KFKI-1983-21

COLLECTION OF SCIENTIFIC PAPERS

IN COLLABORATION WITH JOINT INSTITUTE FOR NUCLEAR RESEARCH, DUBNA, USSR AND CENTRAL RESEARCH INSTITUTE FOR PHYSICS, BUDAPEST, HUNGARY

ALGORITHMS AND PROGRAMS FOR SOLUTION OF SOME PROBLEMS IN PHYSICS

Hungarian Academy of Sciences

CENTRAL RESEARCH INSTITUTE FOR PHYSICS

BUDAPEST

KFKI-1983-21

СОВМЕСТНЫЙ НАУЧНЫЙ СБОРНИК ОБЪЕДИНЕННОГО ИНСТИТУТА ЯДЕРНЫХ ИССЛЕДОВАНИЙ (ДУБНА, СССР) И ЦЕНТРАЛЬНОГО ИНСТИТУТА ФИЗИЧЕСКИХ ИССЛЕДОВАНИЙ (БУДАПЕШТ, ВЕНГРИЯ)

АЛГОРИТМЫ И ПРОГРАММЫ ДЛЯ РЕШЕНИЯ НЕКОТОРЫХ ЗАДАЧ ФИЗИКИ

ВЫПУСК ЧЕТВЕРТЫЙ

Ответственный за выпуск: *Е.П. Жидков*

Редактор: *Г. Немет*х

COLLECTION OF SCIENTIFIC PAPERS IN COLLABORATION WITH JOINT INSTITUTE FOR NUCLEAR RESEARCH, DUBNA, USSR AND CENTRAL RESEARCH INSTITUTE FOR PHYSICS, BUDAPEST, HUNGARY

ALGORITHMS AND PROGRAMS FOR SOLUTION OF SOME PROBLEMS IN PHYSICS

FOURTH VOLUME

Responsible persons in edition: *E.P. Zhidkov*
G. Németh

HU ISSN 0368 5330
ISBN 963 372 052 4

С О Д Е Р Ж А Н И Е

CONTENT

МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ ПРИ РЕШЕНИИ ЗАДАЧ ОБРАБОТКИ И ИНТЕРПРЕТАЦИИ ЭКСПЕРИМЕНТАЛЬНЫХ ФИЗИЧЕСКИХ ДАННЫХ

А.Н. ТИХОНОВ, В.Я. ГАЛКИН

Московский государственный университет им. М.В. Ломоносова,
Москва

АННОТАЦИЯ

Приводятся разработанные за последнее десятилетие статистические модели из практики обработки экспериментальных данных, главным образом по гамма-взаимодействиям. Анализируются математические модели автоматического режима измерений и учета аппаратурных факторов, первичной статистической обработки, разделения множественности ядерных процессов, мессбауэровской спектроскопии и кинетики гамма-излучающих систем. Обсуждаются постановки типичных прямых и обратных задач, здесь возникающих, а также некоторые результаты их решения.

ABSTRACT

Statistical models for experimental physics data processing, mostly for gamma-interactions, are presented. Mathematical models of automatic measurements, hardware influence, multiplicity nuclear processes separation, Mössbauer spectroscopy and gamma-radiation kinetics are studied. Some direct and inverse problems formulations and decision results are discussed.

Введение

Проблемы автоматизированной обработки результатов эксперимента в настоящее время становятся все более актуальными. Тщательное, скурпулезное выполнение эксперимента, конечно, является главным условием успеха исследования ; однако вовсе не безразлично, как обработать и проанализировать полученные эмпирические данные. Надежность и объективная ценность выводов опытного исследования определяется как качеством и количеством, так и последующей обработкой экспериментального материала, на базе которого эти выводы делаются. Любая обработка, предпринимаемая ради решения практических задач, предполагает указание оценок, делающих сопоставления, прогнозы или выводы оправданными в той мере, в какой это вообще возможно в данной ситуации. Такой заключительный и наиболее ответственный момент приложения математических методов технически должен быть обеспечен надлежащими алгоритмами и программами для ЭВМ.

За последние годы в связи с бурным развитием техники и методики проведения эксперимента, особенно в ядерной физике, существенно возросли потоки информации, подвергающиеся обработке. Успехи в области автоматизации эксперимента позволяют получать большое количество первичных материалов, однако их последующая обработка и интерпретация (решение определенных обратных задач с целью сопоставимости с теоретическими посылками) весьма сложны и остаются проблематичным местом. Любой эксперимент, пусть даже самый тонкий и технически сложный, мало провести, из него необходимо еще извлечь обоснованные выводы. А это невозможно сделать без четкой математической постановки задач обработки и интерпретации и без соответствующих методов их решения. Далее, эксперимент, в котором играют роль столь различные обстоятельства и в столь большом числе, что их все учесть невозможно, необходимо предварительно спланировать, дать указания по его организации и на каждом этапе корректировать с учетом уже полученных результатов.

В процессе обработки экспериментальных данных можно выделить несколько основных этапов независимо от вида и характера конкретного эксперимента [I-3]. Важнейшими из них, реализуемыми в автоматизированных системах полной математической обработки, являются первичная статистическая обработка и интерпретация. Математические и практичес-

кие вопросы решения обратных задач интерпретации являются основной тематикой школы первого из авторов настоящей статьи, и им посвящена обширная литература [4]. Вместе с тем проблематика интерпретации - важнейшее но не единственное направление, стимулирующее развитие автоматизированной обработки экспериментальных данных. С методами статистической обработки и решением соответствующих прямых задач тесно связаны вопросы планирования, методики измерений, оценки разрешающей способности, наконец, организации, проектирования и управления экспериментом. Ответы на эти вопросы следуют исключительно из анализа математических моделей, представляющих типичные физические ситуации. Так что одним из основных направлений теории и практики обработки экспериментальных данных является построение математических моделей, постановка на их основе соответствующих задач обработки и разработка методов их численного решения.

Понятно, что изучение моделей, сколь-нибудь полно отражающих реальную физическую ситуацию, приводит к постановке весьма сложных математических задач, решение которых невозможно без применения ЭВМ. В этом смысле вычислительная машина становится неотъемлемой частью экспериментальной установки. При этом важно, чтобы разрабатывались наиболее универсальные и эффективные методы, решающие широкий класс задач обработки экспериментальных данных. Именно описанию некоторых математических моделей, возникающих в задачах обработки экспериментальной информации в ядерной физике, и изложению методов практической обработки данных посвящена статья. Она носит обзорный характер и написана на основе результатов группы сотрудников факультета вычислительной математики и кибернетики и Научно-исследовательского вычислительного центра МГУ.

Первым этапом обработки, как уже отмечалось, является первичная статистическая обработка наблюденных данных с целью получения выходных характеристик эксперимента и оценки их точности. Решение этой важной проблемы, связанной с необходимостью переработки больших объемов числовой информации, основывается на общих принципах математической статистики. После того как эксперимент поставлен и его выходные результаты введены в ЭВМ и расшифрованы, возникают задачи, во-первых, анализа информации на качественность, во-вторых, ее редукции для получения окончательных характеристик выхода, наконец, оценивания того элемента неопределенности, от которого всегда несвободен эксперимент при фиксированном (умеренном или малом) числе опытов. Для больших групп физических экспериментов этот этап обработки имеет много общего, полностью стандартизовать его однако не удается ; и не всегда общепринятые методы приводят к наилучшему решению, по-

скольку положенные в их основу предположения часто выполняются недостаточно точно. Конкретность задачи всегда вносит и свою специфику.

Ниже приводятся разработанные за последнее десятилетие статистические модели из практики обработки экспериментальных данных, главным образом по гамма-взаимодействиям. Постановки задач обработки, интерпретации и проектирования при этом основываются на концепции математической модели и процесса моделирования [3,5,6]. Анализируются математические модели автоматического режима измерений и учета аппаратурных факторов, первичной статистической обработки, разделения множественности ядерных процессов, мессбауэровской спектроскопии и кинетики гамма-лазера. Обсуждаются постановки типичных прямых и обратных задач, здесь возникающих, а также некоторые результаты их решения.

Модель автоматического режима измерений

В экспериментальной практике часто изучается зависимость интенсивности $Y(T)$ некоторого потока случайных событий (например, продуктов ядерных реакций) от некоторого параметра T (скажем, энергии частиц, вызывающих реакции). Измерения $Y_j(T_i)=Y_{ji}$, проводимые в автоматическом режиме на многоканальной аппаратуре, получаются суммированием числа зарегистрированных событий при многократном пробегании n точек сетки $\langle T_i \rangle$, $i=\overline{1,n}$ аргумента. Результат измерений - набор случайных векторов $\vec{Y}_j$ -, вообще говоря, с зависимыми компонентами, что обусловлено согласованным изменением аппаратурных параметров (дрейфом) в близких узлах сетки. При этом точность измерений характеризуется ковариационной матрицей Q, которую обычно оценивают выборочной ковариационной матрицей A. Число оцениваемых (неизвестных) параметров $n(n+1)/2$ в современных экспериментах часто имеет порядок числа измерений N ($j=\overline{1,N}$), так что точность получаемых оценок низка. Основной смысл модели автоматического режима измерений [7,8] заключается в таком способе параметризации ковариационной матрицы, который учитывает специфику указанного класса измеряемых величин: медленность дрейфа по сравнению со скоростью пробегания сетки $\langle T_i \rangle$. При этом число оцениваемых параметров будет иметь уже порядок n.

Предполагается, что $M\vec{Y}=\hat{\vec{Y}}=(\hat{Y}_1,\ldots,\hat{Y}_n)$ известно и $\vec{Y}\sim N(\hat{\vec{Y}},Q)$. Величина $\vec{Y}$ определена на вероятностном пространстве $\Omega=\Omega_1 \times \Omega_2$, где Ω_1 - "физическое" пространство, связанное со стохастичностью самого изучаемого процесса, а Ω_2 - "аппаратурное" пространство, описывающее случайный характер дрейфа параметров аппаратуры. (Каждой реализации измерений соответствует свой элементарный исход $\omega \in \Omega_2$, поскольку

за время измерений состояние аппаратуры определенным образом меняется, при этом ω есть функция времени). Далее, считается, что условное математическое ожидание $M_\omega\vec{Y}=M(\vec{Y}/\omega)$ имеет независимые компоненты, что объясняет коррелированность наблюдений только эффектом согласованного изменения состояния параметров аппаратуры ; сами измеряемые физические величины при различных T_i статистически независимы. Тогда для элементов Q имеем

$$q_{ij}=\begin{cases}\operatorname{cov}\{M_\omega Y_i, M_\omega Y_j\}, & i\neq j,\\ \overline{D_\omega Y_i}+DM_\omega Y_i, & i=j,\end{cases}$$

где операторы cov и D понимаются как осуществленные в Ω_2 и

$$M\{Y_iY_j\}=\int_{\Omega_2}M_\omega\{Y_iY_j\}d\omega\doteq\overline{M_\omega\{Y_iY_j\}}.$$

Условное математическое ожидание можно записать в виде

$$M_\omega Y_i=x_i\hat{Y}_i, \qquad (I)$$

выделив в явном виде безусловное математическое ожидание, так что каждой реализации ω отвечает n-мерная случайная величина $\vec{x}=(x_1,\ldots,x_n)$, определенная на Ω_2. Будем считать ее также подчиняющейся нормальному распределению. При этом

$$q_{ij}=\begin{cases}\hat{Y}_i\hat{Y}_j\operatorname{cov}\{x_ix_j\}, & i\neq j,\\ \overline{D_\omega Y_i}+\hat{Y}_i^2Dx_i, & i=j.\end{cases}$$

Теперь $\vec{Y}$ представим в виде суммы случайных величин

$$\vec{Y}=\hat{Y}+\Lambda\vec{\xi}+\vec{\varepsilon},$$

где Λ - диагональная матрица с элементами $\{\hat{Y}_1,\ldots,\hat{Y}_n\}$, $\vec{\xi}\sim N(0,\|\operatorname{cov}\{x_i,x_j\}\|)$, $\vec{\varepsilon}\sim N(0,V)$, V - диагональная матрица с элементами $\{\overline{D_\omega Y_1},\ldots,\overline{D_\omega Y_n}\}$; величины $\vec{\xi}$ и $\vec{\varepsilon}$ независимы.

Наконец, считается, что сетка $\langle T_i\rangle$ пробегается быстро по отношению к скорости изменения режима аппаратуры, так что $\vec{\xi}$ представима небольшим (по сравнению с n) числом членов разложения по системе ортогональных функций $\{k_s(T)\}$: $\xi_i=\sum_1^m k_s(T_i)u_s$, $i=\overline{1,n}$ или в матричной форме

$$\vec{\xi}=K\vec{u},\quad (K=\|k_s(T_i)\|,\ \vec{u}=(u_1,\ldots,u_m)').$$

Считая, что ранг K равен m, мы можем перейти к m-мерной случайной величине $\vec{\eta}=H^{-1}\vec{u}$, являющейся нормальной: $\vec{\eta}\sim N(0,I)$. Таким образом, для $\vec{Y}$ справедливо представление в виде суммы независимых случайных величин

$$\vec{Y}=\hat{Y}+\Lambda KH\vec{\eta}+\vec{\varepsilon}=\hat{Y}+L\vec{\eta}+\vec{\varepsilon}, \qquad (2)$$

аналогичное основной схеме факторного анализа [39]. Отметим однако принципиальную разницу нашей модели (2) по сравнению с факторным анализом. Как известно, в задаче факторного анализа существует неоднозначность в определении факторов f и их нагрузок (элементов матрицы L): они определены с точностью до вращения (ортогонального преобразования вектора простых факторов). В нашей же модели интерес представляют не сами факторы и их нагрузки, а инвариант ортогональных преобразований матрицы

$$Q = \Lambda K H H' K' \Lambda + V = LL' + V, \qquad (3)$$

а для ее оценивания нужно определить m^2+n параметров - элементов матриц H и V.

Метод максимального правдоподобия приводит к следующей системе непосредственно относительно H и V

$$\mathcal{J} H' = H' R,$$
$$\operatorname{diag}\{V - A - LL' + 2LL'Q^{-1}A\} = 0, \qquad (4)$$

где $\mathcal{J} = L'V^{-1}L$ - диагональная матрица, а

$$R = K'\Lambda V^{-1}(A-V)V^{-1}\Lambda K(K'\Lambda V^{-1}\Lambda K)^{-1}.$$

Полученная нелинейная система может быть решена различными итерационными методами. В важном частном случае квазипуассоновского потока событий: $V = \mu\Lambda$, $\mu = \mathrm{Const}$ первая группа уравнений (4) попрежнему представляет собой задачу на собственные значения, но теперь уже матрицы

$$R = \frac{1}{\mu}\{K'(A-\mu\Lambda)K(K'\Lambda K)^{-1}\},$$

а второе - сводится к кубическому относительно μ уравнению

$$n\mu^3 - \mu^2 Sp\{\Lambda^{-1}L(I+\mathcal{J})^{-1}L' + \Lambda^{-1}A\} + 2\mu Sp\{\Lambda^{-1}A\Lambda^{-1}L(I+\mathcal{J})^{-1}L'\} -$$
$$- Sp\{\Lambda^{-1}L(I+\mathcal{J})^{-1}L'\Lambda^{-1}L(I+\mathcal{J})^{-1}L'\} = 0.$$

Численное решение некоторых специальных модельных задач и анализ полученных результатов показали преимущества выбранной параметризации для оценивания [7,8].Q.

Конкретный учет корреляционных связей в оценке точности измерений особенно существенен при решении обратных задач интерпретации. Показано [8], что при неучете корреляций во входной информации возможно появление ложной флуктуационной структуры в численном решении интегральных уравнений I рода.

Задачи первичной статистической обработки результатов измерений

Другой стороной модели автоматического режима измерений оказывается возможность четкой математической постановки задач первичной статистической обработки реализаций $Y_{1i},\dots,Y_{Ni}$, полученных на дрейфующей аппаратуре. При некоторых дополнительных предположениях о процессе и результатах измерений [9,10], основным из которых является независимость компонент в представлении условного математического ожидания, задачи первичной статистической обработки сводятся: к выявлению и устранению аномальных (резко выделяющихся) наблюдений $Y_{j^*i^*}$, построению оценки интенсивности $\hat{Y}$, проверке исходных допущений (в частности, квазипуассоновости: $V = \mu\Lambda$ потока событий) и т.д. Остановимся кратко на постановке и обосновании методов решения таких задач.

Для построения статистических критериев выявления "сбоя" в совокупности Y_{ji} при фиксированном $i = i_o$ элементы Y_{ji_o} прежде всего нормируются к одному математическому ожиданию $a_{i_o} = MY_{oi}$, где Y_{oi} – либо некоторый фиксированный элемент $Y_{j_o i}$, либо среднее $\bar{Y}_{\cdot i}$. Нормировку можно проводить различными способами, например, через отношения $k_{ji} = Y_{ji}/Y_{oi}$ в предположении, что $p_{ji} = Mk_{ji}$ аппроксимируются в центрированной $2l+1$-окрестности $U^{(l)}_{i_o}$ точки i_o полиномами $\varphi^{(j)}(i) = \sum_{s=1}^{x+1} u^{(j)}_s (i - i_o)^{x+1-s}$ степени x. Элементы набора $\{y_{1i_o},\dots,y_{Ni_o}\}$ нормированных измерений $y_{ji_o} = Y_{ji_o}/\tilde{u}^{(j)}_{x+1}$ ($\tilde{u}^{(j)}_{x+1}$ – оценка $u^{(j)}_{x+1}$ свободного члена полинома $\varphi^{(j)}(i)$) теперь статистически зависимы.

