2} (2\tau r v) f_1 - (2\tau r v) p_1, \quad r = R_1, \end{aligned}$$

where for simplicity we denote $v = y_t$.

Then, taking into consideration easily verified identities

$$\begin{aligned}
2\tau \sum_{\omega_h} \rho y_{tt} v r h + 2\tau (\rho y_{tt} v r \frac{h}{2})|_{R_0} + 2\tau (\rho y_{tt} v r \frac{h}{2})|_{R_1} &= \sum_{\omega_h} \hbar r \rho (y_t - y_i) (y_t + y_i), \\
2\tau [- \sum_{\omega_h} \omega_h r h v \frac{1}{r} (\bar{r} \bar{\sigma}_r)_t + v_N \bar{r}_N (\bar{\sigma}_r)_N - v_0 \bar{r}_1 (\bar{\sigma}_r)_1] &= 2\tau \sum_{\omega_h^+} \bar{r} h \bar{\sigma}_r (\bar{\epsilon}_r)_t, \\
2\tau [\sum_{\omega_h} r h v \frac{\bar{\sigma}_\theta^{(+1)} + \bar{\sigma}_\theta^{(-1)}}{2r} + \frac{h}{2} (rv \frac{\bar{\sigma}_\theta^{(+1)}}{r})|_{R_0} + \frac{h}{2} (rv \frac{\bar{\sigma}_\theta}{r})|_{R_1}] &= \\
2\tau [\sum_{\omega_h^+} \bar{r} h \frac{\bar{\sigma}_\theta}{2\bar{r}} v + \sum_{\omega_h^-} \bar{r}^{(+1)} h \frac{\bar{\sigma}_\theta^{(+1)}}{2\bar{r}^{(+1)}} v] &= 2\tau \sum_{\omega_h^+} \bar{r} h \bar{\sigma}_\theta (\bar{\epsilon}_\theta)_t,
\end{aligned}$$

we get

$$\begin{aligned}
\sum_{\omega_h} \hbar r \rho (y_t - y_i) (y_t + y_i) + 2\tau \sum_{\omega_h^+} h \bar{r} [\bar{\sigma}_r (\bar{\epsilon}_r)_t + \bar{\sigma}_\theta (\bar{\epsilon}_\theta)_t] = \\
2\tau [\sum_{\omega_h} \hbar r y_t f_1 + (R_1 p_1 y_t|_{R_1} - R_0 p_0 y_t|_{R_0})].
\end{aligned} \tag{6.1}$$

According to the approximation of state equations (4.8) we have:

$$\bar{\sigma}_r (\bar{\epsilon}_r)_t + \bar{\sigma}_\theta (\bar{\epsilon}_\theta)_t = c_{11} \bar{\epsilon}_r (\bar{\epsilon}_r)_t + c_{12} [\bar{\epsilon}_\theta (\bar{\epsilon}_r)_t + (\bar{\epsilon}_\theta)_t \bar{\epsilon}_r] + c_{22} \bar{\epsilon}_\theta (\bar{\epsilon}_\theta)_t + \epsilon_{11} \bar{E}_r (\bar{E}_r)_t - (\bar{D}_r)_t \bar{E}_r.$$

Also, we can easily verify that

$$2\tau \sum_{\omega_h^+} h \bar{r} \bar{\epsilon}_r (\bar{\epsilon}_r)_t = \mathcal{A} - \frac{\tau^2}{4} \mathcal{B}, \text{ and } 2\tau \sum_{\omega_h^+} h \bar{r} [\bar{\epsilon}_\theta (\bar{\epsilon}_r)_t + (\bar{\epsilon}_\theta)_t \bar{\epsilon}_r] = \mathcal{I}(t + \tau) - \mathcal{I}(t),$$

where

$$\begin{aligned}
\mathcal{A} &= \frac{1}{4} \sum_{\omega_h^+} h \bar{r} (\bar{\epsilon}_r + \bar{\epsilon}_r)^2 - \frac{1}{4} \sum_{\omega_h^+} h \bar{r} (\bar{\epsilon}_r + \bar{\epsilon}_r)^2, \quad \mathcal{B} = \sum_{\omega_h^+} h \bar{r} ((\bar{\epsilon}_r)_t)^2 - \sum_{\omega_h^+} h \bar{r} ((\bar{\epsilon}_r)_t)^2, \\
\mathcal{I}(t) &= \sum_{\omega_h^+} h \bar{r} [\bar{\epsilon}_r \bar{\epsilon}_\theta + \bar{\epsilon}_r \bar{\epsilon}_\theta - \tau^2 (\bar{\epsilon}_r)_t (\bar{\epsilon}_\theta)_t].
\end{aligned}$$

Then, setting ¹⁸ $(\bar{D}_r)_t = 0$ and introducing the discrete analogue of the total energy of the electromechanical system as

$$\begin{aligned}
\bar{\mathcal{E}}(t) &= \rho \sum_{\omega_h} \hbar r y_t^2 + \sum_{\omega_h^+} h \bar{r} \left\{ c_{11} \left[\frac{1}{4} (\bar{\epsilon}_r)^2 + \bar{\epsilon}_r \right] - \frac{\tau^2}{4} ((\bar{\epsilon}_r)_t)^2 \right\} + \\
c_{12} [\bar{\epsilon}_r \bar{\epsilon}_\theta + \bar{\epsilon}_r \bar{\epsilon}_\theta - \tau^2 (\bar{\epsilon}_r)_t (\bar{\epsilon}_\theta)_t] + c_{22} \left[\frac{1}{4} (\bar{\epsilon}_\theta + \bar{\epsilon}_\theta)^2 - \frac{\tau^2}{4} ((\bar{\epsilon}_\theta)_t)^2 \right] + \\
\sum_{\omega_h^+} h \bar{r} \epsilon_{11} \left[\frac{1}{4} (\bar{E}_r + \bar{E}_r)^2 - \frac{\tau^2}{4} ((\bar{E}_r)_t)^2 \right], \tag{6.2}
\end{aligned}$$

¹⁸because of (5.1)

we get the discrete analogue of the energy identity (5.2)

$$\bar{\mathcal{E}}(t + \tau) = \bar{\mathcal{E}}(t) + 2\tau \sum_{\tilde{\omega}_h} \hbar r y_{\tilde{t}} f_1 + 2\tau [R_1 p_1 y_{\tilde{t}}|_{R_1} - R_0 p_0 y_{\tilde{t}}|_{R_0}]. \quad (6.3)$$

The condition of non-negativeness of the discrete analogue of the energy integral gives the stability condition of the scheme (4.6)-(4.12).

Let us estimate the quantity $\bar{\mathcal{E}}(t)$ assuming its non-negativeness for any values of its arguments. Summing (6.3) in t from τ to a certain t_1 ($\tau < t_1 \leq T$), we obtain

$$\bar{\mathcal{E}}(t + \tau) = \bar{\mathcal{E}}(\tau) + \sum_{t'=\tau}^{t_1} 2\tau \sum_{\tilde{\omega}_h} \hbar r y_{\tilde{t}} f_1 + \sum_{t'=\tau}^{t_1} 2\tau [R_1 p_1 y_{\tilde{t}}|_{R_1} - R_0 p_0 y_{\tilde{t}}|_{R_0}]. \quad (6.4)$$

The second term in (6.4) can be estimated with the ϵ -inequality, whereas for the last term we have:¹⁹

$$\begin{aligned} \sum_{t'=\tau}^{t_1} 2\tau [R_1 p_1 v_N - R_0 p_0 v_0] &= \sum_{t'=\tau}^{t_1} 2\tau \{R_1[p_1(y|_{R_1})_{\tilde{t}} - \frac{1}{2}(\hat{y}_N(p_1)_{\tilde{t}} + \check{y}_N(p_1)_{\tilde{t}})] - \\ &\quad \{R_0[p_0(y|_{R_0})_{\tilde{t}} - \frac{1}{2}(\hat{y}_0(p_0)_{\tilde{t}} + \check{y}_0(p_0)_{\tilde{t}})]\}\} = I_1 - I_2 + I_3, \end{aligned}$$

where

$$\begin{aligned} I_1 &= R_1[p_1(t_1+\tau)y(R_1, t_1+\tau) + p_1(t_1)y(R_1, t_1)] - R_0[p_0(t_1+\tau)y(R_0, t_1+\tau) + p_0(t_1)y(R_0, t_1)], \\ I_2 &= R_1[p_1(\tau)y(R_1, \tau) + p_1(0)y(R_1, 0)] - R_0[p_0(\tau)y(R_0, \tau) + p_0(0)y(R_0, 0)], \\ I_3 &= \sum_{t'=\tau}^{t_1} 2\tau [[R_0 \frac{1}{2}(\hat{y}_0(p_0)_t + \check{y}_0(p_0)_t) - R_1 \frac{1}{2}(\hat{y}_N(p_1)_t + \check{y}_N(p_1)_t)]. \end{aligned}$$

Then application of the ϵ -inequality to the term I_1 gives

$$|I_1| \leq \frac{R_1}{4\epsilon_2} \max_{t=t_1, t_1+\tau} |p_1(t)|^2 + \frac{R_0}{4\epsilon_3} \max_{t=t_1, t_1+\tau} |p_0(t)|^2 + R_1 \epsilon_2 |\hat{y}_N + y_N|^2 + R_0 \epsilon_3 |\hat{y}_0 + y_0|^2.$$

And, since

$$|\hat{y}_N + y_N|^2 + |\hat{y}_0 + y_0|^2 \leq M_1 \sum_{\tilde{\omega}_h^+} h \bar{r} \frac{1}{4} (\bar{\epsilon}_r + \check{\epsilon}_r)^2 |_{t=t_1+\tau} \leq M_2 \bar{\mathcal{E}}(t_1 + \tau)$$

we finally have:

$$|I_1| \leq M \left\{ \max_{t=t_1, t_1+\tau} |p_1(t)|^2 + \max_{t=t_1, t_1+\tau} |p_0(t)|^2 + \bar{\mathcal{E}}(t_1 + \tau) \right\}.$$

In a similar manner we conclude that

$$|I_2| \leq M \left\{ \max_{t=0, \tau} |p_1(t)|^2 + \max_{t=0, \tau} |p_0(t)|^2 + \bar{\mathcal{E}}(\tau) \right\},$$

¹⁹here we use that $u_{\tilde{t}} v = (uv)_{\tilde{t}} - (\hat{u}v_t + \check{u}v_{\tilde{t}})/2$

and

$$|I_3| \leq \epsilon_1 R_0 \sum_{t'=\tau}^{t_1} \tau \max_{t'-\tau, t'+\tau} |y_0|^2 + \frac{R_0}{4\epsilon_1} \sum_{t'=\tau}^{t_1} \tau [((p_0)_t)^2 + ((p_0)_{\bar{t}})^2] + \\ \epsilon_2 R_1 \sum_{t'=\tau}^{t_1} \tau \max_{t'-\tau, t'+\tau} |y_N|^2 + \frac{R_1}{4\epsilon_2} \sum_{t'=\tau}^{t_1} \tau [((p_1)_t)^2 + ((p_1)_{\bar{t}})^2].$$