Выдвигается нуль-гипотеза H_o: $My_{1i_o} = \dots = My_{Ni_o} = a_{i_o}$ о равенстве математических ожиданий против альтернативы H_1 – что математическое ожидание одной из величин $y_{ji_o} = y_{j^*}$ (номер j^* которой неизвестен) не совпадает с a_{i_o}. Критерий ("Вертикаль") основывается на статистике (индекс i_o для простоты опускаем)

$$\zeta = \max_j \zeta_j, \quad \zeta_j = \frac{|y_j - \bar{y}_{\cdot}|}{S_j}, \quad S_j^2 = C_j \sum_{k=1}^{N} p_k (y_k - \bar{y}_{\cdot})^2,$$

где C_j выбираются так, чтобы ζ_j были распределены одинаково. Теперь $y_{j^*i^*}$ (а, значит, и $Y_{j^*i_o}$), соответствующее $\zeta_{j^*} = \zeta$, считается "сбитым", если $\zeta \geq z_\varepsilon$, где z_ε – ε-процентная точка распределения ζ, и – "несбитым" в противном случае. Оказывается, что при выборе констант $C_j = \frac{P - P_j}{(N-1)P_jP}$, ($P = \sum_1^N P_j$) функцией распределения является

$$F_{\zeta_j}(x) = 2S_{N-2}\left(x\sqrt{\frac{N-2}{N-1-x^2}}\right) - 1, \qquad (5)$$

а если воспользоваться приближением $P\{\zeta \geq z_\varepsilon\} \approx NP\{\zeta_1 \geq z\}$, то z_ε определяется из уравнения

$$1-S_{N-2}\left(z\sqrt{\frac{N-2}{N-1-z^2}}\right)=\frac{\varepsilon}{2N},$$

где $S_{N-2}(.)$ - функция распределения Стьюдента с $N-2$ степенями свободы. Несмещенной оценкой дисперсии $DY.$ в произвольной точке i при таком подходе будет

$$S^2\bar{Y}.=\frac{P}{N^2(N-1)(1-h^{x+1,x+1})}\sum_{j=1}^{N}P_j(y_j-\bar{y}.)^2,$$

где h^{sk} - элементы матрицы, обратной матрице нормальных уравнений для оценок $\tilde{u}_s^{(j)}$, $s=\overline{1,x+1}$.

Разумнее однако первичную обработку начинать с анализа на качественность каждой отдельной реализации $Y_{j1},\ldots,Y_{jn}$. Соответствующий критерий ("Горизонталь") строится следующим образом. Для совокупности величин $K_i, i\in U_{i_0}^{(l)}$ (j -фиксировано) гипотеза H_0 состоит в том, что $MK_i=\varphi(i)$, а альтернатива H_1 — что $MK_{i^*}(i^*$ неизвестно зарнее) не совпадает с $\varphi(i^*)$. Критической областью критерия для классификации "сбоя" $K_{i^*}(Y_{ji^*})$ является

$$\xi=\max_i\xi_i=\max_i\frac{|K_i-\tilde{\varphi}(i)|}{\theta_i S_K}\geqslant z_\varepsilon,$$

где $S_K^2=\frac{1}{2l-x}\sum_{i=i_0-l}^{i_0+l}q_i(K_i-\tilde{\varphi}(i))^2$ - несмещенная оценка дисперсии на единичный вес, θ_i играют ту же роль, что и C_j, z_ε — ε-процентная точка функции распределения $F_\xi(t)$. Оказывается, что уравнение для приближенного определения z_ε имеет вид

$$1-S_{2l-x-1}\left(z\sqrt{\frac{2l-x-1}{2l-x-z^2}}\right)=\frac{\varepsilon}{2(2l+1)}. \qquad (6)$$

Несмещенная оценка $D\bar{Y}.$ при этом

$$S^2\bar{Y}.=\left(\frac{Y_0}{N}\right)^2\sum_{j=1}^{N}\frac{S_{Kj}^2}{1+\lambda P},$$

где $\lambda=1$, если $Y_0=Y_{j_0}$; $\lambda=-\frac{1}{m}$, если $Y_0=\bar{Y}$.

На подобных же идеях могут быть построены статистические критерии проверки равноточности процесса измерений и выбора степени полинома $\varphi(i)$. В случае квазипуассоновского потока мера точности μ оказывается дисперсией на единичный вес и для величин K_{ji} при выборе весов $a_i/(1-\frac{1}{N}P_{ji})P_{ji}$. Пусть $\tilde{\mu}_j$ - оценки μ, и за нулевую гипотезу принято равенство математических ожиданий оценок $\tilde{\mu}_j$ величине μ, а за альтернативу H_1: $M\tilde{\mu}_1=\ldots=M\tilde{\mu}_{j^*-1}=M\tilde{\mu}_{j^*+1}=\ldots=M\tilde{\mu}_N=\mu$, $M\tilde{\mu}_{j^*}>\mu$. Тогда критическая область выявления измерений $K_{j^*}(Y_{j^*})$, зарегистрированных с меньшей точностью, задается неравенством

$$G=\max_j G_j=\max_j\frac{\tilde{\mu}_j}{\sum_{k=1}^{N}(1-P_k/N)\tilde{\mu}_k}\geqslant z_\varepsilon,$$

а приближенное нахождение z_ε распределения $F_G(t)$ сводится к решению уравнения

$$F_{(N-2)(2l-\varkappa),\,2l-\varkappa}\left(\frac{1-z}{(N-2)z}\right)=\frac{\varepsilon}{N}\,, \qquad (7)$$

где $F_{.,.}(t)$ – функция распределения Фишера.

При решении задачи выбора степени $\varkappa$ показывается, что распределение статистики

$$V_i^2=\sum_{j=1}^{N}\frac{(\tilde{u}_i^{(j)})^2}{P_j}\Big/h''\sum_{j=1}^{N}\frac{S_{kj}^2}{P_j}$$

совпадает с $F_{N-1,\;(2l-\varkappa)(N-1)}(t)$, так что можно воспользоваться обычным фишеровским критерием выбора степени полинома.

Важным свойством критериев "Вертикаль" и "Горизонталь", позволяющим их сравнивать между собой и со стандартными критериями выявления резко выделяющихся наблюдений, является их мощность. Функция мощности (вероятность выявления аномального измерения в зависимости от величины сбоя в Y_{ji}) строятся и исследуются в [9].

Коротко о вопросах обоснования приведенных выше статистических критериев. Если ε -процентные точки распределений ζ, ξ, G находятся из уравнений (5)-(7), то истинный уровень значимости каждого из критериев отличается от выбираемого ε. Однако при выборе за z_ε, например, $F_\zeta(t)$ решения приближенного уравнения $NP\{\zeta_1\geq z\}=\varepsilon$ (7) $\varepsilon_{ист.}=P\{\zeta\geq z\}$ оказывается всегда заниженным, но меньше, чем на $\varepsilon^2/2$. Точнее, если события $A_j=\{\zeta_j\geq z\}$ попарно положительно некоррелированы $\left(P\{A_jA_k\}\leq P\{A_j\}P\{A_k\},\ j\neq k\right)$, то $\varepsilon-\varepsilon^2/2<\varepsilon_{ист.}\leq\varepsilon$, т.е. относительная погрешность не превышает $\varepsilon/2$. Достаточным условием неположительной коррелированности является

$$z_\varepsilon\geq\sqrt{(N-1)\left(1-\frac{1-r}{1+r}K_{N-3}^4\right.}\,,\quad\left(r^2=\frac{P_jP_k}{(P-P_j)(P-P_k)}\,,\ K_n=\sqrt{\frac{2}{n}}\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\right).$$

Такое условие в случае не очень различающихся весов необременительно: если $P_j\to 1$ и $N\to\infty$, то правая часть стремится к $\sqrt{3}$, что не является, конечно, существенным ограничением, поскольку z_ε при фиксированном ε и $N\to\infty$ неограниченно возрастают.

Говоря о задачах первичной статистической обработки данных, нельзя не остановиться на вопросе использования так называемых робастных (устойчивых по эффективности к отклонениям от нормальности) оценок. Использование таких оценок часто рассматривают как некую альтернативу к статистическим процедурам выявления аномальных наблюдений. В теории робастных методов и оценок далеко продвинута

разработка оценивания параметра сдвига для выборок одномерных случайных величин: предложены различные робастные оценки, изучены их асимптотические свойства, проведено различными методами исследование их поведения для малых и больших выборок, а также эффективности. В то же время многомерным оценкам уделено значительно меньше внимания из-за трудностей аналитического характера. Нами проводилось методом статистических испытаний изучение свойств и сравнение многомерных робастных оценок для рекомендаций об их практическом использовании [II,I2].

Рассматривались многомерные аналоги оценок Ходжесса-Лемана

$$Y^{(i)}_{H\text{-}L} = \underset{1\le k\le l\le n}{\operatorname{median}} \left(Y^{(i)}_k + Y^{(i)}_l\right)/2, \qquad (8)$$

α - "обрубленной"

$$Y^{(i)}_{T(\alpha)} = \frac{1}{N-2[N\alpha]} \sum_{k=[N\alpha]+1}^{N-[N\alpha]} Y^{(i)}_{(k)} \qquad (9)$$

и α - "винзоризованной"

$$Y^{(i)}_{W(\alpha)} = \frac{1}{N}\left\{ \sum_{k=[N\alpha]+1}^{N-[N\alpha]} Y^{(i)}_{(k)} + [N\alpha]\left(Y^{(i)}_{([N\alpha])} + Y^{(i)}_{(N-[N\alpha]+1)}\right)\right\}, \qquad (10)$$

где i - номер компоненты случайной величины, N - объем выборки, $0\le\alpha\le 1/2$, $Y_{(k)}$ - k-ая порядковая статистика для выборки. Исследование свойств оценок (8)-(10) методом Монте-Карло проводилось для смеси нормальных распределений $(1-\varepsilon)\mathcal{N}(0,\rho)+\varepsilon\mathcal{N}(0,hI)$, $0\le\varepsilon\le 0.3$, $h=2,4,\ldots,10$, ρ - матрица корреляций. Такая модель была выбрана в связи с интерпретацией резко выделяющихся наблюдений как примеси нормальных некоррелированных многомерных случайных величин с большими дисперсиями. При численном исследовании моделировалось 100 реализаций выборок объема $N=40$. Для каждой выборки строились оценки $\bar{Y}_\cdot$, (8), (9), (10) при $\alpha=0.1$ и 0.25; затем вычислялись для них выборочные (по реализациям) среднее и дисперсия. Относительная эффективность оценок (по сравнению с выборочным средним) полагалась равной отношению мер для выборочных ковариационных матриц. Использовались три меры:

$$1)\ \det A_{\bar{Y}_\cdot}/\det A_t,\quad 2)\ \operatorname{tr} A_{\bar{Y}_\cdot}/\operatorname{tr} A_t,\quad 3)\left(\operatorname{tr} A^2_{\bar{Y}_\cdot}/\operatorname{tr} A^2_t\right)^{1/2},$$

где $A_{\bar{Y}_\cdot}, A_t$ - выборочная ковариционная матрица для $\bar{Y}_\cdot$ и, соответственно, t какая-нибудь из оценок (8)-(9). Приведем некоторые полученные результаты.

Моделирование смеси нормальных распределений в двумерном случае подтвердило теоретические результаты, полученные ранее некоторыми авторами. При $\varepsilon=0$ эффективность робастных оценок $Y_{T(0.1)}$, $Y_{W(0.1)}$

Y_{H-L} оказалось от 0.9 до 0.95 даже при корреляциях, близких к единице; для α=0.25 эффективность несколько ниже (до 0.7-0.8). Вообще в двумерном случае при последующем робастном оценивании налицо высокая эффективность всех рассмотренных робастных оценок. При этом винзоризованные оценки имеют при том же α более высокую эффективность, чем обрубленные, но чувствительнее последних к большим "загрязнениям". Оценка Ходжеса-Лемана показала высокую устойчивость для всех моделированных корреляционных матриц, дисперсий "загрязняющего" распределения и величин "загрязнения". Для двумерных с.в. был рассмотрен также случай специального вида "загрязнения", когда большие ошибки есть только в одной компоненте. Это сказалось в несколько меньших значениях эффективностей робастных оценок по сравнению с выборками, "загрязненными" одним наблюдением с большими ошибками.

В трехмерном случае эффективность всех рассматриваемых оценок в первой мере оказалась чувствительной к коррелированности компонент нормально распределенной с.в. Так, для выборки из нормального распределения с корреляционной матрицей

$$\rho = \begin{Vmatrix} I & 0.853 & 0.5 \\ 0,853 & I & 0.853 \\ 0.5 & 0.853 & I \end{Vmatrix}$$

($\det\rho$ =0.02) эффективности оценок $Y_{T(0.25)}$, $Y_{T(0.I)}$, $Y_{W(0.25)}$, $Y_{W(0.I)}$, Y_{H-L} равны соответственно 0,445, 0.762, 0.375, 0.668, 0.677, что ниже их эффективностей в одномерном случае. Эффективности этих оценок по второй и третьей мерам близки к одномерным. При моделировании смеси нормально распределенных величин картина иная (видимо, благодаря некоррелированности компонент "загрязняющего" распределения): робастные оценки имеют при всех h эффективность больше I. Если за меру эффективности взять tr^2, то их эффективность в этом случае существенно меньше.

Проведенные численные исследования показывают, что робастные оценки Y_{H-L} и $Y_{W(0.25)}$ имеют высокую эффективность во всем диапазоне рассматриваемых величин ε, h, ρ. Однако вычисление оценки (8) более трудоемко, так как требуется порядка $N^2 \ln N$ операций для упорядочения величин (для вычисления (I0) нужно $N \ln N$ операций).

Математические модели при экспериментальном изучении множественности процессов

Во многих естественно-научных экспериментах при регистрации случайных событий приходится иметь дело с так называемыми множественными процессами. Так, в ядерной физике при облучении образца-мишени могут происходить реакции, различающиеся не типом продуктов,

а числом одинаковых вторичных частиц на I акт взаимодействия: например, (γ,n), $(\gamma,2n),\ldots,(\gamma,kn)$ - фотонейтронные реакции множественности k [I3-I6]. При регистрации вторичные частицы естественно не различимы, и для определения парциальных характеристик процесса (например, сечений реакций) привлекаются статистические методы. Аналогичная проблематика возникает в некоторых областях биологии, теории массового обслуживания и др. [I7].

Первый вопрос заключается в стохастическом описании процесса регистрации событий при наличии множественности, что формализуется в схеме. Пусть ξ_ν, $\nu=\overline{1,k}$ - независимые пуассоновские случайные величины (с.в.), характеризующие число событий с ν частицами на акт, $\xi=\xi_1+2\xi_2+\ldots+k\xi_k$ (число вторичных частиц) и пусть число ζ зарегистрированных при данном ξ частиц подчиняется: а) биномиальному $Bi(\xi,\varepsilon)$, б) пуассоновскому $Po(\varepsilon\xi)$ законам, где ε - вероятность регистрации частицы. Задача нахождения безусловного распределения ζ приводит [I4,I7] к производящим функциям (п.ф.) $\varphi_\zeta(z)$:

а) $\exp\{\sum_{\nu=1}^{k}x_\nu(z^\nu-1)\}$, б) $\exp\{\sum_{\nu=1}^{k}(y_\nu e^{\varepsilon\nu z}-\lambda_\nu)\}$,

$x_\nu=\varepsilon^\nu\sum_{i=\nu}^{k}C_i^\nu(1-\varepsilon)^{i-\nu}\lambda_i$; $y_\nu=\lambda_\nu e^{-\varepsilon\nu}$, $\nu=\overline{1,k}$.

Эквивалентные постановки задачи формулируются в терминах суммирования случайного числа с.в. [I7].

Имеют место следующие представления: для функций вероятностей P_n, $n=0,1,\ldots$

а) $\varphi_\xi(0)\sum_{\sum_{\nu=1}^{k}\nu i_\nu=n}\prod_{\nu=1}^{k}\frac{x_\nu^{i_\nu}}{i_\nu!}$; б) $\varphi_\xi(0)\frac{\varepsilon^n}{n!}\sum_{m=0}^{n}g_n^{(m)}(y_\nu)$, (II)

где $g_n^{(m)}(y_\nu)$ задаются п.ф. $\left(\sum_{\nu=1}^{k}y_\nu(e^{\nu z}-1)\right)^m/m!$; для семиинвариантов $\varkappa_\ell$, $\ell=1,2,\ldots$

а) $\sum_{\nu=1}^{k}\nu^\ell x_\ell$; б) $\sum_{\tau=0}^{\ell}\sum_{\nu=1}^{k}(\varepsilon\nu)^\tau\sigma_\ell^{(\tau)}\lambda_\nu$,

для начальных моментов α_ℓ, $\ell=1,2,\ldots$

а) $\sum_{\nu=0}^{\ell-1}C_{\ell-1}^{\nu}\varkappa_{\nu+1}\alpha_{\ell-\nu-1}$; б) $\sum_{m=0}^{\ell}\sum_{\tau=m}^{\ell}\varepsilon^\tau\sigma_\ell^{(\tau)}g_\ell^{(m)}(\lambda_\nu)$,

где $\sigma_\ell^{(\tau)}$ - числа Стирлинга 2-го рода; для производных по параметрам

а) $\frac{\partial P_n}{\partial x_\nu}=-P_n+P_{n-\nu}$; б) $\frac{\partial P_n}{\partial y_\nu}=\sum_{\ell=0}^{n}\frac{(\varepsilon\nu)^{n-\ell}P_\ell}{(n-\ell)!}-P_n e^{\varepsilon\nu}$, (I2)

$\frac{\partial P_n}{\partial\varepsilon}=\frac{1}{\varepsilon}\left(nP_n-(n+1)P_{n+1}\right)$.

Распределения являются воспроизводящими по $\lambda_1,\ldots,\lambda_k$. При $\alpha_1\to\infty$ с.в. ζ асимптотически нормальна с параметрами α_1, $\sqrt{\varkappa_2}$, если $\varepsilon=Const$; при $\varepsilon\to0$ второе распределение стремится к пуассоновскому $Po(\varepsilon\alpha_1)$.

Отметим, что при $k=1$ распределение а) переходит в $Po(\varepsilon\lambda_1)$, а б) — в введенное Нейманом "инфекционное" распределение типа А от двух параметров.

Решение прямой задачи, т.е. вычисление P_n в зависимости от параметров, сведено к рекуррентным соотношениям:

$$\text{а)}\ nP_n=\sum_{\nu=1}^{k}\nu x_\nu P_{n-\nu};\quad \text{б)}\ nP_n=\sum_{\ell=0}^{n-1}\sum_{\nu=1}^{k}\frac{(\varepsilon\nu)^{n-\ell}y_\nu P_\ell}{(n-\ell-1)!},\ n=1,2,\ldots \quad (13)$$

Установлена их вычислительная устойчивость. Численные расчеты P_n в некоторых важных характерных случаях представлены графически и проанализированы [13-15].