Taking into consideration that

$$\sum_{t'=\tau}^{t_1} \tau \max_{t'-\tau, t'+\tau} (|y_0|^2 + |y_N|^2) \leq M_3 \sum_{t'=0}^T \tau \bar{\mathcal{E}}(t')$$

and applying the discrete analogue of the Gronwal lemma we obtain

$$\begin{aligned} \bar{\mathcal{E}}(t_1 + \tau) &\leq M_4 \bar{\mathcal{E}}(\tau) + M_5 \max_{t=0, \tau, t_1, t_1+\tau} (|p_0(t)|^2 + |p_1(t)|^2) + \\ &M_6 \sum_{t'=\tau}^{t_1} \tau [((p_0)_{\bar{t}})^2 + ((p_1)_{\bar{t}})^2] + M_7 \sum_{t'=\tau}^{t_1} \tau \sum_{\bar{\omega}_h} \hbar r f_1^2. \end{aligned} \quad (6.5)$$

The main difficulty in obtaining an estimate for the discrete analogue of energy integral $\bar{\mathcal{E}}$ at the moment of discrete time $t = \tau$ is the term $\sum_{\bar{\omega}_h} \hbar \bar{r} \epsilon_{11} \bar{E}_r^2(0)$. To estimate it, let us multiply scalarly the discrete analogue of the Maxwell equation (4.7) at $t = 0$ by the discrete approximation of potential μ :

$$((\bar{r} \bar{D}_r)_r, \mu) = (r f_2, \mu).$$

Taking into consideration approximations for the state equations (4.8) and applying the Cauchy-Schwarz inequality (as well as the ϵ -inequality), it is easy to show that

$$\sum_{\bar{\omega}_h} \bar{r} h \bar{E}_r^2(0) \leq M \left\{ \sum_{\bar{\omega}_h} h \bar{r} \bar{\lambda}^2(0) + \sum_{\bar{\omega}_h} h \bar{r} \bar{\epsilon}_r^2(0) + \sum_{\bar{\omega}_h} h \bar{r} \bar{\epsilon}_\theta^2(0) \right\}, \quad (6.6)$$

where $r f_2 = (\bar{r} \bar{\lambda})_r$, $\bar{\lambda}^{(+1)} = \bar{D}_r^{(+1)}$ for $r = R_0$. Then using the definition (6.2), inequality (6.6) and the relationship (3.6) with the inequality (6.5), we obtain the a-priori estimate²⁰

$$\begin{aligned} \bar{\mathcal{E}}(t_1 + \tau) &\leq M \left\{ \rho \sum_{\bar{\omega}_h} \hbar r y_t^2(0) + \sum_{\bar{\omega}_h} h \bar{r} \left\{ c_{11} [\bar{\epsilon}_r^2(0) + \left(\frac{\tau}{2} (\bar{\epsilon}_r)_t(0) \right)^2] + \right. \right. \\ &2c_{12} [\bar{\epsilon}_r(0) \bar{\epsilon}_\theta(0) + \frac{\tau}{2} (\bar{\epsilon}_r(0) (\bar{\epsilon}_\theta)_t(0) + \bar{\epsilon}_\theta(0) (\bar{\epsilon}_r)_t(0))] + c_{22} [\bar{\epsilon}_\theta^2(0) + \left(\frac{\tau}{2} (\bar{\epsilon}_\theta)_t(0) \right)^2] \left. \right\} + \\ &\max_{t=0, \tau, t_1, t_1+\tau} (|p_0(t)|^2 + |p_1(t)|^2) + \sum_{t'=\tau}^{t_1} \tau [((p_0)_{\bar{t}})^2 + ((p_1)_{\bar{t}})^2] + \sum_{\bar{\omega}_h} h \bar{r} \bar{\lambda}^2(0) + \sum_{t'=\tau}^{t_1} \tau \sum_{\bar{\omega}_h} r \hbar f_1^2. \end{aligned} \quad (6.7)$$

²⁰we use also the inequality $\frac{1}{4} (\bar{E}_r + \bar{E}_{\bar{r}})^2(\tau) - \frac{\tau^2}{4} ((\bar{E}_r)_{\bar{t}})^2(\tau) \leq \bar{E}_r^2(0) + \frac{\tau^2}{4} ((\bar{E}_r)_{\bar{t}})^2(0)$, and equality $[\bar{\epsilon}_r \bar{\epsilon}_\theta + \bar{\epsilon}_r \bar{\epsilon}_\theta - \tau^2 (\bar{\epsilon}_r)_{\bar{t}} (\bar{\epsilon}_\theta)_t]_{t=\tau} = 2 \bar{\epsilon}_r(0) \bar{\epsilon}_\theta(0) + \tau [\bar{\epsilon}_r(0) (\bar{\epsilon}_\theta)_t(0) + \bar{\epsilon}_\theta(0) (\bar{\epsilon}_r)_t(0)]$

Let us note that the assumption $V(t) = 0$ adopted here does not restrict the generality of the problem. In fact, introducing a new unknown function $\phi_1 = \phi - L(r, t)$, ²¹ we can easily reduce the original differential problem (3.1)-(3.5) to the problem with homogeneous boundary conditions for the potential. The explicit form of the function $L(r, t)$ is

$$L(r, t) = [\frac{2r}{R_0 - R_1} + \frac{R_1 + R_0}{R_1 - R_0}]V(t).$$

Hence for the problem with homogeneous boundary conditions for potential we should only change the right hand part of the equation of motion and boundary conditions for stresses:

$$f'_1(r, t) = f_1(r, t) + \frac{2e_{12}V(t)}{r(R_1 - R_0)}, \quad p'_i = p_i + \frac{2e_{11}V(t)}{R_1 - R_0}, \quad i = 0, 1.$$

Estimation of the discrete analogue of energy integral of the electromechanical system (6.7) has been obtained under the condition of its non-negativeness. Let us find this condition in an explicit form. Without attracting "deformational" terms to prove non-negativeness of the semi-norm

$$\sum_{\omega_h^+} h\bar{r}[\epsilon_{11}\left(\frac{1}{4}(\bar{E}_r + \dot{\bar{E}}_r)^2 - \frac{\tau^2}{4}((\bar{E}_r)_i)^2\right)](\tau)$$

for the electric field seems to be impossible. Thus, derive an estimate

$$\frac{\tau^2}{4} \sum_{\omega_h^+} h\bar{r}((\bar{E}_r)_i)^2$$

as follows. From the discrete approximation of the Maxwell equation ²²

$$\sum_{\omega_h^+} h\bar{r}\bar{E}_r \bar{D}_r = \sum_{\omega_h^+} h\bar{r}\bar{\lambda}\bar{E}_r$$

we get that

$$\epsilon_{11} \sum_{\omega_h^+} h\bar{r}((\bar{E}_r)_i)^2 \leq \frac{2}{\epsilon_{11}} \{ e_{11}^2 \sum_{\omega_h^+} h\bar{r}((\bar{\epsilon}_r)_i)^2 + e_{12}^2 ((\bar{\epsilon}_r)_i)^2 \}.$$

Here we take into consideration that by virtue of the choice of $\bar{\lambda}$, the Maxwell equation, and the approximation $(\bar{D}_r)_i = 0$ we have that $\bar{\lambda}_i = 0$. As a result, we get the inequality

$$\bar{\mathcal{E}}(t) \geq \rho \sum_{\omega_h^+} h\bar{r}y_i^2 + \sum_{\omega_h^+} h\bar{r}\{[c_{11}\frac{1}{4}(\bar{\epsilon}_r + \dot{\bar{\epsilon}}_r)^2 - \frac{\tau^2}{4}(c_{11} + \frac{2e_{11}^2}{\epsilon_{11}})((\bar{\epsilon}_r)_i)^2] +$$

²¹here $L(r, t)$ is a linear in r that equals $V(t)$ for $r = R_0$, and $-V(t)$ for $r = R_1$

²²and the obvious inequality $(a + b)^2 \leq 2(a^2 + b^2)$

$$c_{12}[\bar{\epsilon}_r \bar{\epsilon}_\theta \dot{\bar{\epsilon}}_r \dot{\bar{\epsilon}}_\theta - \tau^2 (\bar{\epsilon}_r)_t (\bar{\epsilon}_\theta)_t] + [c_{22} \frac{1}{4} (\bar{\epsilon}_\theta + \dot{\bar{\epsilon}}_\theta)^2 - \frac{\tau^2}{4} (c_{22} + \frac{2e_{12}^2}{\epsilon_{11}} ((\bar{\epsilon}_\theta)_t)^2)] + \frac{\epsilon_{11}}{4} \sum_{\omega_h^+} h \bar{r} (\bar{E}_r + \dot{\bar{E}}_r)^2.$$

With regard to (3.6) it is clear that non-negativeness of $\bar{\mathcal{E}}(t)$ will be proved if we prove the inequality

$$\begin{aligned} & \rho \sum_{\omega_h} \hbar r y_t^2 - \tau^2 \left\{ \left[\frac{c_{11}}{4} + \frac{1}{2} \frac{e_{11}^2}{\epsilon_{11}} \right] \sum_{\omega_h^+} h \bar{r} ((\bar{\epsilon}_r)_t)^2 + \right. \\ & \left. \left[\frac{c_{22}}{4} + \frac{1}{2} \frac{e_{12}^2}{\epsilon_{11}} \right] \sum_{\omega_h^+} h \bar{r} ((\bar{\epsilon}_\theta)_t)^2 \right\} + \frac{c_{12}}{2} (\bar{\epsilon}_r)_t (\bar{\epsilon}_\theta)_t \geq 0. \end{aligned} \quad (6.8)$$

Taking into consideration the inequalities

$$\begin{aligned} \sum_{\omega_h^+} h \bar{r} (\bar{\epsilon}_r)^2 &= \sum_{\omega_h^+} h \bar{r} (y_r)^2 \leq \frac{2}{h^2} [\sum_{\omega_h^+} h \bar{r} y^2 + \sum_{\omega_h^+} h \bar{r} (y^{(-1)})^2] = \\ &\frac{2}{h^2} [\sum_{\omega_h^+} h r \frac{\bar{r}}{r} y^2 + \sum_{\omega_h^-} h \bar{r} \frac{\bar{r}^{(+1)}}{r} y^2] \leq \frac{4}{h^2} (1 + \frac{h}{2R_0}) \sum_{\omega_h} \hbar r y^2, \\ \sum_{\omega_h^+} h \bar{r} (\bar{\epsilon}_\theta)^2 &= \sum_{\omega_h^+} h \bar{r} \left(\frac{y + y^{(-1)}}{2\bar{r}} \right)^2 \leq \frac{1}{2} [\sum_{\omega_h^+} h \frac{y^2}{\bar{r}} + \sum_{\omega_h^+} h \frac{(y^{(-1)})^2}{\bar{r}}] = \\ &\frac{1}{2} [\sum_{\omega_h^+} h r \frac{1}{\bar{r}r} y^2 + \sum_{\omega_h^-} h \bar{r} \frac{1}{\bar{r}^{(+1)}r} y^2] \leq \frac{1}{R_0^2} \sum_{\omega_h} \hbar r y^2, \end{aligned}$$

and

$$\sum_{\omega_h^+} \bar{r} h \bar{\epsilon}_r \bar{\epsilon}_\theta = \frac{1}{h} \sum_{\omega_h^+} h r \frac{y^2 - (y^{(-1)})^2}{2\bar{r}} \leq \frac{1}{2R_0 h} \sum_{\omega_h} \hbar r y^2$$

we conclude that (6.8) is satisfied if the inequality

$$\rho - \tau^2 \left[\frac{4}{h^2} (1 + \frac{h}{2R_0}) \left(\frac{c_{11}}{4} + \frac{e_{11}^2}{2\epsilon_{11}} \right) + \frac{1}{R_0^2} \left(\frac{c_{22}}{4} + \frac{e_{12}^2}{2\epsilon_{11}} \right) + \frac{c_{12}}{4R_0 h} \right] \geq \epsilon, \text{ where } \epsilon > 0,$$

holds. If we note that

• $\delta = e_{11}^2 / (\epsilon_{11} c_{11})$ is the coupling coefficient of the electromechanical system ²³ described by the model (3.1)-(3.5), and

• $c = \sqrt{c_{11}(1 + \delta)/\rho}$ is the velocity of propagation of mixed electro-elastic waves that are the solution of the coupled dynamic problem of electroelasticity,

²³it characterizes the effect of power transformation in piezoelectric material better than the set of elastic, dielectric and piezoelectric constants [1]

then the stability condition for the discrete scheme (4.6)-(4.12) can be finally written in the form:

$$\tau \leq \frac{h}{c} \left\{ \left(1 - \frac{\epsilon}{\rho} \right) / \left[\left(1 + \frac{h}{2R_0} \right) \left(1 + \frac{\delta}{1+\delta} \right) + \frac{c_{12}}{4R_0 c_{11}(1+\delta)} h + \frac{h^2}{4R_0^2 c_{11}(1+\delta)} \left(c_{22} + \frac{2\epsilon_{12}^2}{\epsilon_{11}} \right) \right] \right\}^{\frac{1}{2}}. \quad (6.9)$$

This completes the proof of the following theorem:

Theorem 6.1. *Under the stability condition (6.9), the solution of the discrete problem (4.6)-(4.12) satisfies the energy estimate (6.7), where the discrete analogue of energy integral is defined by (6.2).*

Remark 6.1. The stability condition (6.9) for zero coupling coefficient ²⁴ coincides in the dominant part with the stability condition for a discrete scheme obtained for non-coupled problem of electroelasticity [14].

Remark 6.2. Of course, the stability condition (6.9) has a quite definite physical meaning. In the case of circular preliminary polarization, when mechanical and electric fields appear to be uncoupled, the velocity of propagation of pure elastic waves $c_0 = \sqrt{c_{11}/\rho}$ defines the stability of discrete schemes [14]. Whereas in the case of radial preliminary polarization the coupling effect manifests itself and time-step-discretization depends on the velocity of propagation of mixed electro-elastic waves $c = \sqrt{c_{11}(1+\delta)/\rho}$. Hence, the derived condition gives a Courant-Friderichs-Lewy-type stability condition for the case of coupled dynamic electroelasticity.

7. CONVERGENCE OF THE SCHEME ON THE CLASS OF GENERALIZED SOLUTIONS FROM $W_2^3(Q_T)$.

The error of the scheme (4.6)-(4.12)

$$z = y - u, \quad \zeta = \mu - \phi$$

is the solution of the operator-difference scheme

$$\begin{cases} D_1 z_{tt} + A_1 z + C_1 \zeta = \psi, & t \in \omega_\tau, \\ A_2 \zeta + C_2 z = \kappa, & t \in \omega_\tau, \\ z = 0, \quad D_1 z_t = \psi, & t = 0. \end{cases} \quad (7.1)$$

The functions $z = z(t)$, $\zeta = \zeta(t)$ for each $t \in \bar{\omega}_\tau$ are elements of the Hilbert spaces defined in section 4, whereas the operators of the scheme are defined as

$$A_1 z = \begin{cases} -\frac{2}{h} \bar{r}^{(+1)} \check{\sigma}_r^{(+1)} + \check{\sigma}_\theta^{(+1)}, & r = R_0, \\ -(\bar{r} \check{\sigma}_r)_r + \frac{\check{\sigma}_\theta^{(+1)} + \check{\sigma}_\theta}{2}, & R_0 < r < R_1, \\ \frac{2}{h} \bar{r} \check{\sigma}_r + \check{\sigma}_\theta, & r = R_1, \end{cases}$$

²⁴i.e. $\delta = 0$ when coupling between electric and elastic fields is negligible

$$C_1\zeta = \begin{cases} \frac{2e_{11}}{h}\bar{r}^{(+1)}\check{E}_r^{(+1)} - e_{12}\check{E}_r^{(+1)}, & r = R_0, \\ e_{11}(\bar{r}\check{E}_r)_r - e_{12}\frac{\check{E}_r^{(+1)} + \check{E}_r}{2}, & R_0 < r < R_1, \\ -\frac{2e_{11}}{h}\bar{r}\check{E}_r + e_{12}\check{E}_r, & r = R_1, \end{cases}$$

$$D_1 z = r\rho z, \quad A_2 \zeta = e_{11}(\bar{r}\check{E}_r)_r, \quad C_2 z = [\bar{r}(e_{12}\check{\epsilon}_\theta + e_{11}\check{\epsilon}_r)]_r,$$

where

$$\check{\sigma}_r = c_{11}\check{\epsilon}_r + c_{12}\check{\epsilon}_\theta, \quad \check{\sigma}_\theta = c_{12}\check{\epsilon}_r + c_{22}\check{\epsilon}_\theta, \quad \check{E}_r = -\zeta_r, \quad \check{\epsilon}_r = z_r, \quad \check{\epsilon}_\theta = \frac{z + z^{(-1)}}{2\bar{r}}.$$

The approximation error of the scheme (4.6)-(4.12) is defined as

$$\psi = -\rho r u_{tt} + \check{\psi} \text{ for } t \in \omega_\tau, \text{ and } \psi = \rho r u_t - \rho r u_t + \frac{\tau}{2}\psi \text{ for } t = 0,$$

$$\kappa = -(\bar{r}\check{D}_r)_r = -(\bar{r}(e_{11}\check{E}_r + e_{12}\check{\epsilon}_\theta + e_{11}\check{\epsilon}_r))_r,$$

where

$$\begin{aligned} \check{\psi} &= \begin{cases} (\bar{r}\check{\sigma}_r)_r - \frac{\check{\sigma}_\theta^{(+1)} + \check{\sigma}_\theta}{2} + rf_1, & \text{for } r \in \omega_h, \\ \frac{2}{h}\bar{r}^{(+1)}\check{\sigma}_r^{(+1)} - \check{\sigma}_\theta^{(+1)} + rf_1 - \frac{2}{h}rp_0, & \text{for } r = R_0, \\ -\frac{2}{h}\bar{r}\check{\sigma}_r - \check{\sigma}_\theta + rf_1 + \frac{2}{h}rp_1, & \text{for } r = R_1, \end{cases} \\ \check{\sigma}_r &= c_{11}\check{\epsilon}_r + c_{12}\check{\epsilon}_\theta - e_{11}\check{E}_r, \quad \check{\sigma}_\theta = c_{12}\check{\epsilon}_r + c_{22}\check{\epsilon}_\theta - e_{12}\check{E}_r, \\ \check{E}_r &= -\phi_r, \quad \check{\epsilon}_r = u_r, \quad \check{\epsilon}_\theta = \frac{u + u^{(-1)}}{2\bar{r}}. \end{aligned}$$

If the sought for solution of problem (3.1)-(3.5) belongs to the Sobolev space $W_2^4(Q_T)$ the use of Taylor's expansion with the integral form of the remainder leads to the conclusion that the error of approximation for any $t \in \bar{\omega}_\tau$ can be presented in the form

$$\psi = \check{\psi} + \delta(h)\check{\psi},$$

where

$$\delta(h) = \begin{cases} 0, & \text{for } r \in \omega_h, \\ 2/h, & \text{for } r = R_0, \\ -2/h, & \text{for } r = R_1, \end{cases}$$

and that functionals $\check{\psi}, \check{\psi}, \kappa$ have the second order of smallness in grid steps, i.e.

$$\check{\psi} = \mathcal{O}(h^2 + \tau^2), \quad \check{\psi} = \mathcal{O}(h^2 + \tau^2), \quad \kappa = \mathcal{O}(h^2).$$

Now to obtain an accuracy estimate for the discrete scheme (4.6)-(4.12) we use the a-priori estimate (6.7) established under the condition (6.9). After transformations we have the following accuracy estimate:

$$\check{\mathcal{E}}(t_1 + \tau) \leq M \left\{ \sum_{\omega_h} hr\check{\psi}(r, 0) + \sum_{\omega_h^+} h\bar{r}[c_{11}\check{\epsilon}_r^2(0) + 2\check{\epsilon}_r(0)\check{\epsilon}_\theta(0) + c_{22}\check{\epsilon}_\theta^2(0)] + \right.$$

$$\begin{aligned} & \max_{t=0, \tau, t_1, t_1+\tau} (\|\dot{\psi}(R_0, t)\|^2 + \|\dot{\psi}(R_1, t)\|^2) + \sum_{t'=\tau}^{t_1} \tau [(\dot{\psi}_r(R_0, t'))^2 + (\dot{\psi}_l(R_1, t'))^2] + \\ & \sum_{t'=\tau}^{t_1} \tau \sum_{\omega_h^+} \omega_h h r \ddot{\psi}^2(r, t') + \sum_{\omega_h^+} h \bar{r} \kappa_1^2(0), \end{aligned} \quad (7.2)$$

where $\check{\mathcal{E}}(t_1 + \tau)$ is obtained by the replacement in $\check{\mathcal{E}}(t_1 + \tau)$ all y on z , and ϕ on ζ ; $r\kappa = (\bar{r}\kappa_1)_r$; $\kappa_1^{(+1)} = \check{D}_r^{(+1)}$ when $r = R_0$, and the functions $\check{\epsilon}_r, \check{\epsilon}_\theta, \kappa_1$ in the right-hand part (7.2) are computed for $t = 0$. Of course, when the sought for solution is from the class $W_2^4(Q_T)$, both quantities $\dot{\psi}_l$ and κ_1 have also the second order of smallness with respect to grid steps. Hence, we come to the main result of this section.

Theorem 7.1. If the stability condition (6.9) holds the solution of the discrete problem (4.6)-(4.12) converges to the sought for solution of the original problem (3.1)-(3.5) from the class $W_2^4(Q_T)$ with the second order with respect to space-time grid steps. Such solutions satisfy the accuracy estimate (7.2).

Remark 7.1. If the equation of motion and the Maxwell equation are coupled between themselves only by the state equations, but are not coupled by the boundary conditions for stresses,²⁵ then the problem is essentially simplified. In such a case, a stronger result than that stated in theorem 7.1 can be obtained. Namely, if the generalized solution of the original problem is assumed to be from the class $W_2^2(Q_T)$, then the second order of convergence can be preserved for the discrete scheme in a weaker than L^2 metric. Results of this type for a single wave equation were discussed earlier in [4].

8. CONCLUSIONS.

In the general non-stationary case, accounting for coupling effects between electric and elastic fields in anisotropic materials may essentially influence the results of computations. Steep gradients of computed functions require adequate mathematical tools to deal with such phenomena. The concept of generalized solutions provides a unified framework for derivation of such computational models as well as their justification.