Основная проблема экспериментального изучения множественных процессов заключается в оценивании парциальных характеристик [15]. Пусть за N наблюдений зарегистрировано $n_1,\ldots,n_N$ вторичных частиц. Зная функциональный вид $P_n(\lambda_\nu)$ распределений (11), требуется построить оценки $\lambda_1,\ldots,\lambda_k$ или их комбинаций. Уравнения максимального правдоподобия (м.п.) выписываются для произвольной множественности k. Если эффективность регистрации ε известна, то уравнения м.п. для случая а) в системе параметров x_ν имеют следующий экзотический вид

$$\sum_{i=1}^{N}P_{n_i-\nu}/P_{n_i}=N,\quad \nu=\overline{1,k}$$

и приводят к явной оценке для $\alpha_1=\sum_1^k\nu x_\nu$, равной первому выборочному моменту a_1. При неизвестном ε уравнения м.п. в случае б) запишутся

$$\sum_{i=1}^{N}\frac{(n_i+1)P_{n_i+1}}{P_{n_i}}=Na_1,\quad \sum_{i=1}^{N}\sum_{\ell=0}^{n_i}\frac{(\varepsilon\nu)^{n_i-\ell}P_\ell}{(n_i-\ell)!P_{n_i}}=Ne^{\varepsilon\nu},\quad \nu=\overline{1,k}$$

и дают в качестве первого интеграла м.п.- оценку a_1 для $\alpha_1=\varepsilon\sum_1^k\nu\lambda_\nu$.

Для конкретных k получены уравнения метода моментов и его модификаций. При $k=2$ просчитаны дисперсии оценок, позволяющие сравнивать методы по точности и указать их эффективность. При этом на основании неравенства Рао-Крамера с помощью (12) и (13) нижние границы дисперсий оценок вычислены на ЭВМ. Исследование их поведения в зависимости от единственного управляемого в эксперименте параметра α_1(интенсивность счета) позволяет указать оптимальные по точности оценивания условия измерений. Изучены необходимые с точки зрения экспериментальной практики обобщения приведенной математической теории множественности на непостоянство интенсивности первичного пучка. Полученные результаты численного решения соответствующих прямых и обратных задач использованы при проектировании и обработке экспериментов по разделению фотонейтронных реакций

множественности 2 [I4-I6,35].

О математических моделях при обработке мессбауэровских спектров

Ядерный гамма-резонанс (ЯГР) является мощным инструментом исследования физико-химических свойств (кристаллической структуры, характера связи атомов и т.д.) твердых тел. Информация о характеристиках резонансного поглощения извлекается из экспериментальных мессбауэровских спектров. Теоретические посылки о взаимодействии ядер (на которых возможен эффект Мессбауэра) с электрическими и магнитными полями в образце приводят к моделям ЯГР-спектров вида суперпозиции лоренцевских резонансных линий

$$MY_i = N_\infty - \sum_{j=1}^{m} \frac{A_j}{\left(\frac{i-a_j}{\Gamma/2}\right)^2+1}, \quad i=\overline{1,n}, \qquad (I4)$$

либо к некоторому интегральному представлению [36]. При этом геометрические параметры модели определяются физическими характеристиками явления через определенные функциональные связи $A_j = A_j(\beta)$, $a_j = a_j(\beta)$, где β - вектор, компоненты которого есть параметры квадрупольного взаимодействия, изомерного сдвига, асимметрии ГЭП, анизотропии и т.д. [I8].

Прямые задачи при этом сводятся к оценке m в (I4), выводу связей на A_j, a_j и расчету спектра в зависимости от компонент β. Собственно интерпретация экспериментального ЯГР-спектра может проводиться в двух аспектах. При фиксированной модели (I4) необходимо по результатам измерений Y_i оценить компоненты β. Однако физическое исследование часто предполагает ситуацию, когда теоретические посылки приводят к различным представлениям (I4). В математической постановке это формализуется наложением связей на β, и задача интерпретации сводится к проверке определенных статистических гипотез результатами эксперимента [I9].

Разработанная в НИВЦ и включенная в математическое обеспечение ЭВМ МГУ библиотека программ мессбауэровской спектроскопии широко используется при интерпретации экспериментальных спектров. Решены конкретные задачи анализа ЯГР-спектров соединений: перовскитных сложного состава, органических сурьмы, интерметаллических типа фаз Лавеса, сталей и др. [20-23]. Поскольку форма экспериментальных мессбауэровских спектров поглощения и дифракции зависит и от аппаратурных факторов (эффективной толщины поглотителей, геометрических параметров установки и др., возникают задачи анализа этих зависимостей и учета их при проектировании и организа-

ции мессбауэровского эксперимента [24-26].

Численное исследование моделей кинетики гамма-излучающих систем

Создание лазерных систем с длиной волны порядка $1\mathring{A}$ (на основе мессбауэровских изотопов) - одна из важнейших нерешенных научно-технических проблем. В последние годы при этом в теоретических и экспериментальных исследованиях математическое моделирование и проектирование (численный эксперимент) стали играть определяющую роль. В коллективе физиков и математиков МГУ изучены математические модели кинетики генерации и усиления γ-излучения, использующие квазиклассический (поле классическое, среда квантуется) и квантовомеханический (квантуются и среда, и поле) подходы. Описание процессов переноса и взаимодействия γ-излучения с возбужденной резонансной средой сведено [27-32] к начально-краевым задачам для квазилинейных систем первого порядка

$$u_t(x,t)+G\,u_x(x,t)=F(u,x,t) \qquad (15)$$

(u -неизвестная комплекснозначная вектор-функция размерности $n=3$ или 6, G – $n\times n$ -матрица) с начальным и граничным условиями первого рода на концах стержня длины ℓ

$$u(x,0)=\varphi(x);\ u^i(0,t)=\psi_i(t),\ u^j(\ell,t)=\psi_j(t),\ i\in I_1, j\in I_2;\ I_1, I_2\subset\overline{1,n},\ I_1\cap I_2=\emptyset. \qquad (16)$$

Разрешены вопросы существования и единственности классического решения в конечной области задач типа (15), (16), исследованы постановки и математические свойства решений [29-32]. Обоснованы и применены некоторые варианты метода характеристик численного интегрирования [29,32]. Численные алгоритмы реализованы в специализированном программном обеспечении БЭСМ-6, составившем комплекс программ для решения определенных задач кинетики гамма-лазера. Решены следующие конкретные физические задачи.

При квазиклассическом подходе: $u=(A,P,\Delta n)$, A - амплитуда вектор-потенциала, P - ток ядерного перехода, Δn - инверсная населенность, $G_{11}=c$, $G_{ij}=0$, $i+j>2$,

$$F=\left(\mu P-\frac{c}{2\ell_1}A+\varkappa,\ C\Delta nA-(i\varepsilon+\Gamma)P,\ -\mathcal{D}\,\mathrm{Re}(AP^*)-1-\Delta n\right)'.$$

Задача усиления сигнала заданной формы: $\varkappa(x,t)=0$, $\varphi(x)=(0,0,1)'$, $\psi_1(t)=A(t)$. Задача генерации и кинетики спонтанного излучения: $\varkappa(x,t)$ – δ-коррелированный по x случайный процесс с нулевым средним, начальные условия те же, $\psi_1(t)=0$ (внешний сигнал отсутствует).

При квантовомеханическом подходе: $u=(n_1, n_2, f_1, f_2, S, \Delta n)'$, n_1, n_2 — плотности числа квантов волн вправо и влево, f_1, f_2 и S описывают обмен полей со средой и корреляцию излучателей, $G_{11} = -G_{22} = -G_{33} = -G_{44} = I$, остальные $G_{ij} = 0$,

$$F=\left(f_k-n_k, \frac{\beta}{2}\left[n_k\Delta n+\frac{S}{2}+g(1+\Delta n)\right]-\alpha_1 f_k, \Delta n\sum_k f_k-\alpha_1 S, -2\sum_k f_k\right)'$$

(коэффициенты в F определяются параметрами среды). Двухволновая задача сверхизлучательной кинетики: $k=I,2$, $\varphi(x)=(0,0,0,0,0,I)'$, $\psi_i(t)=0$, $i=\overline{I,4}$. Одноволновая задача: $k=I$, $\varphi(x)=(0,0,0,I)'$, $\psi_1(t)=\psi_2(t)=0$. Чисто временная задача: $k=0$, $\varphi(x)=(0,0,0,I)'$, сводящаяся к задаче Коши для системы обыкновенных дифференциальных уравнений.

Интерпретация численного решения этих задач позволила сделать важные практические выводы. Указаны физически реализуемые параметры среды, наиболее выгодные для реализации режимов и типов излучения, проведено сопоставление самих физических подходов [29-31]. Получен так называемый критерий слабого усиления [29-33]. Исследовано влияние процесса сужения линии излучения $\Gamma(t)$ на временное поведение импульсов [29,34]. Проведены численный анализ и физическая интерпретация пространственно-временного характера режимов сверхизлучения и суперлюминесценции в квантовомеханическом подходе [30,34]. Ближайшие перспективы в этом направлении связаны с решением многомерных задач и с учетом температурных факторов [37,38].

Литература

I. **On the everall automatik of data** processing for determining the **photonuclear reaction cross-cection** Tikhonov **A.N.** ,Shevohenko B.G., Galkin V.Ja., Gorjachev B.I., Zaikin P.N.,Ishanov B.S., Kapitonov I.M.- **Information processing** 68.North-Holland Publishing Company-Amsterdam., 1969,p.1549-1551.

2. Система сплошной автоматической обработки результатов эксперимента по исследованию фотоядерных реакций/ Тихонов А.Н., Шевченко В.Г., Галкин В.Я., Заикин П.Н., Горячев Б.И., Ишханов Б.С., Капитонов И.М. - В кн.: Вычислит. методы и программирование. Вып. XIУ. М.: Изд-во Моск. ун-та, I970, с. 3-26.

3. Тихонов А.Н. О математических методах автоматизации обработки наблюдений. - В сб.: Проблемы вычислительной математики. М.: Изд-во Моск. ун-та, I980, с. 3-I7.

4. Тихонов А.Н., Арсенин В.Я. Методы решения некорректных задач. - М.: Наука, I979.

5. Тихонов А.Н. Вычислительная математика. - В кн.: БСЭ, изд. 3-е, т. 5, М., I97I, с. 568-569.

6. Тихонов А.Н. Математическая модель. - В кн.: БСЭ, изд. 3-е, т. I5. М., I974, с. 480.

7. Применение факторного анализа для учета аппаратурного дрейфа при обработке результатов автоматических измерений/ Тихонов А.Н., Большев Л.Н., Галкин В.Я., Орлин В.Н., Уфимцев М.В. - В сб.: Вычисл. методы и программир. Вып. XXI. М.: Изд-во Моск. ун-та, I973, с. II8-I23.

8. Галкин В.Я., Орлин В.Н., Уфимцев М.В. Учет аппаратурного дрейфа при решении обратных задач. - В кн.: Всесоюзная конференция по некорректно поставленным задачам. Фрунзе: Изд-во Илим, I979, с. 39-40.

9. Галкин В.Я., Орлин В.Н. Первичная обработка экспериментальных данных, полученных в автоматическом режиме измерений. - В сб.: Вычисл. методы и программир. Вып. XXI. М.: Изд-во Моск. ун-та, I973, с. 74-I07.

I0. Галкин В.Я. Статистические задачи обработки и интерпретации результатов физических экспериментов. - В кн.: Труды Всесоюзной школы "Методы решения некорректн. задач и их применение". М.: Изд-во Моск. ун-та, I974, с. I20-I28.

II. Галкин В.Я., Уфимцев М.В. Свойства некоторых оценок параметра сдвига, устойчивых к отклонениям от нормальности. - В сб.: Обработка и интерпретация физич. экспериментов. М.: Изд-во Моск. ун-та, I976, с. 25-4I.

I2. Галкин В.Я., Уфимцев М.В. Монте-карловское исследование многомерных робастных оценок параметра сдвига. - Математическая статистика и ее приложения, I98I, № 7, с. 20-23.

I3. Галкин В.Я. Прямые задачи при разделении множественности ядерных процессов. - ДАН СССР, I974, т. 2I6, № 5, с. I0I4-I0I7.

I4. Галкин В.Я., Уфимцев М.В. Исследование одного класса дискретных распределений. - В сб.: Вычислит. методы и программирование. Вып. XXУI. М.: Изд-во Моск.ун-та, I977, с. 36-56.

I5. Галкин В.Я., Горячев Б.И., Орлин В.Н., Уфимцев М.В. Об оптимальной обработке и организации эксперимента по статистическому разделению выходов ядерных реакций различной множественности. - В сб.: Вычислит. методы и программирование. Вып. XУШ. М.: Изд-во Моск. ун-та, I972, с. I6I-I72.

I6. Галкин В.Я. Прямые и обратные задачи при разделении множественности ядерных процессов. Препринт ОИЯИ, ДIO-7707. - Дубна, I974, с. 93-99.

I7. Галкин В.Я., Уфимцев М.В. Распределение продуктов множественных процессов и "инфекционное" распределение Неймана. - Вестн. Моск. ун-та, сер. I5. Вычисл. матем. и киберн., I982, № 2, с. 4I-5I.

I8. Некоторые вопросы обработки мессбауэровских спектров/ Галкин В.Я., Горьков В.П., Заикин П.Н., Кузьмин Р.Н., Новакова А.А. - В сб.: Вычисл. методы и программирование. Вып. XXI. М.: Изд-во Моск. ун-та, I973, с. I08-II7.

I9. Галкин В.Я., Горьков В.П. Об обработке мессбауэровских спектров. Препринт ОИЯИ, ДIO-7707. - Дубна, I974, с. 23I-237.

20. Особенности эффекта Мессбауэра в теллуре/ Опаленко А.А., Авенариус И.А., Горьков В.П., Комиссарова Б.А., Кузьмин Р.Н., Заикин П.Н. - ЖЭТФ, I972, т. 62, № 3, с. I037-I042.

2I. Исследование некоторых органических соединений сурьмы методом резонансной спектроскопии/ Гукасян С.Е., Горьков В.П., Заикин П.Н., Шпинель В.С. - Журн. структурной химии, I973, т. XIУ, № 4, с. 650-655.

22. Анализ спектров мессбауэровского поглощения в интерметаллических соединениях со структурой фаз Лавеса/ Горьков В.П., Митрофанов К.П., Рясный Г.К., Сорокин А.А. - В сб.: Обработка и интерпретация физич. экспериментов. М.: Изд-во Моск. ун-та, I978, с. 79-88.

23. Исследование сложных перовскитских соединений методом мессбауэровской спектроскопии/ Тихонов А.Н., Галкин В.Я., Горьков В.П., Кузьмин Р.Н., Ройнг Нуньес Х., Шагдаров В.Б. - В совм. науч. сборнике ОИЯИ (Дубна, СССР) и ЦИФИ (Будапешт, ВНР), вып. 3. Будапешт: МТА, I979, с. I-8.

24. Mitrofanov K.P. and Gorkov V.P. Asymmetry of Mössbauer Lines due to the finite dimensions of the source.- Nucl. Inst. and Meth., 1973, v. 112, p. 427 - 429.

25. Численный анализ влияния ядерного резонансного поглощения на мессбауэровские спектры дифракции/ Андреева М.А., Галкин В.Я., Горьков В.П., Кузьмин Р.Н., Тихомиров О.Ю. - В сб.: Обработка и интерпретация физич. экспериментов. М.: Изд-во Моск. ун-та, I976, с. III-I42.

26. Горьков В.П. Учет влияния аппаратурных факторов в мессбауэровской спектроскопии. - В сб.: Обработка и интерпретация физич. экспериментов. М.: Изд-во Моск. ун-та, I980, с. 99-I04.

27. Некоторые математические задачи кинетики усиления спонтанного излучения ядер в гамма-лазере/ Тихонов А.Н., Бушуев В.А., Галкин В.Я., Кузьмин Р.Н., Тихомиров О.Ю. Препринт ОИЯИ, ДIO, II-II264. - Дубна, I978, с. 5-9.

28. Математическая модель кинетики высвечивания систем двухуровневых излучателей/ Тихонов А.Н., Андреев А.В., Галкин В.Я., Тихомиров О.Ю. Препринт ОИЯИ, Д10, 11-11264. - Дубна, 1978, с. 10-14.

29. Математическое моделирование процессов усиления и генерации излучения в гамма-лазере/ Тихонов А.Н., Бушуев А.В., Галкин В.Я., Кузьмин Р.Н., Тихомиров О.Ю. - В совм. науч. сборнике ОИЯИ (Дубна, СССР) и ЦИФИ (Будапешт, ВНР), вып. 3. Будапешт: МТА, 1979, с. 147-164.

30. Численный анализ пространственного развития лавины сверхизлучения/ Тихонов А.Н., Андреев А.В., Галкин В.Я., Ильинский Ю.А., Тихомиров О.Ю. - В совм. науч. сборнике ОИЯИ (Дубна, СССР) и ЦИФИ (Будапешт, ВНР), вып. 3, Будапешт: МТА, 1979, с. 131-146.

31. Андреев А.В., Бушуев В.А., Тихомиров О.Ю. Математические модели кинетики генерации и усиления гамма-излучения. - ДАН СССР, 1980, т. 252, № 4, с. 845-848.

32. Галкин В.Я. О численном решении задач кинетики гамма-излучения. - ДАН СССР, 1980, т. 255, № 4, с. 833-836.

33. О критерии слабого усиления в квазиклассической модели кинетики гамма-лазера/ Тихонов А.Н., Бушуев В.А., Галкин В.Я., Кузьмин Р.Н., Тихомиров О.Ю. - В сб.: Обработка и интерпретация физич. экспериментов. М.: Изд-во Моск. ун-та, 1979, с, 121-125.

34. Алгоритмическое и программное обеспечение исследований в проблеме гамма-лазера/ Тихонов А.Н., Андреев А.В., Бушуев В.А., Галкин В.Я., Кузьмин Р.Н., Тихомиров О.Ю. - В кн.: Автоматизация научн. исследований на основе применения ЭВМ. Новосибирск: СО АН СССР, 1979, с. 203-204.

35. Тихонов А.Н., Галкин В.Я., Орлин В.Н. Статистические критерии при интерпретации структуры гигантского резонанса на тяжелых деформированных ядрах. - В сб.: Обработка и интерпретация результатов наблюдений. М.: Изд-во Моск. ун-та, 1981, с. 3-26.

36. Горьков В.П. Интегральное представление мессбауэровского спектра. - В сб.: Обработка и интерпретация результатов наблюдений. М.: Изд-во Моск. ун-та, 1981, с. 108-113.

37. Квазиклассическая модель кинетики -излучения с учетом температурного разогрева/ Бушуев В.А., Галкин В.Я., Кузьмин Р.Н., Манцызов Б.И., Серебряков С.Л., Тихомиров О.Ю. - В сб.: Обработка и интерпретация физических экспериментов. М.: Изд-во Моск. ун-та, 1980, с. 3-17.