The model we have considered is typical in coupled field theory where inter-influence of *physical fields different nature* is essential to obtain a plausible picture of the underlying phenomena. Usually, at least one of the equations in the model is a partial differential equation of *hyperbolic* type. A connection of such a hyperbolic-type operator with elliptic and/or parabolic modes of the model leads to a situation where the application of numerical methods become natural and the most efficient way of solving problems arising from coupled field theory. Mathematical modelling in this area requires approaches which can be applied even if a solution of the problem

²⁵for example, displacements are given on the boundary

does not possess an excessive smoothness often imposed as an *a-priori* assumption. Mathematical challenges and practical importance of the problems stimulate interest in them from mathematicians, engineers and scientists.

Of course, the approach based on the coupling procedures is a natural way to reflect *additional information* about the system and to implement it into mathematical models. However, in mathematical modelling and computational experiments we always use the implicit assumption that the problem can effectively be reduced to a finite set of equations, inequalities or inclusions. Though such an approach will remain a powerful tool for investigation of the real-world problems in the foreseeable future, it is very important to pay attention to models of varying complexity. Fixing a degree of coupling in mathematical models implies thorough investigation of system stability under the *specified coupling level*.

All real process, dynamic systems, and phenomena describe a transformation of different types of energy, which implies that, in general, mathematical models applied to them should have integral rather than differential features. Clearly, for example, a border between two different media might not be described appropriately by any differential equation due to a jump of physical parameters. A similar situation arises when we try to describe a nonhomogeneous medium. Probably one of the most demonstrative examples of difficulties involved in mathematical modelling of such media is provided by non-local type models. Along with classical applications of such models for the description of macrosystems as well as microstructures²⁶ non-local type models are typical when we address physical problems of mathematical modelling using an extended thermodynamic approach [5], [15]. The hyperbolic nature of arising models provide an important area of further investigations. In general, many problems in coupled field theory²⁷ may not obtain an adequate description in mathematical models if *a-priori* assumptions of excessive²⁸ smoothness are imposed on their solutions.

A connection between variational principles and computational models for new areas of applications provides many challenging problems [20]. On the other hand, linear models in coupled field theory give important guidelines for further development theory in nonlinear case. Recently, mixed²⁹ variational principles were proposed in nonlinear electroelasticity [19]. Accounting for discontinuities and steep gradients in the solutions, it is important to be able to achieve a trade-off between *a-priori* assumptions on solution smoothness and computational efficiency of the numerical procedures derived from physical principles.

As an ultimate goal we would like to generalize the presented technique to non-local models arising from micro- and macro- levels of description of real dynamic systems. Some results in this field have been already published in [9].

²⁶as in climate modelling and semiconductor device simulation

²⁷arising from studying microstructures as well as macrosystems

²⁸with respect to real solutions

²⁹in a sense that all main variables such as stresses, electric field, displacements and electric potential are considered as field variables

ACKNOWLEDGEMENT.

I wish to acknowledge the support of the School of Mathematics at the University of South Australia and to thank Dr Stephen Lukas for helpful assistance at the final stage of preparation of this paper.

REFERENCES

- [1] Berlincourt,D.A., Curran,D.R., and Jaffe, H.: Piezoelectric and Piezomagnetic materials and their function in transducers, in *Physical Acoustics*, Vol.1A, pp.204-236, ed., W.P. Mason, New York and London: Academic Press.
- [2] Carthy M.F. and Tiersten H.F.: The growth of wave discontinuities in piezoelectric semiconductors. *J.Math.Phys.*, 20, No.12, 2682-2691 (1979).
- [3] Daher, N.: A continuum energy approach of deformable piezoelectric and ferroelectric semiconductors. *Proc. of 1994 IEEE Ultrasonics Symposium, France*, Vol.2, 701-704 (1994).
- [4] Djuraev, I.N., and Moskalkov, M.N.: Convergence of difference scheme solution to the generalized solution of the wave equation from $W_2^2(Q_T)$. *Differential Equations*, 21, 2145-2152 (1985).
- [5] Jou, D., Casa-Vasquez, J and Lebon, G.: *Extended Irreversible Thermodynamics*, Springer-Verlag, 1993.
- [6] Kulkarni, G. and Hanagud, S.: Modelling issues in the vibration control with piezoceramic actuators, in *Smart Structures and Materials*, pp.7-15, MD-Vol.24/AMD-Vol 123, ASME, New York, 1991.
- [7] Ladyzhenskaya, O.A.: *Boundary-value problems in Mathematical Physics*, Springer-Verlag, 1985.
- [8] Landau, L.D. and Lifshiz, E.M.: *Theoretical Physics, Vol. VIII: Electrodynamics of Continuous Media*, Moscow, Nauka, 1982.
- [9] Melnik, V.N.: Nonconservation law equation in mathematical modelling: aspects of approximation. *Proc. of the Int. Conf. AEMC'96, Sydney*, 423-430 (1996).
- [10] Melnik, V.N.: Existence and uniqueness theorems of the generalized solution for a class of non-stationary problems of coupled electroelasticity. *S. Mathematics (Iz.VUZ)*, 35, 24-32; Allerton Press, 23-30 (1991).
- [11] Melnik, V.N. and Moskalkov, M.N.: On the coupled non-stationary electroelastic oscillations of a piezoceramic cylinder with radial polarization. *Computational Mathematics and Mathematical Physics*, Pergamon Press, 109-110 (1990).
- [12] Melnik, V.N. and Moskalkov, M.N.: Difference schemes for and analysis of approximate solutions of two-dimensional nonstationary problems in coupled electroelasticity. *Differential Equations*, Plenum Publishing,C/B, New York, 860-867 (1992).
- [13] Mindlin, R.: Equations of high frequency vibrations of thermopiezoelectric crystal plates. *Int.J.Solids and Structure*, 10, 625-637 (1974).

- [14] Moskalkov, M.N.: Investigation of a difference scheme for solution of the sound radiator problem for cylindric piezovibrator. *Differential Equations*, 22(7), 1220-1226 (1986).
- [15] Muller, I., Ruggeri, T.: *Extended Thermodynamics*, Springer-Verlag, 1993.
- [16] Nowacki, W.: *Electromagnetic Effects in Solids*, Moscow, Mir Publishers, 1986.
- [17] Samarskii, A.: Some results in theory of difference schemes, *Differential Equations*, Vol.16, No.7, 1155-1171 (1980).
- [17] Samarskii, A., Lazarov, R. and Makarov, V.: *Difference Schemes for Differential Equations with Generalized Solutions*. Moscow, VS Publishers, 1987.
- [18] Yang, J.S. and Batra, R.C.: Mixed variational principles in non-linear electroelasticity. *Int. J. Non-Linear Mechanics*, 30(5), 719-725, (1995).
- [19] Yu Y.Y. Some recent advances in linear and nonlinear dynamical modeling of elastic and piezoelectric plates. *J. of Intelligent Material Systems and Structures*, 6, 237-254, (1995).

UNIVERSITY OF SOUTH AUSTRALIA



TECHNICAL REPORT SERIES

**Analysis of Convergence of the
Operator-Difference Scheme for
Solution of a Nonstationary Problem
Arising from Coupled Field Theory**

by

(R.) V Nick Melnik

Report No. 1996/5

SCHOOL OF MATHEMATICS

Faculty of Applied Science and Technology

The Levels, South Australia 5095, Telephone (08) 302 3343 Facsimile (08) 302 3381

TECHNICAL REPORT SERIES

**Analysis of Convergence of the
Operator-Difference Scheme for
Solution of a Nonstationary Problem
Arising from Coupled Field Theory**

by

(R.) V Nick Melnik

Report No. 1996/5

ANALYSIS OF CONVERGENCE OF THE OPERATOR-DIFFERENCE SCHEME FOR SOLUTION OF A NONSTATIONARY PROBLEM ARISING FROM COUPLED FIELD THEORY¹

V. Nick Melnik

School of Mathematics, University of South Australia,
The Levels Campus, S.A., 5095, Australia
E-mail: matvnm@lv.levels.unisa.edu.au

Abstract

In this paper the operator-difference scheme for the numerical solution of a problem arising from coupled field theory is thoroughly investigated for the case when the classical assumptions of sufficient smoothness cannot be applied. Such a situation, being typical in many applications, is considered for the problem of nonstationary electroelasticity.

A new *a priori* estimation for the numerical solution of the problem has been obtained. A scale of accuracy results for generalized solutions of the problem has been derived, and the convergence theorem has been proved.

Key words: operator-difference scheme, coupled field theory, generalized solutions, CFL condition for electroelastic waves.

AMS subject classifications: 35A40, 35Q72, 65M06, 70G99.

1. Introduction.

In many areas of human endeavour, related to applications of mathematical models, we can often observe a gap between imposed theoretical assumptions on a solution smoothness and an actual smoothness of the solution in a real practical problem.

A necessity of relaxation of such assumptions is a typical feature of many problems where interconnection of *different nature physical fields* is essential in obtaining a plausible picture of the phenomenon under consideration. Such problems are subject

¹The main results of this paper are submitted to *Journal of Difference Equations and Applications*, Gordon and Breach Publishers

to *coupled field theory*. One of the classical examples of this is thermoelasticity [15], where the electric and thermal fields combine into a unified whole which in general cannot be separated. An efficient way to solve the problem in such cases implies the concept of generalized solutions [19] which is intrinsic to coupled field theory and computational models used for the solution of arising problems.

In fact, all real process, dynamical systems and phenomena describe a transformation of different types of energy which implies that, in general, mathematical models applied to them should have integral rather than differential features. Clearly, for example, a border between two different media might not be described by any differential equation due to a jump of physical parameters. A similar situation arises when we try to describe a nonhomogeneous medium. Probably one of the most demonstrative examples of difficulties involved in mathematical modelling of such media are the non-local models arising from climate modelling and semiconductor device simulation (which, in fact, are quite similar to each other from mathematical and computational perspectives) when hydrodynamic (or quasihydrodynamic) type of models is applied (see, for example,[10]).

In general, many problems in coupled field theory ² do not obtain an adequate description in mathematical models if assumptions of excessive smoothness are imposed on their solutions.

Typically, coupled field theory in the nonstationary case deals with systems of partial differential equations (PDE) which do not belong to any classical type of PDEs, yet at least one of the equations of such a system is a PDE of hyperbolic type. Since there is a connection between the hyperbolic (in general, dissipative) equation with PDE of parabolic (for thermoelasticity) or elliptic (for electroelasticity) type ³ analytical solutions of such problems are quite exceptional. This leads to a situation where numerical methods become natural and the most efficient way of solving problems arising from coupled field theory. Mathematical challenges and practical importance of the problems stimulate interest to them from mathematicians, engineers and scientists.

In this paper we deal with a nonstationary problem of coupled electroelasticity the solution of which is of great importance for reliable work of many technical devices such as piezovibrators, different types of transmitters, generators etc (see, for example, [1, 12, 16]). It is also believed that the technique proposed below is applicable to a much wider class of problems arising from coupled field theory.

The paper is organized as follows.

- In Section 2 we state the differential formulation of the problem and point out the difficulties involved in its solution.
- In Section 3 we consider the operator-difference scheme for solution of the problem constructed under variational approach. In this section we explicitly derive a stability condition for such a scheme which can be seen as a generalization of the classical Courant-Friedrichs-Lowy (CFL) condition for the case of coupled electro-elastic waves. A new *a priori* estimation is also obtained in this section.

²arising from studying microstructures as well as macrosystems

³obviously, there are much more complicated cases [10]

- Section 4 is devoted to questions of convergence where we prove a scale of accuracy estimations for the difference problem (considered in section 3) when the solution of the differential problem is from defined generalized classes.

2. Mathematical Model of Coupled Electroelasticity in the Nonstationary Case.

Let us consider a mathematical model of electroelasticity where coupled investigation of electrical and elastic fields under nonstationary conditions is essential to obtain a plausible quantitative (as well as qualitative) picture of physical phenomena in a piezoceramic solid. The existence and uniqueness theorems for the mathematical models of this type as well as numerical experiments on applications can be found in [8, 11, 12].

The process of coupled electroelastic nonstationary oscillations of a piezoceramic cylinder can be described by the system of partial differential equations in the time-space region $Q_T = \{(r, t) : R_0 < r < R_1, 0 < t \leq T\}$ (see, for example, [11]):

$$\rho \frac{\partial^2 u}{\partial t^2} = \frac{1}{r} \frac{\partial}{\partial r}(r \sigma_r) - \frac{\sigma_\theta}{r} + f_1 \quad (2.1)$$

$$\frac{1}{r} \frac{\partial}{\partial r}(r D_r) = f_2 \quad (2.2)$$

which should be completed by initial:

$$u(r, 0) = u_0(r), \frac{\partial u(r, 0)}{\partial t} = u_1(r) \quad (2.3)$$

and boundary conditions:

$$\sigma_r = 0, \varphi = 0 \text{ for } r = R_0, R_1. \quad (2.4)$$

The most difficult (yet the most interesting for practice) case is the radial preliminary polarization. The connection between electric and elastic fields is very strong (see [11,12] for details):

$$\left. \begin{array}{l} \sigma_r = c_{11}\epsilon_r + c_{12}\epsilon_\theta - \epsilon_{11}E_r, \\ \sigma_\theta = c_{12}\epsilon_r + c_{22}\epsilon_\theta - \epsilon_{12}E_r, \\ D_r = \epsilon_{11}E_r + \epsilon_{12}\epsilon_\theta + \epsilon_{11}\epsilon_r. \end{array} \right\} \quad (2.5)$$

There remain only Cauchy relations and formula for electrostatic potential φ to be added to complete the problem formulation:

$$\epsilon_r = \frac{\partial u}{\partial r}, \quad \epsilon_\theta = \frac{u}{r}, \quad E_r = -\frac{\partial \varphi}{\partial r}. \quad (2.6)$$

In (2.1)-(2.6) we use the following notations: u is the radial displacement, E_r and D_r are radial components of electric field strength and electric induction respectively, c_{kl}

are elastic moduli, c_{ij} are piezomoduli, ϵ_{11} is the dielectric permittivity, ρ is the density of piezoceramic material, and f_1, f_2 are density of mass forces and electric charge density of solid respectively. We also assume nonnegativeness of potential energy of deformation, e.g.

$$\delta(\xi_1^2 + \xi_2^2) \leq c_{11}\xi_1^2 + 2c_{12}\xi_1\xi_2 + c_{22}\xi_2^2,$$

which is fulfilled for $\delta > 0$ and $\forall \xi_1, \xi_2$. Therefore we have coupled system of partial differential equations of the type

$$\rho \frac{\partial^2 u}{\partial t^2} = \frac{1}{r} c_{11} \frac{\partial}{\partial r} \left(r \frac{\partial u}{\partial r} \right) - c_{22} \frac{u}{r^2} + \frac{1}{r} \epsilon_{11} \frac{\partial^2 \varphi}{\partial r^2} - \frac{1}{r} \epsilon_{12} \frac{\partial \varphi}{\partial r} + f_1, \quad (2.7)$$

$$- \epsilon_{11} \frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial \varphi}{\partial r} \right) + \epsilon_{11} \frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial u}{\partial r} \right) + \epsilon_{12} \frac{1}{r} \frac{\partial u}{\partial r} = f_2 \quad (2.8)$$

which is strongly connected. In fact, we have the second derivative with respect to φ in the equation (2.7) (this equation is mainly responsible for the elastic field) and the second derivative with respect to u in the equation (2.8) (this equation is mainly responsible for the electric field). This connection of the equations is amplified by the boundary conditions for tensions (2.4).

In many cases to obtain efficient finite difference schemes for the solution of coupled field theory problems we can use variational principles. For example, the Biot variational principle can be of great help in the coupled thermoelasticity, whereas in coupled electroelasticity a similar role plays the conservation energy law for the whole electromechanical system [13]. Earlier we applied these ideas to construct the difference schemes for coupled nonstationary thermo- and electroelasticity problems [9,11,12]. In this paper we are interested in solutions that do not possess such high smoothness (for example, $C_4(Q_T)$) as it is often assumed (see, for example, [11]). Below we develop a technique which is similar to that in [9] where coupled problems of thermoelasticity were the subject of discussion. It was shown by computational experiments [12] that the coupling effect in the case of electroelasticity can manifest significantly and can influence decisively on the output characteristics of the designed devices. It is necessary that the development of such approaches to the solution of the problem can give practically acceptable results even if the solution is not obligatory smooth. A general framework for such approaches gives the concept of generalized solutions.

3. Difference Scheme and a New *A Priori* Estimation for the Numerical Solution.

A discrete space-time analogue of the problem (2.1)-(2.6) can be obtained by the application of the variational approach and the finite difference method (see [11,12] for details) which give the following operator-difference scheme:

$$D_1 y_{it} + A_1 y + C_1 \mu = \varphi_1, \quad (3.1)$$

$$A_2\mu + C_2y = \varphi_2, \quad (3.2)$$

$$y = y_0, D_1y_t = y_1, t = 0, \quad (3.3)$$

where operators of this scheme is defined in the following way:

$$A_1y = \begin{cases} -\frac{2}{h}\bar{r}^{(+1)}\bar{\sigma}_r^{(+1)} + \bar{\sigma}_\theta^{(+1)}, & r = R_0 \\ -(\bar{r}\bar{\sigma}_r)_r + \frac{\bar{\sigma}_\theta^{(+1)} + \bar{\sigma}_\theta}{2}, & R_0 < r < R_1 \\ \frac{2}{h}\bar{r}\bar{\sigma}_r + \bar{\sigma}_\theta, & r = R_1 \end{cases}$$

$$C_1\mu = \begin{cases} \frac{2e_{11}}{h}\bar{r}^{(+1)}\bar{E}_r^{(+1)} - e_{12}\bar{E}_r^{(+1)}, & r = R_0 \\ e_{11}(\bar{r}\bar{E}_r)_r - e_{12}\frac{\bar{E}_r^{(+1)} + \bar{E}_r}{2}, & R_0 < r < R_1 \\ -\frac{2e_{11}}{h}\bar{r}\bar{E}_r + e_{12}\bar{E}_r, & r = R_1 \end{cases}$$

$$D_1y = r\rho y, A_2\mu = \epsilon_{11}(\bar{r}\bar{E}_r)_r, C_2y = [\bar{r}(e_{12}\bar{\epsilon}_\theta + e_{11}\bar{\epsilon}_r)]_r, \bar{r} = r - \frac{h}{2},$$

$$\bar{\sigma}_r = c_{11}\bar{\epsilon}_r + c_{12}\bar{\epsilon}_\theta, \bar{\sigma}_\theta = c_{12}\bar{\epsilon}_r + c_{22}\bar{\epsilon}_\theta, \bar{E}_r = -\mu_{\bar{r}},$$

$$\bar{\epsilon}_r = y_{\bar{r}}, \bar{\epsilon}_\theta = \frac{y + y^{(-1)}}{2\bar{r}}, \varphi_1 = S^r S^t(rf_1), \varphi_2 = S^r(rf_2),$$

Here S^r and S^t are averaging Steklov operators defined by the formulae:

$$S^r u(r, t) = \begin{cases} \frac{2}{h} \int_{R_0}^{R_0 + \frac{h}{2}} u(\xi, t) d\xi, & r = R_0 \\ \frac{1}{h} \int_{r - \frac{h}{2}}^{r + \frac{h}{2}} u(\xi, t) d\xi, & R_0 < r < R_1 \\ \frac{2}{h} \int_{R_1 - \frac{h}{2}}^{R_1} u(\xi, t) d\xi, & r = R_1 \end{cases}$$

$$S^t v(r, t) = \begin{cases} \frac{1}{\tau} \int_{t - \frac{\tau}{2}}^{t + \frac{\tau}{2}} v(r, \mu) d\mu, & t > 0 \\ \frac{2}{\tau} \int_0^{\frac{\tau}{2}} v(r, \mu) d\mu, & t = 0. \end{cases}$$

For any function the notations like $p_{\bar{r}}$, p_r , $p_{\bar{r}r}$ stand for the first backward-difference, the first forward-difference and the second central difference approximation of the function p with respect to r respectively (analogous notations are used for differences with respect to t). The rest of notations are the following: $y(r, t)$ and $\mu(r, t)$ are the functions of two discrete variables defined on the difference mesh

$$\bar{\omega}_{h\tau} = \bar{\omega}_h \times \bar{\omega}_\tau$$

where

$$\bar{\omega}_h = \{r_i = R_0 + ih, h = \frac{R_1 - R_0}{N}, i = \overline{0, N}\}, \bar{\omega}_\tau = \{t_j = j\tau, \tau = T/L, j = \overline{0, L}\}.$$

These functions for each $t \in \bar{\omega}_\tau$ are elements of Hilbert spaces

$$H_1 = \{y(r) : r \in \bar{\omega}_h\}, \quad H_2^0 = \{\mu(r) : r \in \bar{\omega}_h; \mu = 0, r = R_0, R_1\}$$

with the scalar product $(y, v) = \sum_{\bar{\omega}_h} \hbar yv$, where $\hbar = \frac{h}{2}$ for $i = 0, N$ and $\hbar = h$ for $i = \overline{1, N-1}$. They give approximations to the functions of displacement $u(r, t)$ and electrostatic potential $\varphi(r, t)$ respectively.

Let us introduce further notations for norms and semi-norms of these discrete functions:

$$\begin{aligned}\|y(t)\|^2 &= (y(t), y(t)), \quad \|y(t)\|_A^2 = (Ay(t), y(t)), \quad \|y(t)\|_0^2 = \sum_{t=0}^T \tau \|y(t)\|^2, \\ \|y\|_{(1)}^2 &= \|y\|_{D_1 - \frac{\tau^2}{4} A_1}^2 - \tau^2 R_1 \left(\frac{e_{11}^2}{\epsilon_{11}} \|\bar{\epsilon}_r\|^2 + \frac{e_{12}^2}{\epsilon_{11}} \|\bar{\epsilon}_\theta\|^2 \right) + \left\| \sum_{t'=0}^t \tau y(t') \right\|_{A_1}^2, \\ \|\mu\|_{(2)}^2 &= \left\| \sum_{t'=0}^t \tau \mu(t') \right\|_{A_2}^2.\end{aligned}$$

Obtaining a priori estimation for difference solutions in negative norms of the right hand parts is complicated by the strong electromechanical coupling of the initial system (see (2.7)-(2.8)). At the same time for many practically important problems (see [11,12] for some examples on cylindrical acoustic vibrator modelling) we should have a plausible picture on the propagation of the mixed electro-elastic waves. This can be achieved by *dealing with coupling from the very beginning of the numerical analysis stage*.