38. Квазиклассическая модель генерации гамма-излучений в силь-

ноусиливающей среде/ Бушуев В.А., Галкин В.Я., Кузьмин Р.Н., Манцызов Б.И., Серебряков С.Л., Тихомиров О.Ю. - В сб.: Обработка и интерпретация результатов наблюдений. М.: Изд-во Моск. ун-та, 1981, с. 59-73.

39. Лоули Д., Максвелл А. Факторный анализ как статистический метод. - М.: Мир, 1967.

СОВРЕМЕННЫЕ ЧИСЛЕННЫЕ МЕТОДЫ РЕШЕНИЯ СЕТОЧНЫХ УРАВНЕНИЙ

А.А. САМАРСКИЙ, И.Е. КАПОРИН, А.Б. КУЧЕРОВ, Е.С. НИКОЛАЕВ

Московский государственный университет им. М.В. Ломоносова,
Москва

АННОТАЦИЯ

Статья посвящена методам решения важного класса разреженных систем - пятиточечных разностных краевых задач для эллиптических уравнений 2-го порядка на прямоугольной сетке. Дается классификация задач и описание соответствующих методов.

I.<u>Задачи с разделяющимися переменными в прямоугольнике</u>. <u>Экономичные прямые методы</u>.

I.I.Постановки основных подзадач. Все известные алгоритмы решения задач этого пункта базируются на эффективных методах для следующих двух задач.

Первая из них - задача решения системы линейных уравнений $Sx=b$ с матрицей

$$S=\begin{pmatrix} c_1 & -b_1 & 0 & \dots & 0 & -a_1 \\ -a_2 & c_2 & -b_2 & \dots & 0 & 0 \\ 0 & -a_3 & c_3 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & c_{M-1} & -b_{M-1} \\ -b_M & 0 & 0 & \dots & -a_M & c_M \end{pmatrix} \qquad (I)$$

Для решения этой задачи применяются различные варианты разреженного исключения Гаусса (прогонка [I] , циклическая редукция [2] , имеющие асимптотику числа операций $O(M)$.

Вторая задача - умножение на вектор матриц Q и Q^{-1}, приводящих матрицу

$$T=\begin{pmatrix} \xi_1 & -\eta_1 & 0 & \dots & 0 & -\zeta_1 \\ -1 & 2 & -1 & \dots & 0 & 0 \\ 0 & -1 & 2 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 2 & -1 \\ -\zeta_2 & 0 & 0 & \dots & -\eta_2 & \xi_2 \end{pmatrix} \qquad (2а)$$

к диагональному виду:

$$Q^{-1}TQ=\Lambda, \qquad \Lambda=\mathrm{diag}(\lambda_1,\dots,\lambda_M).$$

Для специальных случаев, например

$$(\xi_\alpha,\eta_\alpha,\zeta_\alpha)\in\{(2,-1,0),(2,-2,0)\},\ \alpha=1,2, \qquad (2б)$$

либо

$$(\xi_\alpha,\eta_\alpha,\zeta_\alpha)=(2,-1,-1),\quad \alpha=1,2, \qquad (2в)$$

известны явные выражения элементов Q, Q^{-1} и Λ (через тригонометрические функции) и эффективные алгоритмы решения этой задачи (основанные на применении быстрого преобразования

Фурье [1,2].
Число операций указанных алгоритмов имеет асимптотику $O(M\log M)$.

1.2. Постановки задач с разделяющимися переменными в прямоугольнике. Рассматриваемый класс задач в общем случае есть системы линейных алгебраических уравнений с матрицей порядка MN

$$Ay=f, \quad A = T_1 \otimes I_N + I_M \otimes T_2, \qquad (3)$$

где T_1, T_2 – матрицы вида (1) порядков соответственно M и N, $\otimes$ – кронекеровское произведение:

$$A \otimes B = \begin{pmatrix} b_{11}A & b_{12}A & \dots & b_{1M}A \\ b_{21}A & b_{22}A & \dots & b_{2M}A \\ \dots & \dots & \dots & \dots \\ b_{M1}A & b_{M2}A & \dots & b_{MM}A \end{pmatrix}.$$

Если одна из матриц T_α имеет вид (2), то систему (3) называют задачей с постоянными коэффициентами по одному направлению

1.3. Метод циклической редукции (CR). Метод CR изложим для задачи с постоянными коэффициентами по одному направлению, в том случае, когда

$$T_2 = \begin{pmatrix} 2 & -1 & & 0 \\ -1 & 2 & -1 & \\ & \ddots & \ddots & \ddots \\ 0 & & -1 & 2 \end{pmatrix}.$$

Соответствующую систему (3) можно записать в виде трехточечного векторного уравнения с краевыми условиями первого рода:

$$\begin{aligned} &-Y_{j-1} + SY_j - Y_{j+1} = F_j, \quad 1 \le j \le N, \\ &Y_0 = 0, \quad Y_{N+1} = 0, \end{aligned} \qquad (4)$$

где $Y_j = (y_j(1), \dots, y_j(M))^T$ – j-й неизвестный вектор, $S = 2I + T_1$. Дополнительно предположим, что $N = 2^n - 1$.

Метод CR представляет собой рекурсивную процедуру [3], применяемую к аналогичному уравнению вида

$$\begin{aligned} &-W_{j-1} + CW_j - W_{j+1} = (C + \beta I)P_j, \\ &1 \le j \le 2^{n-k} - 1, \quad W_0 = 0, \quad W_{2^{n-k}} = 0, \end{aligned} \qquad (5)$$

где

$$C = \overset{k}{C} = 2\,T_{2^k}\left(\frac{S}{2}\right), \quad T_r(t)=\cos(r\arccos t), \qquad (6а)$$

$$\beta = \overset{k}{\beta}, \quad \overset{k+1}{\beta} + \overset{k}{\beta}{}^2 = 2, \quad k \geqslant 0. \qquad (6б)$$

Заметим, что при $k=0$,

$$P_j = (C - \overset{0}{\beta} I)^{-1} F_j, \qquad (*)$$

получается исходная задача.

Параметром, характеризующим размер задачи (5), является число k. Рекурсивный шаг метода CR заключается в переходе к системе (5) с параметром размера, на единицу бóльшим:

$$-\hat{W}_{j-1} + \hat{C}\hat{W}_j - \hat{W}_{j+1} = (\hat{C} + \hat{\beta} I)\hat{P}_j,$$
$$1 \leqslant j \leqslant 2^{n-k-1} - 1, \quad W_0 = 0, \quad W_{2^{n-k-1}} = 0, \qquad (7)$$

где

$$W_{2j} = \hat{W}_j, \quad 0 \leqslant j \leqslant 2^{n-k-1}, \quad \hat{C} = \overset{k+1}{C}, \quad \hat{\beta} = \overset{k+1}{\beta}. \qquad (8)$$

Это осуществляется путем подстановки выражений

$$W_{2j-1} = P_{2j-1} + C^{-1}(W_{2j-2} + W_{2j} + \beta P_{2j-1}), \quad 1 \leqslant j \leqslant 2^{n-k-1}, \qquad (9)$$

в уравнении (5) для четного j и домножения на C. После преобразований, учитывающих (6), получаем формулы для $\hat{P}_j$:

$$\hat{P}_j = P_{2j} + (C - \beta I)^{-1}(P_{2j-1} + \beta P_{2j} + P_{2j+1}), \qquad (10)$$
$$1 \leqslant j \leqslant 2^{n-k-1} - 1.$$

Таким образом, рекурсивный шаг метода CR описывается формулами (10), (8), (9). Шаги выполняются до тех пор, пока не получится задача (5) при $k=n-1$. Если последовательность $\overset{k}{\beta}$ такова, что $\overset{n-1}{\beta} = 0$ (например, $\overset{k}{\beta} = 2\cos\left[\frac{\pi}{3}\left(1 - \left(-\frac{1}{2}\right)^{n-k}\right)\right]$), то решение этой задачи уже получено: $W_j = P_j$, и на его отыскание не затрачиваются арифметические операции. Как известно [3], число операций $\mathcal{N}(k)$ рекурсивного алгоритма для решения задачи размера k удовлетворяет рекуррентному соотношению

$$\mathcal{N}(k) = \mathcal{N}(k+1) + \mathcal{N}_1(k),$$

где $\mathcal{N}_1(k)$ - число операций, затрачиваемое на выполнение рекурсивного шага. Отсюда, учитывая, что $\mathcal{N}(n-1) = 0$, получаем

$$\mathcal{N}(k) = \sum_{j=k}^{n-2} \mathcal{N}_1(j).$$

Так как исходная задача (4) приводится к задаче (5) с $k=0$ путем преобразования $(*)$ правых частей, на которое затрачива-

ется $\mathcal{N}_0$ операций, то общее число операций метода CR есть

$$\mathcal{N} = \mathcal{N}_0 + \sum_{j=0}^{n-2} \mathcal{N}_1(j). \qquad (11)$$

Величина $\mathcal{N}_0$ есть $O(MN)$, так как (*) выполняется путем применения, например, метода прогонки. Вычисления (9), (10) также сводятся к умножению матрицы, обратной к трехдиагональной, на вектор и векторному сложению. Для этого достаточно представить $(C-\alpha I)^{-1}$ в виде суммы простейших дробей:

$$(C-\alpha I)^{-1} = (2T_{2^k}(\frac{S}{2}) - \alpha I)^{-1} = \sum_{j=1}^{2^k} \xi_j (S+\eta_j I)^{-1}. \qquad (12)$$

Такое представление существует, поскольку все корни многочленов $2T_{2^k}(t/2)-\alpha$, $\alpha \in \{0, \beta^k\}$, - простые. Отсюда становится ясным, что величина $\mathcal{N}_1(k)$ также есть $O(MN)$. Поэтому из (11) получаем асимптотику $O(MN \log N)$ для числа операций метода CR.

Отметим, что описанная методика построения алгоритма успешно обобщается и на те случаи, когда матрица T_2 имеет любую из форм (2) при соответствующих ограничениях на N.

Известные к настоящему времени варианты метода CR используют, как правило, представление правой части (5) в виде $Cp_j + q_j$ *) и требуют в полтора раза больше памяти при реализации на ЭВМ. Алгоритмы указанного типа распространены на случай произвольного N [4], а также на общий случай, когда как T_2, так и T_1 имеют вид (1) [5]. При этом асимптотика числа операций сохраняется.

I.4. Дискретное преобразование Фурье (FA). Рассмотрим задачу типа (3) с матрицей

$$\mathcal{A} = (C-2I_M)\otimes I_L + I_M \otimes T_2, \qquad (13)$$

где C - полином степени K от матрицы вида (1), а T_2 является $L\times L$ - матрицей типа (2). Подставляя разложение $T_2 = Q\Lambda Q^{-1}$ в (13), получаем $\mathcal{A} = (C-2I_M)\otimes I_L + I_M \otimes Q\Lambda Q^{-1}$, откуда, используя формулу $(A\otimes B)(C\otimes D) = (AC)\otimes(BD)$,

---

*) А также факторизацию $(C-\alpha I)^{-1} = \prod_{j=1}^{2^k}(S+\eta_j I)^{-1}$ вместо разложения (12), что менее выгодно в смысле численной устойчивости.

нетрудно получить

$$\mathcal{A} = (I_M \otimes Q)((C-2I_M)\otimes I_L + I_M \otimes \Lambda)(I_M \otimes Q^{-1}).$$

Отсюда, учитывая что средний сомножитель есть блочно-диагональная матрица (с блоками $C + (\lambda_j - 2) I_M$ на диагонали), можно получить искомый алгоритм FA для задачи (3), (I3):

$$y = (I_M \otimes Q) \begin{pmatrix} [C+(\lambda_1-2)I_M]^{-1} & & \\ & \ddots & \\ & & [C+(\lambda_L-2)I_M]^{-1} \end{pmatrix} (I_M \otimes Q^{-1}) f.$$

Алгоритм умножения на вектор первого и третьего сомножителя очевидным образом получается применением M раз второго алгоритма п°1.1 , а умножение на вектор второго сомножителя строится по формуле (I2) и требует KL применений первого алгоритма п°1.1 . Таким образом, число операций метода FA для задачи (3) с матрицей (I3) имеет асимптотику

$$O(KLM + ML\log L).$$

Так, если мы применим построенный метод непосредственно к задаче (4), для которой $K=1$, $L=N$, то получим метод с асиммптотикой числа операций $O(MN\log N)$. Если же мы используем метод FA в рекурсивной процедуре CR - выполним ℓ шагов рекурсии и для полученной задачи (5) с параметром размера $k=\ell$ (которая есть задача (3), (I3) при $K=2^\ell$, $L = \frac{N+1}{2^\ell} - 1$) применим метод FA - то число операций такого метода (получившего название $FACR(\ell)$) будет удовлетворять рекуррентному соотношению

$\mathcal{N}(k) = \mathcal{N}(k+1) + \mathcal{N}_1(k)$, $k=0,1,\dots,\ell-1$, $\mathcal{N}(\ell) = O(MN(1+2^{-\ell}\log N))$.

Отсюда получаем $\mathcal{N}(0) = O(\ell MN + 2^{-\ell} MN \log N)$, так что при $\ell = c_0 \log\log N$ метод $FACR(\ell)$ имеет асимптотику числа операций $O(MN\log\log N)$, т.е. лучшую, чем для CR и FA.

Другой вариант применения алгоритмов п°1.1 изложен в следующем пункте.

I.5. Маршевый метод (MA). Рассмотрим задачу (3), где $T_1 = S$ - матрица вида (2). Запишем ее в виде трехточечного векторного уравнения с периодическими краевыми условиями:

$$-\alpha_j Y_{j-1} + (\gamma_j I + S) Y_j - \beta_j Y_{j+1} = F_j', \quad 1 \le j \le N, \qquad (I4a)$$

$$Y_0 = Y_N, \quad Y_{N+1} = Y_1. \qquad (I4б)$$

Основную роль при построении маршевого метода играют соотношения, связывающие векторы Y_m, Y_{m+1} и Y_j для любых допустимых m, j. Именно, имеет место

Лемма. Пусть $\beta_j \neq 0$, $1 \le j \le N-1$, $\alpha_j \neq 0$, $2 \le j \le N$, и, кроме того,

а) матрицы $\mathcal{R}_q^p$ и $\mathcal{L}_q^p$ определены рекуррентно как

$$\mathcal{R}_{p-2}^p = 0,\ \mathcal{R}_{p-1}^p = I,\ \beta_q \mathcal{R}_q^p = (\gamma_q I + S)\mathcal{R}_{q-1}^p - \alpha_q \mathcal{R}_{q-2}^p,\ q = p, p+1, \ldots; \quad (15а)$$

$$\mathcal{L}_q^{q+2} = 0,\ \mathcal{L}_q^{q+1} = I,\ \alpha_p \mathcal{L}_q^p = (\gamma_p I + S)\mathcal{L}_q^{p+1} - \beta_p \mathcal{L}_q^{p+2},\ p = q, q-1, \ldots; \quad (15б)$$

б) векторы P_j^r, $r \ge 2$, и Q_ℓ^j, $\ell \le N-1$, определены рекуррентно как

$$P_{r-1}^r = 0,\ \beta_r P_r^r = F_r,\ \beta_j P_j^r = (\gamma_j I + S)P_{j-1}^r - \alpha_j P_{j-2}^r + F_j,\ j = r+1, r+2, \ldots \quad (16а)$$

$$Q_\ell^{\ell+1} = 0,\ \alpha_\ell Q_\ell^\ell = F_\ell,\ \alpha_j Q_\ell^j = (\gamma_j I + S)Q_\ell^{j+1} - \beta_j Q_\ell^{j+2} + F_j,\ j = \ell-1, \ell-2, \ldots, \quad (16б)$$

г) имеют место уравнения (14а).

Тогда справедливы равенства

$$-\frac{\beta_r}{\alpha_r}\mathcal{R}_j^{r+1} Y_{r-1} + \mathcal{R}_j^r Y_r - Y_{j+1} = P_j^r,\ r-1 \le j \le N-1; \quad (17а)$$

$$-Y_{j-1} + \mathcal{L}_\ell^j Y_\ell - \frac{\beta_\ell}{\alpha_\ell}\mathcal{L}_{\ell-1}^j Y_{\ell+1} = Q_\ell^j,\ 2 \le j \le \ell+1. \quad (17б)$$

Доказательство леммы проводится так же, как и для аналогичного утверждения в [6].

Используя сформулированный результат, построим редуцированную систему, связывающую отстоящие друг от друга пары соседних неизвестных векторов: $\ldots, Y_{j'-1}, Y_{j'}, Y_{j''-1}, Y_{j''}, \ldots$. Введем последовательность целых чисел m_t, $0 \le t \le 2L$, таких, что

$$1 = m_0 < m_1 < m_2 < \ldots < m_{2L} = N+1,$$

и положим

$$\Phi_{j-1} = P_{q-1}^p - Q_{r-1}^{q+1},\quad \Phi_j = -P_{q-2}^p + Q_{r-1}^q, \quad (18)$$

$$p = m_{j-2},\ q = m_{j-1},\ r = m_j,\ j = 2, 4, 6, \ldots, 2L.$$

Из (I7) получаем, что Φ_j суть правые части искомой редуцированной системы, которую мы запишем в виде

$$\begin{pmatrix} -\frac{\beta_p}{\alpha_p}\mathcal{R}_{q-1}^{p+1} & \mathcal{R}_{q-1}^{p} & -\mathcal{L}_{r-1}^{q+1} & \frac{\alpha_{r-1}}{\beta_{r-1}}\mathcal{L}_{r-2}^{q+1} \\ \frac{\beta_p}{\alpha_p}\mathcal{R}_{q-2}^{p+1} & -\mathcal{R}_{q-2}^{p} & \mathcal{L}_{r-1}^{q} & -\frac{\alpha_{r-1}}{\beta_{r-1}}\mathcal{L}_{r-2}^{q} \end{pmatrix} \begin{pmatrix} Y_{p-1} \\ Y_p \\ Y_{r-1} \\ Y_r \end{pmatrix} = \begin{pmatrix} \Phi_{j-1} \\ \Phi_j \end{pmatrix}, \tag{I9}$$

$p = m_{j-2}$, $q = m_{j-1}$, $r = m_j$;

$j = 2, 4, 6, \ldots, 2L$, $\quad Y_0 = Y_N$, $Y_{N+1} = Y_1$.

Прежде чем излагать метод решения системы (I9), заметим, что если она решена, то векторы Y_j, не входящие в (I9), определяются из рекуррентных соотношений

$$\begin{aligned} &\beta_j Y_{j+1} = (\gamma_j I + S) Y_j - \alpha_j Y_{j-1} - F_j, \quad j = p, p+1, \ldots, q-2; \\ &\alpha_j Y_{j-1} = (\gamma_j I + S) Y_j - \beta_j Y_{j+1} - F_j, \quad j = r-1, r-2, \ldots, q+1; \\ &p = m_{t-2}, \quad q = m_{t-1}, \quad r = m_t, \quad t = 2, 4, 6, \ldots, 2L. \end{aligned} \tag{20}$$

Таким образом, маршевый метод заключается в выполнении трех этапов:

I) вычисление по формулам (I6) векторов P_{q-1}^{p} и Q_{r-1}^{q+1} при соответствующих p, q, r и получение правых частей системы (I9) по формулам (I8);

2) решение системы (I9);

3) вычисление остальных векторов Y_j согласно (20).

Поскольку первый и третий этапы в нашем случае требуют, очевидно, $O(MN)$ операций, число операций маршевого метода составит $O(MN + \gamma_2)$, где γ_2 - затраты на решение системы (I9) Оценим последнюю величину.

Предполагая, что S имеет вид (2), подставим соответствующее разложение $S = Q \Lambda Q^{-1}$ в (15). Мы получим, что

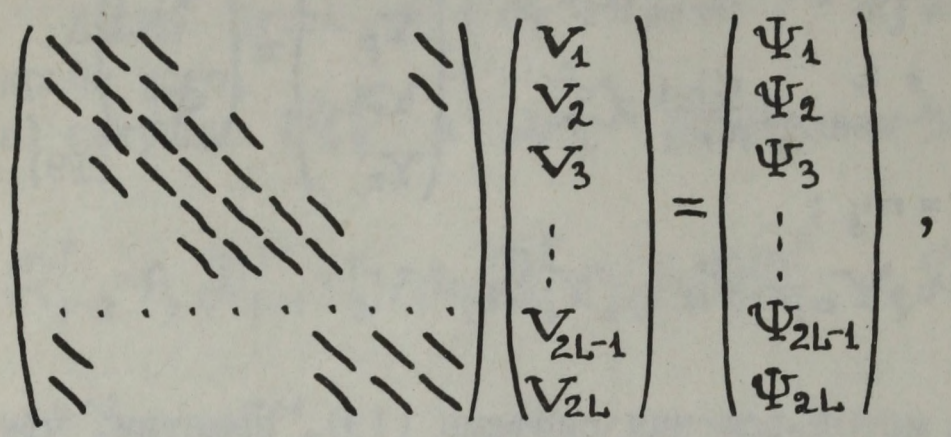
$$\mathcal{R}_q^p = Q \Lambda_q^p Q^{-1}, \quad \mathcal{L}_q^p = Q M_q^p Q^{-1}, \tag{2I}$$

где Λ_q^p, M_q^p - диагональные матрицы, которые можно вычислить по формулам

$$\Lambda_{p-2}^{p} = 0, \ \Lambda_{p-1}^{p} = I, \ \beta_q \Lambda_q^p = (\gamma_q I + \Lambda)\Lambda_{q-1}^{p} - \alpha_q \Lambda_{q-2}^{p}, \ q = p, p+1, \ldots \tag{22}$$

$$M_q^{q+2} = 0, \ M_q^{q+1} = I, \ \alpha_p M_q^p = (\gamma_p I + \Lambda) M_q^{p+1} - \beta_p M_q^{p+2}, \ p = q, q-1, \ldots.$$

Подставляя (2I) в систему (I9) и домножая уравнения на Q^{-1}, получаем систему с матрицей, составленной из диагональных блоков (элементы которых вычисляются по формулам (22)):



$$\begin{pmatrix} & & & & \\ & & & & \\ & & & & \\ & & & & \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ & & & & \\ & & & & \end{pmatrix} \begin{pmatrix} V_1 \\ V_2 \\ V_3 \\ \vdots \\ V_{2L-1} \\ V_{2L} \end{pmatrix} = \begin{pmatrix} \Psi_1 \\ \Psi_2 \\ \Psi_3 \\ \vdots \\ \Psi_{2L-1} \\ \Psi_{2L} \end{pmatrix}, \tag{23}$$

где

$$QV_{2j-1} = Y_{m_{j-1}}, \quad QV_{2j} = Y_{m_j-1}, \quad 1 \le j \le L, \tag{24}$$

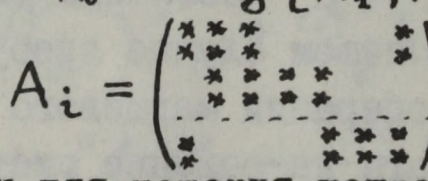
$$\Psi_j = Q^{-1}\Phi_j, \quad 1 \le j \le 2L. \tag{25}$$

Таким образом, метод решения редуцированной системы сводится к выполнению следующих четырех этапов:

I) преобразование правых частей по формулам (25);

2) вычисление коэффициентов матрицы системы (23) с использованием формул (22);

3) решение системы (23); заметим, что соответствующая перестановка строк и столоцов приводит ее матрицу к блочно-диагональному виду $diag\{A_1, \ldots, A_M\}$ с $2L \times 2L$ – блоками вида



$$A_i = \begin{pmatrix} & & & & \\ & & & & \\ & & & & \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ & & & & \end{pmatrix},$$

и поэтому для решения каждой системы $A_i x_i = b_i$ достаточно $O(L)$ операций (применяется метод исключения Гаусса, учитывающий структуру разреженности A_i);

4) вычисление искомых векторов Y_j по формулам (24).

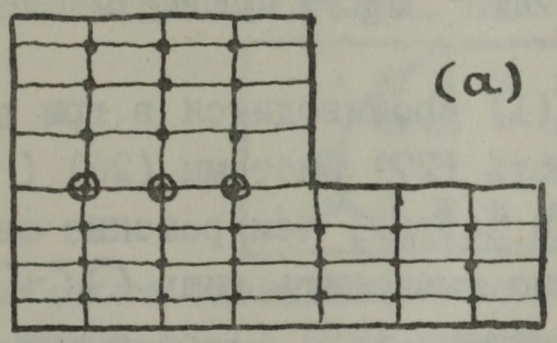
В силу п°1.1 общее число операций на решение задачи (I9) составит $O(LM\log M)+O(MN)+O(ML)+O(LM\log M)$ операций, т.е. $\mathcal{N}_2 = O(MN + LM\log M)$.Отсюда, учитывая предыдущие оценки, получаем асимптотику числа операций маршевого метода вида $O(MN + LM\log M)$.

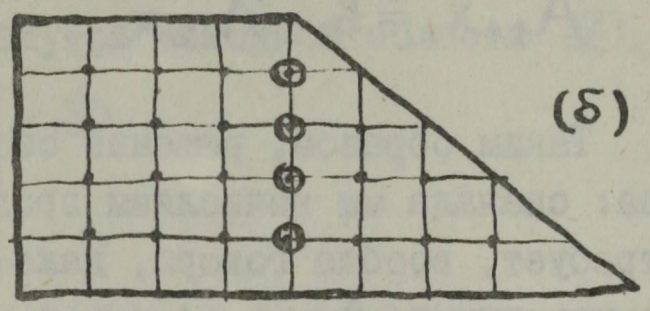
Для рассматриваемого класса задач характерна ситуация, когда матрицы T_1 и T_2 в задаче (3) положительно определены. В этом случае большие длины маршей $m_j - m_{j-1}$, желательные с точки

зрения уменьшения L (т.е. ускорения метода) недопустимы, так как приводят к численной неустойчивости метода. Учитывая результаты исследований устойчивости, проведенных в [7], будем считать, что при $M=O(N)$ выполнено $L=O(N/\log N)$ (т.е. $m_j - m_{j-1} = O(\log N)$). Тогда мы получаем численно устойчивый маршевый метод с асимптотикой числа операций $O(N^2)$ при $M=O(N)$.

2. Прямые и итерационные методы, основанные на разрезании сеточной области.

На практике часто встречаются краевые задачи, поставленные в областях, составленных из прямоугольников либо других областей, где может использоваться сетка, эквивалентная прямоугольной:





Возникающие при этом системы N линейных уравнений характеризуются тем, что нахождение $O(\sqrt{N})$ значений неизвестной сеточной функции в узлах на границе раздела, очевидно, сводит задачу к нескольким задачам, рассмотренным выше (в случае (а)) При этом построение системы, связывающей эти неизвестные, и ее решение также производится с использованием указанных алгоритмов.

Проиллюстрируем сказанное на примере задачи (а). В этом случае матрица системы имеет вид

$$A = \begin{pmatrix} S' & -I' & & & & & & \\ -I' & S' & -I' & & & & & \\ & -I' & S' & -I' & & & & \\ & & -I' & S' & -I' & 0 & & \\ & & & -I' & S'' & -I'' & & \\ & & & 0 & -I'' & S'' & -I'' & \\ & & & & & -I'' & S'' & \end{pmatrix}.$$

Вводя более крупное блочное разбиение, систему $Ax=b$ запишем в виде

$$\begin{pmatrix} A_{11} & A_{12} & 0 \\ A_{21} & A_{22} & A_{23} \\ 0 & A_{32} & A_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}, \quad A_{22} = S'. \tag{26}$$

Исключая из второго уравнения x_1 и x_3, имеем

$$\hat{b}_2 = b_2 - A_{21}A_{11}^{-1}b_1 - A_{23}A_{33}^{-1}b_3, \tag{27}$$

$$Cx_2 \equiv (A_{22} - A_{21}A_{11}^{-1}A_{12} - A_{23}A_{33}^{-1}A_{32})x_2 = \hat{b}_2, \tag{28}$$

$$A_{11}x_1 = b_1 - A_{12}x_2, \quad A_{33}x_3 = b_3 - A_{32}x_2. \tag{29}$$

Таким образом, решение системы (I) производится в три этапа: сначала мы вычисляем правую часть (27) системы (28) (это требует, вообще говоря, даже меньше затрат, чем решение систем с матрицами $A_{\alpha\alpha}$, $\alpha=1,3$, т.к. нужно вычислить лишь $O(\sqrt{N})$ компонент решения), затем решаем систему (28) (это самый сложный этап) и , наконец, решаем две системы (29).

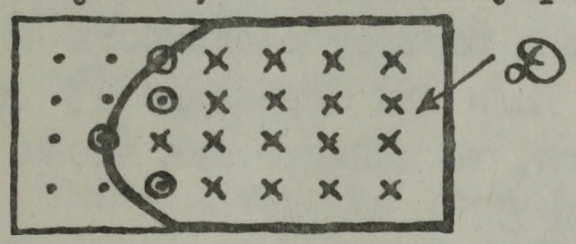
Систему (28) можно решать либо прямыми, либо итерационными методами. В первом случае вычисляются те $O(\sqrt{N})$ компонент векторов $A_{\alpha\alpha}^{-1}a_{\alpha 2(j)}$ (где $a_{\alpha 2(j)}$ – j-й столбец $A_{\alpha 2}$, $\alpha=1,3$), которые необходимы для формирования произведения матриц $A_{2\alpha} \times A_{\alpha\alpha}^{-1}A_{\alpha 2}$. Если матрицы S', S'' имеют вид (2), то для этого достаточно $O(N\log N)$ операций; в противном случае – $O(N\sqrt{N})$ операций. Далее, полученная матрица емкости C разлагается в произведение $LD^{-1}L^T$ по алгоритму симметричного исключения Гаусса с затратой $O(N\sqrt{N})$ операций. Этот вариант алгоритма ориентирован на решение задачи (26) со многими правыми частями, так как после вычисления разложения $C = LD^{-1}L^T$ для решения системы (26) достаточно не более $O(N\log N)$ операций ($O(N)$, если используется маршевый метод).

Другой вариант метода использует итерационный метод решения системы (28) и не требует явного вычисления C. Так, нетрудно убедиться в том, что для умножения матрицы C на вектор достаточно $O(\sqrt{N}\log N)$ операций, если матрицы S', S'' имеют вид (2) (если нет, то $O(N)$ операций). Поэтому явный метод сопряжённых градиентов для (28) имеет число операций $O(N\log N)$,

т.к. должен дать точное решение за $O(\sqrt{N})$ итераций. Если же построить неявный итерационный процесс (31) с регуляризатором $B = \frac{1}{2} A_{22} - A_{21} A_{11}^{-1} A_{12} \equiv R(S')$ где R -рациональная функция степени $O(\sqrt{N})$, то скорость его сходимости не будет зависеть от N [8] и решение системы (28) можно будет вычислить с точностью ε с затратами от $O(\sqrt{N} \log N \log \varepsilon^{-1})$ до $O(N \log \varepsilon^{-1})$ операций (в зависимости от свойств S').

3. <u>Прямые и итерационные методы, основанные на дополнении сеточной области</u>.

В тех случаях, когда область $\mathcal{D}$, где поставлена дифференциальная краевая задача, имеет сложную геометрическую конфигурацию, целесообразно ее не разрезать, а дополнять до объемлющего прямоугольника, вводить в нем сетку и строить разностную аппроксимацию на узлах, попавших внутрь исходной области $\mathcal{D}$.



После этого оказывается возможным дополнить полученные уравнения аналогичными, записанными в остальных узлах прямоугольника, что приводит к системе

$$A x = b. \qquad (30)$$

При этом нетрудно добиться того, чтобы 1) решение построенной задачи в прямоугольнике совпадало с искомым решением в узлах исходной области; 2) уравнение полученной системы отличались от уравнений стандартной пятиточечной задачи $Bw = r$ в прямоугольнике лишь в $O(\sqrt{N})$ узлах на границе между областью и ее дополнением. Кроме того, зачастую удается обеспечить "хорошие" свойства матрицы $B^{-1}A$, что позволяет построить экономичный итерационный метод для задачи (30) со скоростью сходимости, не зависящей от N.

Таким образом, исходная задача сводится к задаче с матрицей

$$A = B + p q^T, \quad nz(p) = O(\sqrt{N}), \quad nz(q) = O(\sqrt{N}),$$

т.е. A отличается от B лишь на слагаемое малого ранга и высокой разреженности. Это обстоятельство позволяет, как и выше, применить метод матрицы емкости, основанный на формуле для модифицированной обратной:

$$A^{-1} = B^{-1}\left(I - p\left(\tilde{I} - q^T B^{-1} p\right)^{-1} q^T B^{-1}\right).$$

Здесь размер матрицы емкости

$$C = \tilde{I} - q^T B^{-1} p$$

также асимптотически равен $O(\sqrt{N})$, но реально может достигать больших величин. Это замедляет вычисление и факторизацию C и затрудняет хранение полученного разложения. Как в случае разрезания, число операций прямого метода матрицы емкости составляет $O(N)$ на каждую из задач в серии (если использовать маршевой метод).

Для одноразового решения задачи целесообразно применять неявные итерационные методы с регуляризатором B. В некоторых важных случаях (третья краевая задача (9), первая краевая задача (10)) удается найти такой способ построения матрицы A, который позволяет добиться быстрой сходимости соответствующего итерационного процесса. Если для обращения B на векторе использовать маршевый метод, то асимптотика числа операций таких методов есть $O(N \log \varepsilon^{-1})$.

4. Неявные итерационные методы.

При решении разностных краевых задач для общих эллиптических уравнений (трактуемых как операторное уравнение $Au = f$ в гильбертовом пространстве H) применяются, в основном, неявные итерационные методы. Широкое использование в этих методах, базирующихся на двухслойной

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + A y_k = f, \quad k = 0, 1, \ldots, n-1, \quad y_0 \in H, \qquad (31)$$

(или трехслойной) итерационной схеме, нашли полные и факторизованные операторы верхнего слоя B.

В качестве полных применяются операторы, соответствующие разностным краевым задачам с разделяющимися переменными. Такие операторы могут быть получены, например, дополнением исходной сеточной задачи до расширенной в объемлющей регулярной области и усреднением коэффициентов расширенной задачи по одной пространственной переменной. Применение экономичных прямых алгоритмов (см. п°1) позволяет реализовать каждую итерацию за $O(MN)$ или $O(MN \log\log N)$ операций. В случае краевых задач для уравнения вида

$$\sum_{\alpha=1}^{2} \frac{\partial}{\partial x_\alpha} \left(k_\alpha(x) \frac{\partial u}{\partial x_\alpha} \right) = -f(x), \qquad x = (x_1, x_2) \in G,$$

число итераций $n = n(\varepsilon) = 0.5\sqrt{c_0}\,\ln(2/\varepsilon)$ данного метода не зависит от шага сетки h. Однако, в силу зависомости от $c_0 = \max_{\alpha}\left(\max_{x\in G} k_\alpha(x) / \min_{x\in G} k_\alpha(x)\right)$, итерационные методы с полным оператором верхнего слоя эффективно применимы лишь к решению разностных уравнений, коэффициенты которых меняются слабо.

К настоящему времени разработано несколько подходов к построению факторизованных операторов B для схемы (3I).

Первый подход связан с использованием специальной структуры матрицы оператора A. На основе этого подхода построены методы переменных исправлений - ADI и верхней релаксации - SOR, для которых проблема выбора итерационных параметров решена лишь при весьма жестких ограничениях на свойства матрицы A [I]. Применение метода SOR требует достаточно точного и экономичного вычисления оптимального параметра релаксации. Способ оценки параметра релаксации, предложенный в [II], порождает зависи - мость числа итераций SOR от экстремальных характеристик коэффициентов разностных уравнений. Указанная зависимость устраняется применением метода [I], позволяющего вычислить априорную информацию для метода SOR за $O(MN)$ операций.

Наряду с ограниченной областью эффективного применения рассмотренные методы обладают достаточно низкой скоростью сходи - мости с числом итераций $O(h^{-1}\log\varepsilon^{-1})$.

Конструктивный подход к построению факторизованного оператора верхнего слоя лежит в основе попеременно- треугольного метода А.А. Самарского [I2], который применим для решения уравнения с самосопряженным и положительно определенным оператором A. В итерационных схемах используется оператор

$$B = (\mathcal{D} + \omega R_1)\mathcal{D}^{-1}(\mathcal{D} + \omega R_2), \qquad (32)$$

построенный по разложению оператора-регуляризатора R, $c_1 R \le A \le c_2 R$, в сумму сопряженных друг к другу операторов R_1 и R_2, $R_1 = R_2^*$. Здесь $\mathcal{D} = \mathcal{D}^* > 0$ - операторный параметр, выбираемый из условия максимума $\eta = \delta/\Delta$ где δ и Δ - числа (априорная информация) из операторных неравенств $\delta\mathcal{D} \le R$, $\delta > 0$, $R_1\mathcal{D}^{-1}R_2 \le \frac{\Delta}{4}R$.