Firstly, note that the condition of the semi-norm $\|\cdot\|_{(1)}$ existence is positiveness for any y the following quadratic form:

$$((D_1 - \frac{\tau^2}{4} A_1)y, y) - \tau^2 R_1 \left(\frac{e_{11}^2}{\epsilon_{11}} \|\bar{\epsilon}_r\|^2 + \frac{e_{12}^2}{\epsilon_{11}} \|\bar{\epsilon}_\theta\|^2 \right) > 0. \quad (3.4)$$

Taking into consideration easily proved inequalities:

$$\begin{aligned}\|\bar{\epsilon}_r\|^2 &= \sum_{\omega_h^+} h(y_{\bar{r}})^2 \leq \frac{4}{h^2} \sum_{\omega_h} \hbar y^2, \\ \|\bar{\epsilon}_\theta\|^2 &= \sum_{\omega_h^+} h \left(\frac{y + y^{(-1)}}{2\bar{r}} \right)^2 \leq \frac{1}{R_0^2} \sum_{\omega_h} \hbar y^2, \\ \sum_{\omega_h^+} h \bar{\epsilon}_r \bar{\epsilon}_\theta &= \frac{1}{h} \sum_{\omega_h^+} h \frac{y^2 - (y^{(-1)})^2}{2\bar{r}} \leq \frac{1}{2R_0 h} \sum_{\omega_h} \hbar y^2,\end{aligned}$$

where

$$\omega_h^+ = \{r_i = R_0 + ih, i = \overline{1, N}\}$$

as well as the equality:

$$(A_1 y, y) = r(\bar{\sigma}_r, \bar{\epsilon}_r) + r(\bar{\sigma}_\theta, \bar{\epsilon}_\theta) = r(c_{11} \|\bar{\epsilon}_r\|^2 + 2c_{12} (\bar{\epsilon}_r, \bar{\epsilon}_\theta) + c_{22} \|\bar{\epsilon}_\theta\|^2),$$

it is not difficult to conclude that the condition (3.4) will be satisfied if the following inequality is fulfilled:

$$\tau \leq \frac{h}{c} \left\{ \left(1 - \frac{\epsilon}{\rho} \right) / \left[1 + \frac{3\delta}{1 + \delta} + \frac{c_{12}}{4R_0 c_{11}(1 + \delta)} h + \frac{c_{22} + 4e_{12}^2/\epsilon_{11}}{4R_0^2 c_{11}(1 + \delta)} h^2 \right] \right\}^{\frac{1}{2}}, \quad (3.5)$$

where $c = [c_{11}(1 + \delta)/\rho]^{\frac{1}{2}}$ is the velocity of the mixed electro-elastic wave propagation and $\delta = \frac{\epsilon_{11}^2}{\epsilon_{11} c_{11}}$ is the coupling coefficient of electromechanical system, $\epsilon > 0$.

Let us take the scalar product of the equation (3.1) and $\tau w(t)$, where $w(t)$ is defined as in [14]:

$$w(t) = \sum_{t'=\tau}^{t_1} \tau(y(t') + \dot{y}(t')), \quad \dot{y}(t') = y(t' - \tau).$$

Let us also define the following function:

$$w_1(t) = \sum_{t'=\tau}^t \tau(y(t') + \dot{y}(t')).$$

The functions $w(t)$ and $w_1(t)$ have the following properties:

$$\begin{aligned} w_t &= -(y + \dot{y}) \text{ for all } 0 \leq t < t_1; \quad w(t) = 0, \text{ for all } t_1 \leq t \leq T; \\ w(t) &= w_1(t_1) - w_1(t), \text{ and } w(0) = w_1(t_1). \end{aligned}$$

Using the easily verified identity

$$\tau(D_1 y_t, w) = (D_1 y_t, w) - (D_1 y_t, \dot{w}) - \tau(D_1 y_t, w_t),$$

we get:

$$(D_1 y_t, w) - \tau(D_1 y_t, w_t) + \tau(A_1 y, w) + (C_1 \mu, w) = (D_1 y_t, \dot{w}) + \tau(\varphi_1, w).$$

Summing the last identity over t from τ to a certain t_1 ($0 < t_1 \leq T$) and taking into account that $w(t_1) = 0$ we derive the following **energy identity**:

$$\begin{aligned} \sum_{t'=\tau}^{t_1} \tau(D_1 y_t, y + \dot{y})(t') + \sum_{t'=\tau}^{t_1} \tau(A_1 y, w)(t') + \sum_{t'=\tau}^{t_1} \tau(C_1 \mu, w)(t') = \\ (D_1 y_t, w)(0) + \sum_{t'=\tau}^{t_1} \tau(\varphi_1, w)(t'). \end{aligned} \quad (3.6)$$

Let us introduce now the following auxiliary functions:

$$g(t) = \frac{1}{2}w(t) - \frac{\tau}{2}y(t), \quad j(t) = \frac{1}{2}v(t) - \frac{\tau}{2}\mu(t)$$

and the function $v(t)$ defined by the formula:

$$v(t) = \sum_{t'=\tau}^{t_1} \tau(\mu(t') + \dot{\mu}(t')),$$

which has the properties analogous to the properties of the function $w(t)$. Using obvious equalities:

$$y = \frac{1}{2}(y + \dot{y}) + \frac{\tau}{2}y_t = -g_t, \quad w = g + \dot{g},$$

the consequence of the equation (3.2):

$$C_2 w = -A_2 v + \sum_{t'=\tau}^{t_1} \tau(\varphi_2 + \check{\varphi}_2)$$

and properties of operators of the scheme (3.1)-(3.3), the second and the third terms of the left hand part of (3.6) can be transformed in the following way:

$$\begin{aligned} \sum_{t'=\tau}^{t_1} \tau(A_1 y, w)(t') &= - \sum_{t'=\tau}^{t_1} \tau(A_1 g_i, g + \check{g})(t') = -(A_1 g, g)(t_1) + (A_1 g, g)(0) = \\ &= -\frac{\tau^2}{4} (A_1 y, y)(t_1) + (A_1 g, g)(0), \end{aligned}$$

$$\begin{aligned} \sum_{t'=\tau}^{t_1} \tau(C_1 \mu, w) &= - \sum_{t'=\tau}^{t_1} \tau(\mu, C_2 w) = \sum_{t'=\tau}^{t_1} \tau(\mu, A_2 v) - \sum_{t'=\tau}^{t_1} \tau(\mu, \sum_{t''=t'+\tau}^{t_1} \tau(\varphi_2 + \check{\varphi}_2)) = \\ &= -\frac{\tau^2}{4} (A_2 \mu, \mu)(t_1) + (A_2 j, j)(0) - \sum_{t'=\tau}^{t_1} \tau(\mu, \sum_{t''=t'+\tau}^{t_1} \tau(\varphi + \check{\varphi}_2)). \end{aligned}$$

Then the identity (3.6) can be rewritten in the form:

$$\begin{aligned} ((D_1 - \frac{\tau^2}{4} A_1) y, y)(t_1) + (A_1 g, g)(0) + (A_2 j, j)(0) - \\ \frac{\tau^2}{4} (A_2 \mu, \mu)(t_1) &= (D_1 y, y)(0) + (D_1 y_t, w)(0) + \sum_{t'=\tau}^{t_1} \tau(\varphi_1, w)(t') + \\ &\quad \sum_{t'=\tau}^{t_1} \tau(\mu, \sum_{t''=t'+\tau}^{t_1} \tau(\varphi_2 + \check{\varphi}_2)) . \quad (3.7) \end{aligned}$$

The estimation of the term $-\frac{\tau^2}{4} (A_2 \mu, \mu)(t_1)$ can be performed if we take into consideration the equation (3.2):

$$-\frac{\tau^2}{4} (A_2 \mu, \mu) \geq 3R_1 \frac{\tau^2}{4} \left(\frac{1}{R_0^2 \epsilon_{11}} \|\hat{\lambda}\|^2 + \frac{e_{11}^2}{\epsilon_{11}} \|\bar{\epsilon}_r\|^2 + \frac{e_{12}^2}{\epsilon_{11}} \|\bar{\epsilon}_g\|^2 \right),$$

where $\hat{\lambda}(t) = \lambda(t + \tau)$, $\varphi_2 = \hat{\lambda}_r$ and $\hat{\lambda} = 0$ for $r = R_0$. From the obvious inequality $(a - b)^2 \geq \frac{1}{2}a^2 - b^2$ the following estimations are implied:

$$\|g(0)\|^2 \geq \frac{1}{8} \|w_1(t_1)\|^2 - \frac{\tau^2}{4} \|y(0)\|^2, \quad \|j(0)\|^2 \geq \frac{1}{8} \|v_1(t_1)\|^2 - \frac{\tau^2}{4} \|\mu(0)\|^2,$$

where the function $v_1(t)$ is defined analogously to the function $w_1(t)$ by the replacement y for μ .

Transforming the rest of terms in the right hand part of (3.7) with the help of the Cauchy-Buniakovskiy inequality, the ϵ -inequality [18], and assuming that $\varphi_1 =$

$(\xi_1)_t + (\xi)_r$, we come to the following result (after application of the difference analogue of Gronwal's lemma) which we derive from (3.7).

Theorem 3.1.

If the condition (3.5) is satisfied, then the following *a priori* estimation:

$$\begin{aligned} \|y(t_1)\|_{(1)}^2 + \|\mu(t_1)\|_{(2)}^2 &\leq M \left\{ ((D_1 + \frac{\tau^2}{4} A_1)y, y)(0) + \frac{\tau^2}{4} (A_2 \mu, \mu)(0) + \right. \\ &\quad \left. \|D_1 y_t(0)\|^2 + \sum_{t'=0}^T \tau (\|\xi_1\|^2 + \|\xi_2\|^2) + \|\hat{\lambda}\|^2 + \|\hat{\lambda}\|_0^2 \right\} \end{aligned} \quad (3.8)$$

is true for the solution of the problem (3.1)-(3.3)

Remark 3.1. In fact the condition (3.5) is the CFL-type stability condition in this case of coupled field theory. It contains the velocity of the coupled electro-elastic wave. If $\delta \rightarrow 0$ this stability condition coincides in the main term with the stability condition obtained in [11].

4. Difference Scheme Convergence in the Generalized Solution Classes

Let us apply *a priori* estimation (3.8) obtained for the investigation of the difference scheme (3.1)-(3.3) convergence. We consider here the generalized solutions from Sobolev's spaces $W_2^k(Q_T)$, $k = \overline{2, 4}$ and the space $V(\bar{Q}_T) = C(\bar{Q}_T) \cap Q_1(\bar{Q}_T)$ of continuous functions with piecewise first derivatives, which in the domain of continuity have integrable with square generalized derivatives.