Попеременно-треугольный метод (ПТМ) показал высокую эффективность при решении разностных краевых задач для общих (самосопряженных) эллиптических уравнений 2-го порядка с сильноменяющимися разрывными коэффициенты, заданных как в прямоугольнике,

так и в области сложной формы [I3]. Для числа итераций ПТМ справедлива оценка $n(\varepsilon)=O(h^{-\gamma}\ln\frac{1}{\varepsilon})$ с $\gamma=0.5$, причем число итераций определяется интегральными характеристиками коэффициентов, параметра эллиптичности и шагов сетки. В частности, наличие разрывов коэффициентов не приводит к увеличению показателя γ, форма области не влияет на число итераций.

Сравнение ПТМ с рядом других итерационных методов свидетельствует о его преимуществе. Отметим, что для метода симметричной верхней релаксации SSOR $0.5\le\gamma\le0.75$ [I4], причем значение $\gamma=0.5$ достигается лишь в случае гладких коэффициентов; для метода неполного разложения Холецкого ICCG(0)[I5] $\gamma=1$; для варианта метода неполной факторизации [I6] $\gamma=0.5$.

Разработанные точечные и блочные варианты ПТМ позволяют решать разностные смешанные краевые задачи для эллиптических уравнений, заданных в криволинейных ортогональных системах координат [I7], системы линейных алгебраических уравнений с матрицей Стилтьеса [I8], к которым приводятся разностные уравнения на треугольных сетках и девятиточечные разностные схемы, полученные частичным исключением неизвестных в пятиточечных схемах. ПТМ с выбором регуляризатора позволяет решать разностные краевые задачи для эллиптического уравнения со смешанными производными для систем уравнений теории упругости [I] и др. Факторизованный оператор ПТМ (32) нашел эффективное применение в итерационных схемах для решения задач на собственные значения [I9].

Другим конструктивным подходом к построению оператора верхнего слоя является неполное разложение матрицы оператора A. Методы неполного разложения - МНР, основанные на данном подходе находят широкое применение при решении разностных уравнений с матрицей сложной структуры, в частности, разреженных систем линейных уравнений, возникающих в схемах метода конечных элементов.

В предложении, что выбор ведущего элемента не требуется, полное разложение $N\times N$-матрицы $A=LDU$ на треугольные множители L и U определяется следующим образом:

$$L=\prod_{n=1}^{N}\begin{pmatrix} I_{n-1} & 0 \\ 0 & {}^{n}L \end{pmatrix},\quad U=\prod_{n=1}^{N}\begin{pmatrix} I_{n-1} & 0 \\ 0 & {}^{n}U \end{pmatrix},\quad D=\begin{pmatrix} {}^{1}a_{11} & & \\ & \ddots & \\ & & {}^{N}a_{NN} \end{pmatrix},$$

где $$ {}^{n}A := \begin{pmatrix} {}^{n}a_{nn} & {}^{n}q^{T} \\ {}^{n}p & {}^{n}a \end{pmatrix} = \begin{pmatrix} {}^{n}a_{nn} & 0 \\ {}^{n}p & I_{N-n} \end{pmatrix} \begin{pmatrix} {}^{n}a_{nn}^{-1} & 0 \\ 0 & {}^{n-1}A \end{pmatrix} \begin{pmatrix} {}^{n}a_{nn} & {}^{n}q^{T} \\ 0 & I_{N-n} \end{pmatrix}, $$

$$ {}^{n-1}A = {}^{n}a - \frac{1}{{}^{n}a_{nn}} {}^{n}p \cdot {}^{n}q^{T}, \quad n = 1, 2, \ldots, N, \quad {}^{0}A := A. $$

Видно, что матрицы L и U являются, вообще говоря, заполненными, что и сдерживает применение прямых методов для решения разностных уравнений с большим числом неизвестных в плоском случае, и тем более в трехмерном случае.

В неполном разложении посредством усечения относительно малых элементов матриц ${}^{n}L$ и ${}^{n}U$ сохраняется разряженность множителей L и U в операторе $B = LDU$. Возможны два способа усечения элементов: а) усечение после разложения; б) усечение при разложении.

В каждом из этих способов возможны следующие стратегии усечения:

I) по индексу усекаемого элемента, когда задано множество $\mathcal{P}(A)$ индексов элементов, не подлежащих отбрасыванию;

2) по абсолютной величине усекаемого элемента, когда задан параметр усечения c такой, что отбрасываются все элементы ${}^{n}a_{ij}$, удовлетворяющие неравенству

$$ ({}^{n}a_{ij})^{2} < c \, {}^{n}a_{ii} \, {}^{n}a_{jj}, \quad i \neq j. $$

Не имея возможности останавливаться подробно на достоинствах и недостатках вариантов неполного разложения, укажем лишь основные.

Существенным недостатком первого способа усечения является лишь необходимость предварительного полного разложения, требующего больших объемов вычислений и памяти. С другой стороны, такой способ усечения всегда приводит к невырожденному оператору B, достаточно близкому к исходному оператору A в случае разумного объема усечения. Данное неполное разложение можеть быть с успехом применено при решении задач эволюционного типа с постоянным оператором временного слоя. МНР с усечением после разложения, а также ряд интересных приложений этого метода рассмотрен в [20].

В разложении с усечением требуется память меньшего объема, фиксированного в случае усечения по индексу и заранее не извес-

тного при усечении по значению. Последний случай требует, вообще говоря, разработки специальных стратегий изменения параметра усечения при исчерпании доступной памяти [21], Существенным недостатком данного способа усечения является возможное вырождение процесса разложения или потеря знакоопределенности оператора B, если A знакоопределен [22]. Для класса $\mathcal{H}$ - матриц, включающего важные случаи разностных уравнений с диагональным преобладанием, устойчивость и невырожденность неполного разложения доказана в[23]. В иных ситуациях предварительный сдвиг матрицы A - приведение к $\mathcal{H}$ - матрице или переопределение диагональных элементов матриц ${}^{n}L$ и ${}^{n}U$ позволяет избежать вырождения процесса неполного разложения [23,24].

В заключение отметим, что для методов, использующих факторизованный оператор на верхнем слое, разработаны эффективные алгоритмы, в которых существенное сокращение вычислительной работы получено за счет более простой, чем традиционная, реализации каждой итерации [25,26].

ЛИТЕРАТУРА

I. Самарский А.А., Николаев Е.С. Методы решения сеточных уравнений. - М.: Наука, I978.

2. Хокни Р. Методы расчета потенциала и их приложения. - В сб. Вычислительные методы в физике плазмы. - М.: Мир, I974, I43-2I2.

3. Ахо А., Хонкрофт Дж., Ульман Дж. Построение и анализ вычислительных алгоритмов. - М.: Мир, I979.

4. Sweet R. A cyclic reduction algorithm for solving block tridiagonal systems of arbitrary dimension. -SIAM J. Numer. Anal., 1977, 14, N4, 706-710.

5. Swarztrauber P.N. A direct method for the discrete solution of separable elliptic equations. - SIAM J. Numer. Anal., 1974, 11, N6, 1136-1150.

6. Капорин И.Е. Модифицированный марш-алгоритм решения разностной задачи Дирихле для уравнения Пуассона в прямоугольнике. - В сб. Разностные методы математической физики. - М.: Изд-во Моск. ун-та, I980, II-2I.

7. Bank R.E., Rose D.J. Marching algorithms for elliptic boundary value problems. I: The constant coefficient case. – SIAM J. Numer. Anal., 1977, v.14, 792–829.
8. Мацокин А.М. Об одном методе решения систем сеточных уравнений. – В сб. Методы решения систем вариационно-разностных уравнений. – ВЦ СО АН СССР, Новосибирск, 1979, вып. 5, 136–138.
9. Астраханцев Г.П. Метод фиктивных областей для эллиптического уравнения второго порядка с естественными граничными условиями. – ЖВМ и МФ, 1978, т. 18, №1, 118–125.
10. Капорин И.Е., Николаев Е.С. Метод фиктивных неизвестных для решения разностных эллиптических краевых задач в нерегулярных областях. – Дифференц. урвнения , 1980, т. 16, №7, 1211–1225.
11. Young D.M. On the accelerated SSOR method for solving large linear systems. – Advances in Math., 1977, v.23, N3, 215–271.
12. Самарский А.А. Об одном экономичном алгоритме численного решения систем дифференциальных и алгебраических уравнений. – ЖВМ и МФ, 1964, т. 4, №3, 580–585.
13. Кучеров А.Б., Николаев Е.С. Попеременно-треугольный итерационный метод решения сеточных эллиптических уравнений в произвольной области. – ЖВМ и МФ, 1977, т. 17, №3, 664–675.
14. Wang H.H. The application of the symmetric SOR and the symmetric SIP methods for the numerical solution of the neutron diffusion equation. – Nuclear Sci. and Eng., 1978, 67, 162–171.
15. Некоторые неявные итерационные методы. Анализ и сравнение/ Богданова М.С., Кучеров А.Б., Николаев Е.С. и др. Препринт №115. – М.: ИПМ АН СССР, 1978.
16. Dupont T., Kendall R.P., Rachford H.H. An approximate factorization procedure for solution self-adjoint elliptic difference equations. – SIAM J. Numer. Anal., 1968, 5, N3, 559–573.
17. Кучеров А.Б. Попеременно-треугольный итерационный метод решения разностных уравнений: Автореферат дисс. на соиск. уч. ст. канд. физ.-мат. наук. – М.: Изд-во Моск. ун-та, 1979.
18. Кучеров А.Б. Попеременно-треугольный итерационный метод для решения линейных систем с ленточной S -матрицей. – В сб. Численные методы линейной алгебры. – М.: Изд-во Моск. ун-та, 1982, 94–102.
19. Приказчиков В.Г. Прототипы итерационных процессов в задаче на

собственные значения. - Дифференц. уравнения, I980, т. I6, №9, I688-I697.

20. Glowinski R., Periaux J., Pironneau O. An efficient preconditioning scheme for iterative numerical solution of partial differential equations. — Appl. Math. Modell., 1980, 4, N3, 187-192.
21. Munksgaard N. Solving sparse symmetric sets of linear equations by preconditioned conjugate gradients. — ACM TOMS, 1980, v.6, N2, 206-219.
22. Kershaw D. An incomplete Cholesky conjugate gradient method for iterative solution of systems of linear equations. - J.Comput. Phys., 1978, 26, N1, 43-65.
23. Manteuffel T.A. An incomplete factorization technique for positive definite linear systems. — Math. Comp. 1980, 34, N150, 473-497.
24. Kershaw D. On the problem of unstable pivots in the incomplete LU-conjugate gradient method. - J.Comput. Phys., 1980, 38, N1, 114-123.
25. Кучеров А.Б. Алгоритмы итерационных схем с факторизованным оператором. - В сб. Разностные методы математической физики. - М.: Изд-во Моск. ун-та, I98I, 23-30.
26. Eisenstat S.C. Efficient implementation of a class of preconditioned conjugate gradient methods. — SIAM J. Sci. Stat. Comp., 1981, v.2, N1, 1-4.

МЕТОДЫ ПОВЫШЕНИЯ ТОЧНОСТИ ПРИБЛИЖЕННЫХ РЕШЕНИЙ ЗАДАЧ МАТЕМАТИЧЕСКОЙ ФИЗИКИ

Е.П. ЖИДКОВ, Б.Н. ХОРОМСКИЙ

Объединенный институт ядерных исследований
Лаборатория вычислительной техники и информации

АННОТАЦИЯ

Рассматриваются вопросы применения экстраполяции Ричардсона, и ускорения сходимости итерационных процессов на последовательности сеток. Практическая эффективность приведенных алгоритмов для ряда физических задач иллюстрируется численными расчетами.

ABSTRACT

Some aspects of application of Richardson's extrapolation method and of increasing the convergence rate of iteration processes on the sequence of grids are considered. Numerical calculations show the practical efficiency of given algorithms for solution of some physical problems.

Пусть для решения операторного уравнения построена разностная схема, для которой погрешность приближенных решений разлагается по степеням шага дискретизации. Такое разложение можно использовать для уточнения приближенных решений (экстраполяция Ричардсона), для оценки погрешности через известные величины (правило Рунге), для ускорения сходимости итерационных процессов с использованием последовательности сеток.

В обзоре рассмотрены общие условия для операторных уравнений, гарантирующие разложение погрешности по степеням шага дискретизации (§ I). Как следствие общей теоремы получены разложения для уравнений типа Чу-Лоу, одного класса задач на собственные значения, для краевых задач, связанных с оператором Лапласа. Проиллюстрирована эффективность экстраполяции Ричардсона для этих задач (§ 2). Приведен специальный метод экстраполяции для операторов, инвариантных относительно поворота системы координат (§ 3). Рассмотрены методы ускорения сходимости итерационных процессов на последовательности сеток с учетом регулярного поведения погрешности относительно шага дискретизации (§ 4).

Вопросам, связанным с методом Ричардсона и оптимизацией вычислений на последовательности сеток, посвящено большое количество научных публикаций. Среди них отметим монографию /I/ и цитируемые там работы, а также /2-I0/. Рассмотренные здесь результаты, в основном, содержатся в работах /II-20/.

§ I. Асимптотическое разложение погрешности

Рассмотрим вопрос о приближенном решении уравнения

(I.I) $Au = y$, $u \in X$, $y \in Y$,

где нелинейный оператор A действует из B-пространства X в B-пространство Y. Пусть задана последовательность операторов P_n и Q_n, проектирующих X и Y на n-мерные подпространства X_n и Y_n, а уравнению (I.I) соответствует следующее уравнение для приближенного решения

(I.2) $A_n u_n = Q_n y$; $A_n : X_n \to Y_n$.

Оператор A_n аппроксимирует A. Обозначим через u^* и u_n^* решения уравнений (I.I),(I.2) соответственно. Пусть в пространствах X и Y заданы линейные множества $B_k \subset X$ и $B_k' \subset Y$, $k = 0, 1, \dots, N$, так что $B_0 \subset B_1 \subset \dots \subset B_N = X$, $B_0' \subset \dots \subset B_N' = Y$. Эти множества определяются классами разрешимости задачи (I.I): если для $y \in B_k'$ существует решение u^* уравнения (I.I), то $u^* \in B_k$, и, кроме того, $A(B_k) \subset B_k'$. Аналогичные соотношения справедливы для производных Фреше порядка l ($l = 1, \dots, N$) оператора A: если $v_{k_i} \in B_{k_i}$, $i = 1, \dots, l$, то

$$A^{(l)}(u^*)(v_{k_1}, \dots, v_{k_l}) \in B'_{k_m}, \quad k_m = \max_{1 \le i \le l} k_i .$$

Наша цель – получить разложение величины

(I.3) $\Delta_n = P_n u^* - u_n^*$

по степеням n^{-1} :

(I.4) $$\Delta_n = P_n \left(\sum_{k=1}^{N} c_k n^{-k} \right) + \Omega_n , \quad \|\Omega_n\| = o(n^{-N}),$$

где элементы $c_k \in X$ зависят только от u^* и не зависят от n. При этом разложение (I.4) связано с последовательностью множеств $B_k \subset X$:

$$c_k \in B_k \subset \dots \subset B_N = X , \quad k = 1, \dots, N ; \quad u^* \in B_0 .$$

Относительно операторов A и A_n предположим следующие свойства:

(A) Уравнения (I.I) и (I.2) имеют решения $u^* \in B_0 \subset X$, $u_n^* \in X_n$, и выполнено соотношение

(I.5) $$\lim_{n\to\infty} \| P_n u^* - u_n^* \| = 0 .$$

(B) Операторы A и A_n имеют равномерно непрерывные производные Фреше до порядка N в окрестностях $S_R = \{ x : \| x - u^* \| \le R \}$ и $S_R^n = \{ x_n : \| x_n - P_n u^* \| \le R \}$ соответственно, так что $\| A^{(k)}(u^*) \| \le M$, $\| A_n^{(k)}(P_n u^*) \| \le M$, $k \le N$, и для операторов $A'(u^*)$, $A_n'(P_n u^*)$ существуют ограниченные обратные

$$\| A'(u^*)^{-1} \| \le M_1 , \quad \| A_n'(P_n u^*)^{-1} \| \le M_1 ,$$

и из равенства

$$A'(u^*) v = g , \quad g \in B_k'$$

следует, что $v \in B_k$.

(C) Справедливо разложение

(I.6) $$Q_n A u^* = A_n P_n u^* + Q_n \left(\sum_{k=1}^{N} a_k(u^*) n^{-k} \right) + \Omega_{n,0} ,$$

где элементы $a_k \in B_k'$, $k = 1, \dots, N$ определяются через u^* и не зависят от n . Величина $\Omega_{n,0}$ удовлетворяет соотношению

$$\| \Omega_{n,0} \| = o(n^{-N}) , \quad n \to \infty .$$

(D) Для производных $A_n^{(\ell)}(P_n u^*)$ и $A^{(\ell)}(u^*)$, $\ell = 1, \dots, N$ при заданных $v_{k_1}, \dots, v_{k_\ell}$; $v_{k_i} \in B_{k_i}$, $i = 1, \dots, N$ справедливо представление

(I.7) $$Q_n A^{(\ell)}(u^*)(v_{k_1}, \dots, v_{k_\ell}) - A_n^{(\ell)}(P_n u^*)(P_n v_{k_1}, \dots, P_n v_{k_\ell}) =$$

$$= Q_n \left(\sum_{|\alpha|=1}^{\beta_\ell} A_{\alpha,\ell}(v_{k_1}, \dots, v_{k_\ell}) n^{-|\alpha|} \right) + \Omega_{n,\alpha_\ell}(v_{k_1}, \dots, v_{k_\ell}),$$

$$\| \Omega_{n,\alpha_\ell} \| = o(n^{\|\alpha_\ell\| - N}) , \quad n \to \infty$$

Здесь α - мультииндекс, т.е. $\alpha = (i_1, \dots, i_\ell)$, $\alpha_\ell = (k_1, \dots, k_\ell)$, $\beta_\ell = (N - k_1, \dots, N - k_\ell)$, а равенство двух индексов рассматривается покомпонентно. При этом $|\alpha| = i_1 + \dots + i_\ell$, $\|\alpha\| = \max\limits_{1 \le p \le \ell} |i_p|$.