To study the questions of problem (3.1)-(3.3) convergence, let us first consider an approximation error of the difference scheme:

$$z = y - u, \quad \zeta = \mu - \varphi,$$

which is the solution of the following operator-difference scheme:

$$\begin{cases} D_1 z_{tt} + A_1 z + C_1 \zeta = \psi, & t \in \omega_\tau \\ A_2 \zeta + C_2 z = \chi, & t \in \bar{\omega}_\tau \\ z = 0, \quad D_1 z_t = \psi, & t = 0. \end{cases}$$

We introduce notations:

$$u(r \pm \frac{h}{2}, t) = u^{(\pm 0.5)}, \quad u(r, t + \frac{\tau}{2}) = \bar{u}, \quad u(r, t - \frac{\tau}{2}) = \check{u}$$

and consider the error of approximation of equation (2.1) in the inner nodes of the mesh:

$$\psi = \varphi_1 - r \rho u_{tt} + c_{11}(\bar{r} u_r)_r - c_{22} \left[\frac{u^{(+1)} + u}{4\bar{r}^{(+1)}} + \frac{u + u^{(-1)}}{4\bar{r}} \right] +$$

$$c_{11}(\bar{r}\varphi_{\bar{r}})_r - e_{12} \frac{\varphi_r + \varphi_{\bar{r}}}{2}. \quad (4.1)$$

The sequence of the next operations is as follows:

- apply to equation (2.1) the composition of the averaging operators $S^r S^t$;
- express value of φ_1 from the obtained equality;
- substitute φ_1 into (4.1);
- use the main property of the averaging operators:

$$S^r \frac{\partial u}{\partial r} = \frac{1}{h} [u^{(+0.5)} - u^{(-0.5)}] = (u^{(-0.5)})_r, \quad S^t \frac{\partial u}{\partial t} = \frac{1}{\tau} [\bar{u} - \check{u}] = (\check{u})_t,$$

to get:

$$\psi = \rho(\eta_{11})_t + c_{12}(\eta_{12})_r + c_{22}(\eta_{13})_r + e_{11}(\eta_{14})_r + e_{12}(\eta_{15})_r,$$

where

$$\begin{aligned} \eta_{11} &= S^r \left(r \frac{\partial \check{u}}{\partial t} \right) - r u_t, \quad \eta_{12} = \bar{r} u_{\bar{r}} - S^t \left(\left(r \frac{\partial u}{\partial r} \right)^{(-0.5)} \right), \\ (\eta_{13})_r &= \psi_{13} = S^r S^t \left(\frac{u}{r} \right) - \left(\frac{u^{(+1)} + u}{4\bar{r}^{(+1)}} + \frac{u + u^{(-1)}}{4\bar{r}} \right), \\ (\eta_{13})_i &= \sum_{j=0}^{i-1} h(\psi_{13})_j, \quad (\eta_{13})_0 = 0, \\ \eta_{14} &= \bar{r} \varphi_{\bar{r}} - S^t \left(\left(r \frac{\partial \varphi}{\partial r} \right)^{(-0.5)} \right), \quad \eta_{15} = S^t((\varphi)^{(-0.5)}) - \frac{\varphi + \varphi^{(-1)}}{2}. \end{aligned}$$

Now let us consider the case when the solution of the problem (2.1)-(2.6) belongs to the space $W_2^2(Q_T)$. We shall estimate the functionals η_{1k} , $k = \overline{1, 5}$, using Bramble-Hilbert lemma [2;19;20,p.146]. First let us consider the functional η_{12} . It is easy to see that it is the linear functional bounded in the space $W_2^2(Q_T)$. Moreover,

$$|\eta_{12}| \leq Mh^{-1} \|u\|_{W_2^2(e_1)}.$$

The linear substitution $\xi_1 = r + s_1 h$, $\xi_2 = t + s_2 \tau$ permits us to transgress from the region $e_1 = \{(r', t') : r - h < r' < r, t - \frac{\tau}{2} < t' < t + \frac{\tau}{2}\}$ to the region $E = \{(s_1, s_2) : -1 < s_1 < 0, -\frac{1}{2} < s_2 < \frac{1}{2}\}$. It is well-known that a linear substitution does not change the class of functions, and therefore,

$$|\eta_{12}| \leq Mh^{-1} \|u\|_{W_2^2(E)}.$$

Further one can verify that the functional

$$\eta_{12} = -\frac{1}{2h} \{ \tilde{u}(0, 0) - \tilde{u}(-1, 0) - \int_{-0.5}^{0.5} \frac{\partial \tilde{u}(-\frac{1}{2}, s_2)}{\partial s_1} ds_2 \}, \text{ where } \tilde{u}(s) = u(r(\xi_1), t(\xi_2))$$

turns into zero for all polynomials up to the first degree inclusively. That is why according to the Bramble-Hilbert lemma we have:

$$|\eta_{12}| \leq Mh^{-1} |\tilde{u}|_{W_2^2(E)},$$

Transgressing to the variables (r, t) we get:

$$|\eta_{12}| \leq M \frac{h^2 + \tau^2}{h} (h\tau)^{-\frac{1}{2}} |u|_{W_2^2(e_1)}.$$

In the same way it is not difficult to show that

$$|\eta_{12}| \leq M \frac{(h^2 + \tau^2)^{\frac{k}{2}}}{h} (h\tau)^{-\frac{1}{2}} |u|_{W_2^k(e_1)}, \text{ where } k = 3, 4. \quad (4.2)$$

Analogously, using Bramble-Hilbert lemma technique one can obtain estimations for other functionals.

Let us consider now the error of approximation on the boundary, for example if $r = R_0$:

$$\psi|_{R_0} = \varphi_1 - \rho r u_{\bar{t}} + \frac{2}{h} \bar{r}^{(+1)} \check{\sigma}_r^{(+1)} - \check{\sigma}_{\theta}^{(+1)},$$

where

$$\check{\sigma}_r = c_{11} u_r + c_{12} \frac{u + u^{(-1)}}{2\bar{r}} + e_{11} \varphi_{\bar{r}}, \quad \check{\sigma}_{\theta} = c_{12} u_{\bar{r}} + c_{22} \frac{u + u^{(-1)}}{2\bar{r}} + e_{12} \varphi_{\bar{r}}.$$

It can be shown that

$$\psi|_{R_0} = \rho(\eta'_{11})_t + \frac{2}{h} (\eta'_{12})_r + (\eta'_{13})_r,$$

where

$$\begin{aligned} \eta'_{11} &= \frac{2}{h} \int_{R_0}^{R_0 + \frac{h}{2}} r \frac{\partial \tilde{u}}{\partial t} dr - r u_{\bar{t}}, \quad (\eta'_{12})_r = \psi'_{12} = \bar{r}^{(+1)} \check{\sigma}_r^{(+1)} - \frac{1}{\tau} \int_{t - \frac{\tau}{2}}^{t + \frac{\tau}{2}} r \sigma_r|_{R_0 + \frac{h}{2}} dt, \\ (\eta'_{13})_r &= \psi'_{13} = \frac{2}{h\tau} \int_{t - \frac{\tau}{2}}^{t + \frac{\tau}{2}} \int_{R_0}^{R_0 + \frac{h}{2}} \sigma_{\theta} dr dt - \check{\sigma}_{\theta}^{(+1)}. \end{aligned}$$

We consider for the sake of brevity only the functional ψ'_{13} . If we represent the functional in the form:

$$\psi'_{13} = \frac{1}{\tau} \int_{t - \frac{\tau}{2}}^{t + \frac{\tau}{2}} \left[\frac{2}{h} \int_{R_0}^{R_0 + \frac{h}{2}} \sigma_{\theta} dr - \sigma_{\theta}(R_0 + \frac{h}{2}, t) \right] dt + \left[\frac{1}{\tau} \int_{t - \frac{\tau}{2}}^{t + \frac{\tau}{2}} \sigma_{\theta}(R_0 + \frac{h}{2}, t) dt - \check{\sigma}_{\theta}^{(+1)} \right]$$

(where in the brackets of the first term we have the error of right rectangular quadrature formula), then its estimation can be performed with the help of Bramble-Hilbert lemma:

$$|\psi'_{13}| \leq M(h^2 + \tau^2)^{\frac{p}{2}} (h\tau)^{-\frac{1}{2}} |\sigma_{\theta}|_{W_2^p(e'_1)}, \quad p = \overline{1, 3}, \quad (4.3)$$

where

$$e'_1 = \{(r', t') : R_0 < r' < R_0 + \frac{h}{2}, t - \frac{\tau}{2} < t' < t + \frac{\tau}{2}\}.$$

At last we should take into consideration the following:

$$(\eta'_{13})_i = \sum_{j=0}^{i-1} h(\psi'_{13})_j, \quad (\eta'_{13})_0 = 0, \quad \text{and } \sigma_{\theta} = c_{12} \frac{\partial u}{\partial r} + c_{22} \frac{u}{r} + e_{12} \frac{\partial \varphi}{\partial r}.$$

The approximation error of initial conditions has the form:

$$\begin{aligned}\psi|_{t=0} = & \rho r u_1 - S^r(r\rho \frac{\partial u}{\partial t}(0)) + [S^r(r\rho \frac{\partial u}{\partial t}(\frac{r}{2})) - r\rho u_t] + \\ & \frac{\tau}{2}[S^t S^r(\frac{\partial}{\partial r}(r\sigma_r) - \sigma_\theta) - A_1 u - C_1 \varphi].\end{aligned}\quad (4.4)$$

It has been obtained taking into consideration the equation (2.1), on which we preliminary acted by the composition of operators $S^r S^t$ (where S^t is defined for $t = 0$).

The approximation error for the equation (3.2) is readily obtained if we act on the equation (2.2) by the averaging operator S^r :

$$\begin{aligned}\chi = & -\epsilon_{11}[\frac{1}{h} \int_{r-\frac{h}{2}}^{r+\frac{h}{2}} \frac{\partial}{\partial r}(r \frac{\partial \varphi}{\partial r}) dr - (\bar{r} \varphi_{\bar{r}})_r] + e_{11}[\frac{1}{h} \int_{r-\frac{h}{2}}^{r+\frac{h}{2}} \frac{\partial}{\partial r}(r \frac{\partial u}{\partial r}) dr - \\ & (\bar{r} u_{\bar{r}})_r] + e_{12}[\frac{1}{h} \int_{r-\frac{h}{2}}^{r+\frac{h}{2}} \frac{\partial u}{\partial r} dr - (\frac{u + u^{(-1)}}{2})_r].\end{aligned}\quad (4.5)$$

The estimations of the right hand parts (4.4), (4.5) do not cause any difficulties. They are obtained by the above technique.

To obtain accuracy estimation of difference scheme (3.1)-(3.3) in cases where the required solution belongs to Sobolev's spaces $W_2^k(Q_T)$, $k = \overline{2, 4}$, we should take into consideration a *priori* estimation obtained in the theorem 3.1, which for the scheme error has the form:

$$\|z\|_{(1)}^2 + \|\zeta\|_{(2)}^2 \leq M \{ \|\psi|_{t=0}\|^2 + \sum_{t'=0}^T \tau (\|\bar{\xi}_1\|^2 + \|\bar{\xi}_2\|^2) + \|\bar{\lambda}\|^2 + \|\bar{\lambda}\|_0^2 \}, \quad (4.6)$$

where

$$\psi = (\bar{\xi}_1)_t + (\bar{\xi}_2)_r, \quad \chi = (\bar{\lambda})_r.$$

The right hand part (4.6) is estimated with the help of inequalities like (4.2), (4.3). For example, using the estimations of the functionals η_{11} , η'_{11} we can get:

$$\begin{aligned}(\sum_{t'=0}^T \tau \|\bar{\xi}_1\|^2)^{\frac{1}{2}} &= (\sum_{t'=0}^T \tau \sum_{\omega_h} h \rho^2 |\eta_{11}|^2)^{\frac{1}{2}} \leq \\ M(\sum_{t'=0}^T \sum_{\omega_h} \tau h \frac{(h^2 + \tau^2)^k}{\tau^2} (h\tau)^{-1} |u|_{W_2^k(e_m)}^2)^{\frac{1}{2}} &\leq M \frac{(h^2 + \tau^2)^{\frac{k}{2}}}{\tau} |u|_{W_2^k(Q_T)}, \quad k = \overline{2, 4},\end{aligned}$$

where

$$\begin{aligned}e_m &= e_2 = \{(r', t') : r - \frac{h}{2} < r' < r + \frac{h}{2}, t - \tau < t' < t\} && \text{for } r \in \omega_h, \\ e'_m &= e'_1 = \{(r', t') : R_1 - \frac{h}{2} < r' < R_1, t - \frac{\tau}{2} < t' < t + \frac{\tau}{2}\} && \text{for } r = R_0, \\ e_m &= e'_2 = \{(r', t') : R_1 - \frac{h}{2} < r' < R_1, t - \frac{\tau}{2} < t' < t + \frac{\tau}{2}\} && \text{for } r = R_1.\end{aligned}$$

In the same way the norms of the rest functionals of the right hand part of (4.6) are estimated. The only exception is the estimation of the functional ψ'_{13} . It does not

become zero on polynomials of the first degree if the required solution belongs to the space $W_2^3(Q_T)$. However it can be estimated by the Ilyin inequality that gives an integral estimation on the near-boundary strip of the region (see chapter 1 in [19]).

Some applied problems of coupled theory of dynamical electroelasticity are characterized by the fact that the solution derivatives have a discontinuity of the first kind. Thus we shall consider the case when the solution of the problem (2.1)-(2.6) belongs to the space $V(\bar{Q}_T)$.

A priori estimation (4.6) for the error of the scheme remains true in this case as well, and

$$\sum_{t'=0}^T \tau \|\bar{\xi}_i\|^2 = \sum_{t'=0}^T \tau \sum_{r' \in \omega_p} h \bar{\xi}_i^2(r', t') + \sum_{t'=0}^T \tau \sum_{r' \in \bar{\omega}_h / \omega_p} h \bar{\xi}_i^2(r', t'), \quad i = 1, 2,$$

where ω_p are the points of the mesh, the neighborhood of which contains points of discontinuity of the first derivatives of the solution.

In the domain of continuity of the first derivatives, corresponding functionals which occur in the approximation error of ψ and χ have been estimated earlier. In the points of their discontinuity the functional are bounded. So far, as the total number of points ω_p (where the first derivatives have discontinuity) is finite, then

$$\sum_{t'=0}^T \tau \sum_{r' \in \omega_p} h \bar{\xi}_i^2(r', t') = O(\tau + h).$$

As the result we have proved the following

Theorem 4.2.

Under the stability condition (3.5) the solution of the difference scheme (3.1)-(3.3) converges to the generalized solution of the coupled dynamical problem of electroelasticity at the rate of $O(h^k + \tau^k)$. The following accuracy estimation:

$$\|z\|_{(1)} + \|\zeta\|_{(2)} \leq M(h^k + \tau^k), \quad (4.7)$$

where $k = \frac{1}{2}$ if the solution of the problem (2.1)-(2.6) belongs to the space $V(Q_T)$ and $k = \frac{p}{2}$ if the solution is from the class $W_2^p(Q_T)$, $p = \overline{2, 4}$ is true.

Remark 4.1. When the equations (2.1) and (2.2) are connected only by the state equations (2.5), but there is no connection through the boundary conditions for tensions, for example, when displacements are given on the boundary, the accuracy estimation (4.7) can be improved. Using the technique presented in [3] the difference scheme convergence of the second order (to the generalized solution from $W_2^2(Q_T)$) can be proved in a weaker then $L_2(\omega)$ metric for this semi-coupled case.

Remark 4.2. Similar results have also been obtained in the coupled theory of thermoelasticity [9] where mixed parabolic and hyperbolic operators are nonseparable globally.

5. Conclusions and Future Directions.

Mathematical modelling in coupled field theory requires approaches which can be applied even if a solution of the problem does not possess an excessive smoothness often imposed as an *a priori* assumption. In this paper we have developed such an approach with respect to the problem arising from nonstationary electroelasticity. We explicitly derived the stability condition for the operator-difference scheme applied to the numerical solution of the problem. We also proved convergence of such solution to a generalized solution of the original problem. Depending on smoothness of the latter we obtained a scale of accuracy estimations. Such scales give important *a-priori* characteristics of *computational efficiency* of underlying numerical procedures. The approach based on coupling procedures is a natural way to reflect additional information about the system in mathematical models. In describing system dynamics or transient periods of system behaviour it is very important to use such information rather than ignore it. Since our approach in mathematical modelling and computational experiment uses the implicit assumption that the problem can effectively be reduced to a finite set of equations, inequalities or inclusions, it is very important to investigate models of *varying complexity*. On the other hand, fixing the degree of coupling in mathematical models implies thorough investigation of system stability under the specified coupling level.

The continuing efforts to improve outcomes of mathematical modelling by improving associated computational models and numerical algorithms as well as methods of data collection and analysis, computer software and hardware are important parts of scientific progress. The resulting mathematical models and available computer resources epitomize the grand challenge in human endeavours of modelling dynamics. Nevertheless, it should be realised that the refinements of the coupling approach may continue indefinitely. As a result we need to maintain a balance between *computational efficiency* and the *complexity of such coupling* in the mathematical and computational models we use. At a certain level of refinement of coupling process we cannot carry out controlled experiments on any real dynamical system because the system may only be observable in its transient states. Therefore, available⁴ informational datasets can never completely validate the simulation no matter how seemingly simple dynamics is. However, we always can construct models of "proxy system" on which controlled experiments can be conducted. Such mathematical models of "proxy system" reflect our attempts to describe real processes, dynamical systems and phenomena as transformation of different types of energy. As a result, variational approaches in construction of computational procedures for such mathematical models appear to be natural being, in fact, well established. One, however, should address the difficulty of such approach arising from the approximate character of conservation laws [6]. Even if we accept the optimistically exaggerated assumption that predictability of the "proxy system" is equal to the predictability of the system itself, we should face the fact that the accuracy of the approximation of the upper bound

⁴at a certain level of refinement of coupling process

on predictability decreases when the model improves under higher level of coupling [4,5,17].

We would like to mention two directions for possible development of the presented in this paper technique.

- A similar approach can be applied to more general mathematical models of coupled field theory. A straightforward generalisation might be obtained for dynamical problems of thermoelectroelasticity. A generalization of the presented technique can be also applied to non-local models arising from both micro- and macro- levels of description of real dynamical systems. Applications in semiconductor device theory and climate modelling will be published elsewhere.

- We attempted to approach some problems in nonsmooth (including stochastic) optimal control theory on the basis of Steklov's operators idea. In particular, we derived a local optimality principle which allowed us to reformulate the original problems in such a way that the application of some ideas presented here seems to be encouraging. Some results in this field are published in [6,7].

Acknowledgement

I wish to acknowledge the support of the School of Mathematics at the University of South Australia as well as to thank Paul S. Gaertner for helpful assistance at the final stage of preparation of this paper.

REFERENCES

- [1] Berlincourt,D.A., Curran,D.R., and Jaffe, H. "Piezoelectric and Piezomagnetic Materials and Their Function in Transducers" in *Physical Acoustics*, Ed.W.P.Mason, Vol.1A, New York and London: Academic Press, 1964, 204-236.
- [2] Bramble, J.H. and Hilbert, S.R. "Estimation of Linear Functionals on Sobolev Spaces with Application to Fourier Transforms and Spline Interpolation". *SIAM J.Num.Anal.* 7, 1970, 113-124.
- [3] Djuraev, I.N., and Moskalkov, M.N. "Convergence of Difference Scheme Solution to the Generalized Solution of the Wave Equation from $W_2^2(Q_T)$ ". *Differential Equations*, 21, 1985, 2145-2152.
- [4] Lorenz, E.N. "Atmospheric predictability experiments with a large numerical model". *Tellus*, 34, 505-513.
- [5] May, R. "Necessity and Change: Deterministic Chaos in Ecology and Evolution." *Bull. of the Americ.Math.Soc.*, Vol.32, No.3, 291-308.
- [6] Melnik, V.N. "Nonconservation Law Equation in Mathematical Modelling: Aspects of Approximation". *Proc. of the Int. Conf. AEMC'96*, 423-430.
- [7] Melnik, V.N. "Generalised Solutions of Hamilton-Jacobi-Bellman Equation from Sobolev's Classes". *ANZIAM Conference, Masterton, New Zealand*, 1996.
- [8] Melnik, V.N. "Existence and Uniqueness Theorems of the Generalized Solution for a Class of Non-Stationary Problems of Coupled Electroelasticity". *Izvestiya VUZ.Matematika*, 35, 1991, 24-32(original); or by Allerton Press: Soviet Mathematics, 1991, 23-30.
- [9] Melnik, V.N. "Accuracy of Difference Schemes in Coupled Nonstationary Thermoelasticity for Problems in Tension Formulation". *Mathematical Methods and Physico-Mechanical Fields; ISSN 0130-9420*, 34, 1991, 95-99.
- [10] Melnik, V.N. "Semi-Implicit Difference Schemes with FCT for Quasihydrodynamic Model of Semiconductor Devices". *Electronic Modelling*, No.5, 111-117.; or by Gordon and Breach: *Engineering Simulation*, 12, 1995, 856-865.
- [11] Melnik, V.N. and Moskalkov, M.N. "On the Coupled Non-Stationary Electroelastic Oscillations of a Piezoceramic Cylinder with Radial Polarization". *Computational Mathematics and Mathematical Physics*, 28, 1988, 1755-1756(original); or by Pergamon Press: June(1990), 109-110.
- [12] Melnik, V.N. and Moskalkov, M.N. "Difference Schemes for and Analysis of Approximate Solutions of Two-Dimensional Nonstationary Problems in Coupled Electroelasticity". *Differential Equations*, 27, 1991, 1220-1229(original); or by C/B, New York: (1992), 860-867.

- [13] Mindlin, R. "Equations of High Frequency Vibrations of Thermopiezoelectric Crystal Plates". *Int.J.Solids and Structure*, 10, 1974, 625-637.
- [14] Moskalkov, M.N. "On Accuracy of Difference Schemes for Approximation of Wave Equation with Piecewise Smooth Solutions". *Computational Mathematics and Mathematical Physics*, 14, 1974, 390-401.
- [15] Nowacki, W. (1975), "Dynamic Problems of Thermoelasticity". *Noordhoff International Publishing,Leyden, The Netherland and PWN*.
- [16] Nowacki, W. (1986), "Electromagnetic Effects in Solids", *Moscow, Mir Publishers*.
- [17] Reynolds, C.A., Webster, P.J. and Kalnay, E. (1994) "Random Error growth in NMC's Global Forecasts", *MWR*, 122, 1281-1305.
- [18] Samarskiy, A. (1989), "Theory of Difference Schemes". *Moscow, Nauka*.
- [19] Samarskiy, A., Lazarov, R. and Makarov, V. (1987), "Difference Schemes for Differential Equations with Generalized Solutions". *Moscow, VS Publishers*.
- [20] Strang, G. and Fix, G. (1973), "An Analysis of the Finite Element Method". *Prentice-Hall*.