Ограниченные полилинейные формы определены следующим образом:

$$A_{\alpha,\ell} : (B_{k_1} \times \dots \times B_{k_\ell}) \to B'_{\|\alpha + \alpha_\ell\|}$$

и не зависят от n. В частности, при $\ell = 1$ и $v \in B_k$, $k = 1, \dots, N$ разложение (I.7) имеет вид:

$$(I.8)\quad Q_n A'(u^*)v - A_n'(P_n u^*)P_n v = Q_n\Big(\sum_{i=1}^{N-k} A_{i,1}(u^*)v\, n^{-i}\Big) + \Omega_{n,1}\,;\quad \|\Omega_{n,1}\| = o(n^{-N+k}),\ n\to\infty,$$

где непрерывные линейные операторы $A_{i,1}$ действуют из B_k в B'_{k+i} и не зависят от n и v.

Разложение вида (I.4) устанавливает

Теорема I. Пусть для уравнений (I.I), (I.2) выполнены условия (A)-($\mathcal{D}$). Тогда для решения u_n^* уравнения (I.2) справедливо разложение

$$(I.9)\quad u_n^* = P_n u^* + P_n\Big(\sum_{k=1}^{N} c_k(u^*)n^{-k}\Big) + \Omega_n,\quad \|\Omega_n\| = o(n^{-N}),\ n\to\infty,$$

где элементы $c_k(u^*) \in B_k \subset X$, $k = 1, \dots, N$ и не зависят от n.

Опуская полное доказательство [II], приведем лишь систему уравнений для коэффициентов $c_k(u^*)$ и Ω_n:

$$(I.10)\quad \begin{aligned} &A'(u^*)c_1 = a_1\\ &A'(u^*)c_2 + A_{1,1}(u^*)c_1 + \tfrac{1}{2}A''(u^*)(c_1,c_2) = a_2\\ &\dots\dots\dots\dots\dots\\ &A'(u^*)c_m + \sum_{i+k=m} A_{i,1}(u^*)c_k + \\ &+\sum_{\ell=2}^{m}\frac{1}{\ell!}\Big\{\sum_{|\alpha_\ell|=m} A^{(\ell)}(u^*)(c_{k_1},\dots,c_{k_\ell}) - \sum_{|\alpha+\alpha_\ell|=m} A_{\alpha,\ell}(c_{k_1},\dots,c_{k_\ell})\Big\} = a_m\\ &m = 1,\dots,N,\\ &A_h'(P_h u^*)\Omega_n + n^{-1}G(c_1,\dots,c_N)\Omega_n + F(\Omega_n) = o(n^{-N}), \end{aligned}$$

где линейный оператор G - ограничен, а $\|F(x)\| = o(\|x\|)$.

Приведем два важных следствия из Теоремы I [II, I4]. Первое из них относится к случаю сильно монотонных операторов. Напомним, что отображение $A: X \to X^*$ называется сильно монотонным [2I], если

(I.II) $$\langle A(x+h)-A(x),\, h\rangle \geq \gamma(\|h\|)\,\|h\|,$$

для любых $x, h \in D(A)$, где $\gamma(t)$ – вещественная неотрицательная функция при $t \geq 0$; $\gamma(t) \to \infty$, $t \to \infty$ и из равенства $\gamma(t)=0$ следует $t=0$. Здесь $\langle y, h\rangle$ – значение линейного функционала $y \in X^*$ на векторе $h \in X$.

Следствие I. Пусть отображения $A : X \to X^*$ и $A_n : X_n \to X_n^* = X_n$ удовлетворяют условию (I.II) при $\gamma(t)=ct$, $c>0$, и N раз равномерно непрерывно дифференцированы по Фреше в окрестностях точек u^* и $P_n u^*$. Если при этом выполнены условия (C) и (D) , то справедливо разложение (I.9).

Доказательство свойства (A) теоремы I следует из общих теорем о монотонных операторах /21/, а (I.5) легко получить из соотношения

$$A_n P_n u^* - A_n u_n^* = Q_n\Big(\sum_{k=1}^{\tilde{N}} a_k(u^*)\,n^{-k}\Big) + \Omega_{n,0}.$$

Условие (B) следует из того, что производная Фреше оператора A при условии (I.II) сильно положительна

$$\langle A'(u^*)h, h\rangle \geq c\,\|h\|^2, \quad \forall h \in X.$$

Следствие I позволяет получить разложения погрешности для квазилинейных эллиптических уравнений, установленные в /5,10/, а также для системы интегральных уравнений типа Чу-Лоу /13/.

Еще одно следствие теоремы I относится к задаче на собственные значения для линейного самосопряженного оператора $A = A^*$, определенного в вещественном гильбертовом пространстве H , $A \in (H \to H)$ /14/. Как известно, эта задача представляет собой нелинейное уравнение

(I.I2) $$\Phi(z) \equiv \begin{Bmatrix} Ax - \lambda x \\ (x,x) - 1 \end{Bmatrix} = 0; \quad z = (x, \lambda) \in H^1 = H + R ,$$

где $\Phi : H^1 \to H^1$. Скалярное произведение в H^1 определяется формулой

$$(z_1, z_2) = (x_1, x_2) + \lambda_1 \lambda_2 .$$

Первая производная Фреше оператора (I.I2) определяется выражением

$$\Phi'(z_0)z = \begin{Bmatrix} Ax - \lambda_0 x - \lambda x_0 \\ (x_0, x) \end{Bmatrix}, \quad z_0 = (x_0, \lambda_0).$$

Производные порядка $\ell > 2$ равны нулю. Пусть задан проектор P_n : $H' \to X_n = H_n + R$, а уравнение (I.I2) заменяется конечномерной системой

(I.I3)
$$\Phi_n(z_n) \equiv \begin{Bmatrix} A_n x_n - \lambda_n x_n \\ (x_n, x_n)_{H_n} - 1 \end{Bmatrix} = 0, \quad A_n \in (H_n \to H_n).$$

Зададим классы разрешимости задачи

$$\Phi'(z^*)z = f; \quad f = (g, \mu), \quad z = (x, \lambda);$$

если $g \in B_k'$, то $x \in B_k$, $k = 1, \dots, N$; $A(B_k) \subset B_k'$.

Следствие 2. Пусть существует однократное изолированное собственное значение λ^* задачи (I.I2) с собственной функцией $x^* \in B_0$, $z^* = (x^*, \lambda^*)$ и выполнено $A_n = A_n^*$. Пусть для некоторого решения $z_n^* = (x_n^*, \lambda_n^*)$ задачи (I.I3) выполнено условие

$$\lim \| P_n z^* - z_n^* \| = 0, \quad n \to \infty,$$

и кроме того, для достаточно больших n

$$\| \Phi_n'(P_n z^*)^{-1} \|_{X_n} \le c,$$

где C не зависит от n. Если при этом существуют разложения:

$$P_n Ax - A_n P_n x = P_n \left(\sum_{k=1}^{N-p} a_k(x) n^{-k} \right) + o(n^{-N+p}),$$

для $x \in B_p$, $a_k \in B_{k+p}'$; a_k не зависят от n

$$(y, x) - (P_n y, P_n x)_{H_n} = \sum_{k=1}^{N-p} \mu_k n^{-k} + o(n^{-N+p}),$$

для $x \in B_p$, $y \in B_q$, $q \le p$; μ_k не зависят от n, то имеет место представление (I.9), где $c_k(z^*) \in B_k + R$.

Доказательство следствия 2, основанное на применение теоремы I, использует известную оценку [22]

$$\| \Phi'(z^*)^{-1} \| \le M, \quad M = \max(1, m^{-1}),$$

где $m = \inf_{\lambda \in \sigma(A) \setminus \lambda^*} |\lambda^* - \lambda|$, $\sigma(A)$ - спектр оператора A .
Отметим, что теорема о разложимости погрешности для оператора Штурма-Лиувилля получена в [1].

§ 2. О практической эффективности экстраполяции Ричардсона

I. Уравнение Чу-Лоу

Рассмотрим систему линейных интегральных уравнений типа Чу-Лоу, возникающую в теории дисперсионных соотношений. Согласно [12] используем следующую формулировку: определить вектор-функцию $v(t) = (v_1, \dots, v_N)$, $t \in [0, \pi/2]$ из уравнений

$$(2.1) \quad Av \equiv v(t) - g(t)\{v^2(t) + (\lambda + Fv(t))^2\} = 0 ,$$

где

$$Fv(t) = K_1 v(t) + C K_2 v(t) + (E + C) l(t) v(t),$$

$$K_1 v = \frac{1}{2\pi} \int_0^{\pi/2} [v(\tau) - v(t)] \operatorname{ctg} \frac{\tau - t}{2} d\tau + \frac{1}{2\pi} \int_0^{\pi/2} [v(\tau) + v(t)] \operatorname{ctg} \frac{\tau + t}{2} d\tau ,$$

$$K_2 v = \frac{1}{2\pi} \int_0^{\pi/2} [v(\tau) - v(t)] \operatorname{tg} \frac{\tau - t}{2} d\tau + \frac{1}{2\pi} \int_0^{\pi/2} [v(\tau) + v(t)] \operatorname{tg} \frac{\tau + t}{2} d\tau ,$$

вектор параметров λ и квадратная матрица C порядка $N \times N$ заданы, E - единичная матрица,

$$l(t) = \frac{1}{\pi} \ln \frac{\cos t}{1 + \sin t} , \quad t \in [0, \tfrac{\pi}{2}] , \quad v^2 = (v_1^2, \dots, v_N^2) .$$

Определим пространство $H_0^{m+\alpha}$ - m раз дифференцируемых функций, обращающихся в нуль на концах отрезка $[0, \pi/2]$, m-я производная которых принадлежит H_0^{α} , $0 < \alpha < 1$, где

$$\|x\|_{H^\alpha} = \max_{0 \le t \le \pi/2} |x(t)| + \sup_{0 \le t, \tau \le \pi/2} |x(t) - x(\tau)| |t - \tau|^{-\alpha} .$$

Соответствующее пространство вектор-функций обозначим $H_{0,N}^{m+\alpha}$, при этом

$$\|v\|_\alpha = \max_{1 \le i \le N} \|v_i\|_\alpha , \quad v \in H_{0,N}^{\alpha} ,$$

$$\|\lambda\| = \max_{1\le i\le N} |\lambda_i|, \quad S_R = \{v(t) / v \in H_{0,N}^{m+\alpha}, \|v\|_\alpha \le R, \ 0<R\}$$

Пусть также $g(t) \in H_0^{2\nu+\alpha+1}$, $\nu \ge 2$. Производная Фреше оператора Av из (2.I) имеет вид

$$A'(v_0)v = v - 2g(t)[v_0 v + (\lambda + Fv_0) F(v(t))]. \tag{2.2}$$

На сетке $\omega_h = \{t_i \in [0, \pi/2], \ i = 0, \dots, n, \ t_i = ih, \ nh = \pi/2\}$ рассмотрим разностную задачу

$$A_h v_h \equiv v_h(t_j) - g(t_j)\{v_h^2(t_j) + [\lambda + F_h v_h(t_j)]^2\} = 0, \tag{2.3}$$
$$t_j \in \omega_h,$$

где $v_h(t_j)$ - сеточная функция на ω_h, а

$$F_h v_h(t_j) = B_h^1 v_h(t_j) + C B_h^2 v_h(t_j) + (C+E) l(t_j) v_h(t_j), \tag{2.4}$$

$$B_h^1 v_h(t_j) = h\{v_h(t_j)(a_j + \tfrac{1}{2} a_{n+j} - \tfrac{1}{2} a_{n-j}) +$$
$$+ \sum_{k=1}^{n-1} (v_h(t_k) - v_h(t_j)) a_{k-j} + \sum_{k=1}^{n-1} (v_h(t_k) + v_h(t_j)) a_{n+j}\},$$

$$B_h^2 v_h(t_j) = h\{v_h(t_j)(b_j + \tfrac{1}{2} b_{n+j} - \tfrac{1}{2} b_{n-j}) +$$
$$+ \sum_{k=1}^{n-1} [(v_h(t_k) - v_h(t_j)) b_{k-j} + (v_h(t_k) + v_h(t_j)) b_{n+j}]\},$$
$$t_j \in \omega_h, \ a_i = \operatorname{ctg}(ih/2), \ b_i = a_i^{-1}, \ i = \pm 1, \dots, \pm(2n-1).$$

При этом слагаемое $(v_h(t_k) - v_h(t_j)) a_{k-j}$ при $k = j$ заменяется на величину $2 v'(t_j)$, вычисляемую по формуле

$$v'(t_j) = (2h)^{-1}[v_h(t_{j+1}) - v_h(t_{j-1})] + \sum_{k=1}^{\nu-1} e_{2k}(t_j) h^{2k} + O(h^{2\nu}), \tag{2.5}$$

где $e_{2k}(t)$ не зависят от h.

Пусть P_h – оператор простого сноса на сетку ω_h, а $v(t) \in H_{0,N}^{2\nu+\alpha+1}$, $\nu \geq 2$. Тогда из (2.5) и формулы Эйлера-Маклорена получим

$$(2.6)\quad \gamma_h(v) \equiv P_h A v - A_h P_h v = P_h \sum_{k=2}^{2\nu-1} d_k(t) h^k + O(h^{2\nu}),$$

где $v \in H_{0,N}^{2\nu+\alpha+1}$, $d_k(t) \in H_{0,N}^{2\nu+\alpha-k+1}$ и не зависит от h.

Теорема 2. Пусть число $R > 0$ таково, что

$$(2.7)\quad 2\|g\|_\alpha \left[R + (\|\lambda\| + \|F\|_\alpha R)\|F\|_\alpha\right] < 1$$

Тогда существует единственное решение $v^*(t) \in S_R$ уравнения (2.I) $v \in H_{0,N}^{2\nu+1+\alpha}$, $\nu \geq 2$ которое может быть получено методом простой итерации. Для приближенного решения v_h^* уравнения (2.3) справедлива оценка

$$(2.8)\quad |v_h^* - v^*(t)| \leq M \gamma_h(v^*), \quad t \in \omega_h$$

и представление

$$v^*(t) = v_h^*(t) + \sum_{k=2}^{2\nu-2} c_k(t) h^k + \varepsilon_h; \quad t \in \omega_h, \ |\varepsilon_h| = O(h^{2\nu-1}),$$

где $c_k(t) \in H_{0,N}^{2\nu+\alpha-k+1}$ и не зависят от h.

Доказательство получается из теоремы I, если заметить, что при условии (2.7) оператор $A'(v^*)$ из (2.2) имеет ограниченный обратный. В силу результатов [I2] и (2.8) тоже справедливо и для $A_h'(v_h^*)$ при достаточно малом h. Остальные условия теоремы следуют из (2.6), (2. 8).

Для $N=2$, $C = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, $g(t) = \cos^2 t \sin t$, $\lambda_1 = -\lambda_2 = 0.99$ в [I3] приводятся следующие результаты уточнения решения

$$v_i^*(t) = 4\lambda_i^2 \cos^2 t \sin t \left[4 - 4\lambda_i \cos t (2\cos^2 t - 1) + \lambda_i^2 \cos^2 t\right]^{-1}, \ i = 1,2$$

на двух сетках ω_h и ω_{2h}, $h = 10, 20, 40, 80, 160$

Таблица I.

n	$\sigma_1(n)$	$\sigma(n, 2n)$	$\sigma_S(n)$
I0	$9.I4\cdot I0^{-3}$	-	$6.23\cdot I0^{-4}$
20	$I.I2\cdot I0^{-4}$	$3.20\cdot I0^{-5}$	$3.I5\cdot I0^{-5}$
40	$2.77\cdot I0^{-5}$	$2.52\cdot I0^{-6}$	$2.43\cdot I0^{-6}$
80	$7,0I\cdot I0^{-6}$	$I.60\cdot I0^{-7}$	$I.62\cdot I0^{-7}$
I60	$9.II\cdot I0^{-7}$	$4.82\cdot I0^{-9}$	$4.93\cdot I0^{-9}$

Здесь $\sigma_1(n) = \max |v_n^*(t) - v^*(t)|$, $t \in \omega_h$,
$\sigma(n, 2n) = \max |4/3\, v_{2n}^* - 1/3\, v_n^* - v^*(t)|$, $t \in \omega_{2n}$,
$\sigma_S(n)$ соответствует точности разностной задачи (2.3), когда интегралы в (2.4) вычисляются пооформуле Симпсона.

П. Задача на собственные значения

В качестве примера использования Следствия 2 рассмотрим задачу на собственные значения

$$(2.9)\quad -y'' + q(x)y + \int_0^1 K(x,\xi)y(\xi)d\xi = \lambda y, \quad 0<x<1,$$
$$y(0) = y(1) = 0, \quad \int_0^1 y^2(x)dx = 1, \quad y'(0) > 0,$$

где $q(x) \geq 0$, $q \in C^N[0,1]$, $K(x,\xi) \in C^N([0,1]\times[0,1])$, $K(x,\xi) = K(\xi, x)$, $K(x,\xi) \geq 0$.

Задачу (2.9) сформулируем в виде нелинейного операторного уравнения (I.I2) в пространстве $H^1 = \mathring{W}_2^1 + \mathcal{R}$, где $\mathring{W}_2^1[0,1]$ - подпространство функций из W_2^1 с нулями на концах отрезка $[0,I]$. Для этого перейдем к обобщенному решению задачи (2.9). Напомним, что обобщенным решением из $\mathring{W}_2^1$ задачи (2.9) называется /I3/ функция $y(x) \in \mathring{W}_2^1$, удовлетворяющая тождеству

$$(2.I0)\quad \mathcal{L}(y,z) \equiv \int_0^1 (y_x z_x + qyz + S_y z)dx = \lambda \int_0^1 yz\,dx,$$
$$\forall z \in \mathring{W}_2^1, \quad S_y = \int_0^1 K(x,\xi)y(\xi)d\xi.$$

Используя скалярное произведение

$$[y,v]=\int_0^1 (y_x v_x + qyv + S_y \cdot v)\,dx,\quad \|y\|_1^2=[y,y]$$

и теорему Ф.Рисса запишем (2.10) в виде

$$[y,v]=\lambda[Ay,v],\quad \forall v\in \mathring{W}_2^1$$

причем A есть ограниченный линейный, самосопряженный оператор $A:\mathring{W}_2^1\to\mathring{W}_2^1$ и является вполне непрерывным в $\mathring{W}_2^1$ /23/. Итак имеем равенства типа (I.I2)

$$Ay=\lambda^{-1}y,\quad [y,y]=1 \tag{2.11}$$

в пространстве $H^1=\mathring{W}_2^1+R$.

На сетке $\omega_h=\{x_i=ih,\ i=0,1,..,n,\ x_n=1\},\ h=n^{-1}$ рассмотрим аппроксимацию (2.II), с помощью сеточного аналога тождества (2.I0). Аппроксимируем интегральный оператор S_y с помощью формулы трапеций

$$S_h y=\sum_{i=1}^{n-1} K(x,\xi_i)y_i h+\frac{h}{2}\left(K(x,0)y_0+K(x,1)y_n\right).$$

Придерживаясь обозначений из /I4/ для скалярных произведений и сеточных норм, запишем разностный аналог тождества (2.I0)

$$\Lambda_h(y,v)\equiv\frac{1}{2}\left((y_{\bar x},v_{\bar x}]+[y_{\bar x},v_{\bar x})\right)+[qy,v]+$$

$$+[S_h y,v]=\lambda_h[y,v],\quad [y,y]=1.$$

В пространстве B_{+1} с нормой $\|y\|_{+1}^2=\|y\|^2+[y_x,y_x]$ введем эквивалентную перенормировку

$$[y,v]_1=\frac{1}{2}[y_x,v_x)+\frac{1}{2}(y_{\bar x},v_{\bar x}]+[qy,v]+[S_h y,v],$$

откуда с учетом $|[y,v]|\le c\|y\|_1\|v\|_1$ можно записать

$$[y,v]_1=\lambda_h[A_h y,v]_1;\quad [y,y]=1,$$

где $A_h=A_h^*$, $A_h\ge\alpha E$, $\alpha>0$. В итоге получаем разностную задачу типа (I.I3) на собственные значения

$$A_h y=\lambda_h^{-1}y,\quad [y,y]-1=0. \tag{2.12}$$

Явный вид (2.I2) в точке x_i определяется равенством

$$-(y_{i+1} - 2y_i + y_{i-1})h^{-2} + q_i y_i + [S_n y]_i = \lambda_n y_i ,$$

$$i = 1, \dots, n-1 , \quad y_0 = y_n = 0 .$$

Проверяя условия Следствия 2 для пары уравнений (2.II), (2.I2) /I4/, получаем

Теорема 3. Пусть λ^* есть однократное собственное значение задачи (2.9) с собственной функцией x^*. Тогда существует такое n_0, что при всяком $n \geq n_0$ найдется решение (λ_n^*, y_n^*) разностной задачи (2.I2), для которого справедливы разложения

$$\lambda_n^* = \lambda^* + \sum_{k=1}^{\ell} h^{2k} \mu_k + \rho_n , \quad |\rho_n| = O(h^{2\ell}), \tag{2.I3}$$

$$y_n^* = P_n x^* + P_n \Big(\sum_{k=1}^{\ell} h^{2k} v_k(t) \Big) + \gamma_n ; \quad \|\gamma_n\|_{+1} = O(h^N),$$

где $v_k(t) \in C^{N-2k+2}[0,1]$, $\ell = [N/2]$, а функции $v_k(t)$ и числа μ_k не зависят от n.

Уточнение решений на основе (2.I3) проводится в /I4/.

Ш. Интегральное уравнение

Рассмотрим далее уравнение с вполне непрерывным интегральным оператором. Его исследование идейно примыкает к случаю уравнений (2.II), (2.I2). В качестве приложения рассмотрим граничные интегральные уравнения для гармонических функций.

Рассмотрим интегральное уравнение

$$u(x) - \lambda \int_0^1 K(x,t) u(t) dt = f(x), \quad x \in \Omega = [0,1]. \tag{2.I4}$$

Обозначим $Ku = \int_0^1 K(x,t) u(t) dt$ и предположим, что λ не является собственным числом оператора K. Пусть

$$K(x,t) \in C^{2m+1}(\Omega \times \Omega), \quad u(x) \in C^{2m+1}[\Omega]. \tag{2.I5}$$

При этом уравнение (2.I4) однозначно разрешимо для любой правой части $f \in C[\Omega]$. Приближенное решение $u_h(x)$ ищем в виде кусочно-линейной функции на сетке $\omega_h = \{x_i = ih, i = 0,1,\dots,n; h = n^{-1}\}$, определяемой из уравнения

(2.I6) $$u_h(x_i) - \lambda \int_0^1 K(x_i, t)\, u_h(t)\, dt = f(x_i),$$

$$x_i \in \omega_h, \quad i = 0, 1, \ldots, n,$$

которое запишем в виде

$$(E - \lambda K_h)\, u_h = F, \quad K_h = \{k_{ij}\}; \quad i, j = 0, 1, \ldots, n,$$

$$k_{ij} = \int_{x_{i-1}}^{x_{i+1}} K(x_i, t)\, \varphi_j(t)\, dt,$$

где $\varphi_j(s)$ - кусочно-линейные базисные функции.
Свойство (С), (I.6) выполняется в силу равенства

$$P_h(E - \lambda K)u - (E - \lambda K_h) P_h u =$$

(2.I7)

$$= -\lambda \sum_{k=2}^{2m} h^k \nu_k(x) + \eta_h, \quad |\eta_h| \le c h^{2m+1}, \quad x \in \omega_h,$$

$$\nu_k(x) = \sum_{\substack{p+n=k \\ n \ge 2}} \frac{1}{p!\, n!} \frac{n-1}{(p+2)(p+n+1)} \Phi_{pn}(x), \quad x \in \Omega,$$

где

$$\Phi_{pn}(x) = \int_0^1 \frac{\partial^p}{\partial t^p} K(x, t)\, u^{(n)}(t)\, dt, \quad n \ge 2.$$

Для проверки свойства (В) Теоремы I отметим, что если правая часть $f(x) \in C^k(\Omega)$, то $u(x) \in C^k(\Omega)$, $k \le 2m+1$, т.е. можно положить $B_k = B_k' = C^k(\Omega)$. При этом $(E - \lambda K)^{-1}$ ограничен в каждом из пространств B_k.

Лемма I. Оператор $E - \lambda K_h : R^n \to R^n$ имеет равномерно ограниченный по h (или по n) обратный оператор в равномерной метрике.

Доказательство опирается на свойство полной непрерывности оператора K. Обозначим G_n - оператор кусочно-линейного восполнения с сетки ω_h. Далее, предполагая противное, имеем

$$(E - \lambda K_h) u_n = \varphi_n, \quad |\varphi_n| \to 0, \quad \|u_n\| = 1, \quad n \to \infty.$$

Рассмотрим $v_n = G_n u_n \in C(\Gamma)$, тогда

$$P_h(E - \lambda K) v_n = P_h Q_n \varphi_n \equiv P_h \Phi_n,$$

причем $|\Phi_h| \to 0$, $\|v_h\| = 1$. Если $\xi = x_i + \Delta$, $0 \le \Delta \le h$, $x_i \in \omega_h$, то

$$(E - \lambda K) v_h(\xi) = v_h(\xi) - \lambda \int_{\Omega} K(\xi, t)\, v_h(t)\, dt =$$

$$= v_h(\xi) - \lambda \int_0^1 \left[K(x_i, t)(1 - \Delta h^{-1}) + K(x_{i+1}, t)\Delta h^{-1} \right] v_h(t)\, dt +$$

$$+ O(h^2)\, |v_h| = \Phi_h(\xi) + O(h^2),$$

откуда следует $|(E - \lambda K) v_h| \to 0$, $h \to \infty$. Но так как оператор K вполне непрерывен, то последовательность $z_h = K v_h$ можно считать сходящейся в $C(\Omega)$. Но тогда $v_h \to \bar{v} \in C(\Omega)$, т.к. $|v_h - \lambda z_h| \to 0$. Поэтому $K v_h \to K \bar{v}$, откуда

$$(E - \lambda K) \bar{v} = 0 \quad , \quad \|\bar{v}\| = 1.$$

Это противоречит ограниченности $(E - \lambda K)^{-1}$. Лемма доказана.

Теорема 4. Пусть решение уравнения (2.I4) находится из системы (2.I6). Если выполнено (2.I5), а λ не является собственным числом оператора K , то справедливо разложение

(2.I8) $$u_h(x) = u(x) + \sum_{k=2}^{2m} h^k c_k(u) + O(h^{2m+1}), \quad x \in \omega_h$$

где $c_k(x) \in B_k = C^k(\Omega)$ и не зависят от h .

Примером уравнения типа (2.I4) может служить граничное интегральное уравнения (ГИУ) внешней задачи Неймана для оператора Лапласа. Пусть контур Г с непрерывной кривизной разбивает плоскость R^2 на две области Ω_i и Ω_e, $\Omega_i \cup \Gamma \cup \Omega_e = R^2$, а область Ω_e содержит бесконечно удаленную точку. Обозначим через $\partial/\partial n$ - производную по внутренней нормали к границе Г. Тогда для функции $u(x)$, гармонической в Ω_e и такой, что $u(\infty) = 0$ имеет место ГИУ

(2.I9) $$(E + K) u(x) = L \frac{\partial}{\partial n} u(x) \quad , \quad x \in \Gamma,$$

где

(2.20) $$Ku = \int_\Gamma K(x,s)\, u(s)\, ds \quad , \quad Lv = \int_\Gamma \mathcal{L}(x,s)\, v(s)\, ds,$$

$$K(x,s) = \frac{1}{\pi} \frac{\partial}{\partial n_s} \ln r^{-1}(x,s) \quad ; \quad \mathcal{L}(x,s) = \frac{1}{\pi} \ln r^{-1}(x,s),$$

$$r(x,s) = |x - s| \quad ; \quad x, s \in \Gamma.$$

, Если правая часть в (2.19) задана, то определив из этого уравнения $u(x)$, $x \in \Gamma$, легко восстановить решение во внутренних точках области Ω_e по формуле Грина. В [16] установлено, что для выполнения $u(\infty) = 0$ необходимо и достаточно, чтобы $\int \frac{\partial u}{\partial n} ds = 0$. Известно, что $\lambda = -1$ не является собственным значением оператора K из (2.20). Поэтому для приближенного решения u_h, полученного методом коллокации (2.16), справедливо разложение (2.18) при достаточной гладкости функций $u(x)$ и $K(x,s)$ (2.15).

В табл.2 приведены результаты экстраполяции Ричардсона для решения u_h в точках границы Г и во внутренних точках области Ω_e. Полагаем $\Omega_i = \{ x,y : |x| \le 1, |y| \le 0.5 \}$, $u(x,y) = x(x^2+y^2)^{-1}$, $\delta(h) = \max\limits_{x \in \omega_h} |u_h - u|$, $\delta(h_1, \dots, h_k)$ – погрешность решения, экстраполированного на последовательности сеток $\omega_{h_1}, \dots, \omega_{h_k} \in \Gamma$ с равномерным шагом h_k, $\delta(h) = \max\limits_{x \in D_e} |u_h - u|$, $\delta(h_1, \dots, h_k)$ – максимум погрешности экстраполированных решений на некотором множестве D_e – внутренних точек области Ω_e. В силу симметрий расчеты проводились лишь на четверти контура Г.

Таблица 2.

h	I/8	I/I6	I/32	I/64	
$\delta(h)$	$1.33 \cdot 10^{-2}$	$3.56 \cdot 10^{-3}$	$9.25 \cdot 10^{-4}$	$2.34 \cdot 10^{-4}$	внутренние точки
$\delta(h, h/2)$	$3.47 \cdot 10^{-4}$	$3.48 \cdot 10^{-5}$	$4.05 \cdot 10^{-6}$	$1.71 \cdot 10^{-8}$	$\delta(h, \frac{h}{2}, \frac{h}{4}, \frac{h}{8})$
$\delta(h, h/2, h/4)$	$9.76 \cdot 10^{-6}$	$5.47 \cdot 10^{-7}$	$1.57 \cdot 10^{-7}$	$2.35 \cdot 10^{-6}$	$\delta(h, \frac{h}{2}, \frac{h}{4})$
$\delta(h, \frac{h}{2}, \frac{h}{4}, \frac{h}{8})$	$6.7 \cdot 10^{-8}$	$4.61 \cdot 10^{-7}$	$4.8 \cdot 10^{-6}$	$5.48 \cdot 10^{-5}$	$\delta(h, \frac{h}{2})$
точки границы	$1.04 \cdot 10^{-4}$	$4.15 \cdot 10^{-4}$	$1.64 \cdot 10^{-3}$	$6.42 \cdot 10^{-3}$	$\delta(h)$
	I/64	I/32	I/I6	I/8	h

Использовался комплекс программ, составленный в ЛВТА ОИЯИ для решения внутренних и внешних задач Дирихле и Неймана методом ГИУ.

§ 3. Экстраполяция при помощи поворота системы координат

При уточнении приближенного решения u_h^* на основе разложения (I.9) исключаются несколько первых слагаемых $c_i(u^*)h^{\alpha_i}$, $i=1,\dots,p$ в (I.9) путем комбинирования $p+1$ выражения (I.9) при попарно различных значениях h. При этом необходимо построение различных сеточных областей и интерполирование с нужной точностью.

В случае, если оператор A инвариантен относительно поворота системы координат, проводить уточнение можно на одной и той же сеточной области. Если известна зависимость коэффициентов $c_i(x)$ от угла поворота системы координат, то также можно исключить несколько первых слагаемых в (I.9). Рассмотрим уравнение Пуассона в случае $N=2,3,4$ независимых переменных и бигармоническое уравнение.

Начнем с задачи Дирихле для уравнения

(3.I) $$\Delta u = f(x,y), \quad u_\Gamma = \varphi(\xi), \ \xi\in\Gamma$$

в прямоугольнике $\Pi=\{0\le x\le a;\ 0\le y\le b\}$, покрытом квадратной разностной сеткой ω_h с шагом h. Легко видеть, что для аппроксимации оператора Δ наряду с обычным шаблоном типа "крест", можно использовать также шаблоны с шагом $\tau_k=h\sqrt{k^2+1}$, $k=1,2,\dots$, получающиеся при поворотах на угол $\varphi_k=\pm\operatorname{arctg}k$, $k=1,2,\dots$. В случае $k=1$, соответствующем углу $\varphi_1=\pi/4$ построим на сетке ω_h наряду с оператором Δ_h другую разностную аппроксимацию Δ_τ по формуле

(3.2) $$\Delta_\tau u(x,y)=\tau^{-2}[u(x-h,y-h)+u(x+h,y+h)+u(x-h,y+h)+$$
$$+u(x+h,y-h)-4u(x,y)], \quad \tau=h\sqrt{2}.$$

Рассмотрим решения следующих разностных задач

(3.3) $$\Delta_h u_h = P_h f(x,y), \quad u_{h,\Gamma}=\varphi(\xi)$$
$$\Delta_\tau u_\tau = P_\tau f(x,y), \quad u_{\tau,\Gamma}=\varphi(\xi); \quad P_\tau=P_h.$$

Теорема 4. Пусть $f(x,y)=0$ при $(x,y)\in\Gamma$ и выполнено условие $\frac{d^4u}{dx^4}+\frac{d^4u}{dy^4}=0$ в углах прямоугольника. Тогда для $u(x,y)\in C^6(\bar\Pi)$ справедливо представление

$$(3.4)\qquad \frac{2u_h+u_\tau}{3}+\frac{h^2}{12}f(x,y)=u(x,y)+O(h^4);\ (x,y)\in\omega_h,$$

а для $u(x,y)\in C^8(\bar\Pi)$ выполнено

$$\frac{1}{45}\left[2(16u_{h/2}-u_h)+16u_{\tau/2}-u_\tau\right]+\frac{h^2}{60}f(x,y)=u(x,y)+\mu_h,$$

$$|\mu_h|\le ch^6\ln h;\ (x,y)\in\omega_h.$$

Для доказательства достаточно установить разложение (I.9) для решений u_h и u_τ из (3.3) и найти зависимость между коэффициентами $c_i(x,y)$ при одинаковых степенях h и τ /I5/.

При организации расчетов следует учитывать, что сеточная область при повороте на угол φ_k, $k=1,2,\dots$ распадается на k^2+1 независимых шаблонов, что позволяет проводить расчеты решения u_τ лишь на массиве размерности $N/_{k^2+1}$, где N число точек на сетке ω_h, а затем использовать u_τ как начальное приближение для u_h и таким образом экономить ресурсы ЭВМ. Поворот на угол $\varphi=\pm\operatorname{arctg}2$ рассмотрен в /I5/.

Перейдем к случаю трех и четырех независимых переменных. В прямоугольном параллелепипеде $\Pi=\{0\le x_i\le a_i,\ i=1,2,\dots,N\}$, где либо $N=3$, либо $N=4$, покрытом кубической сеткой ω_h с шагом h, наряду с обычной семиточечной схемой для $N=3$ рассмотрим три разностных оператора Δ_{μ_i}, $i=1,2,3$, которые строятся аналогично оператору Δ_τ из (3.2) при повороте относительно одной из трех осей, параллельных Ox_i, $i=1,2,3$:

$$(3.4)\qquad \Delta_{\mu_i}u(x)=h^{-2}\left(u(x+\Delta_i)-2u(x)+u(x-\Delta_i)\right)+\Delta_\tau u(x),$$

где Δ_τ действует по формуле (3.2) в плоскости ортогональной оси Ox_i, а вектор $\Delta_i=he_i$, где e_i - единичный вектор, параллельный Ox_i. Уравнения

$$(3.5)\qquad \Delta_{\mu_i}u_{\mu_i}=P_nf,\quad u_{\mu_i,\Gamma}=\varphi(\xi),\quad i=1,2,3$$

однозначно разрешимы и для u_{μ_i} справедливы априорные оценки

$$|u_{\mu_i}|\le c\max|f,\varphi|.$$

Теорема 5. Пусть $u(x) \in C^6(\bar{\Pi})$, $f_r = 0$ и решения $v_i(x)$, $i = 1,2,3$ уравнения

$$\Delta v_i = -\frac{1}{12}\left(\sum_{k=1}^{3} \frac{d^4}{dx_k^4} + 6\frac{d^4}{dx_m^2 dx_n^2}\right)u; \quad m \neq n \neq i, \quad v_{i,r} = 0$$

таковы что $v_i \in C^4(\bar{\Pi})$. Тогда для решений u_{μ_i} (3.5) справедливо

$$\frac{1}{3}\sum_{i=1}^{3} u_{\mu_i}(x) + \frac{h^2}{12} f(x) = u(x) + O(h^4), \quad x \in \omega_h.$$

В случае $N = 4$ оператор Лапласа тремя разными способами представим в виде суммы двумерных

$$\Delta = \Delta(x_1, x_2) + \Delta(x_3, x_4) = \Delta(x_1, x_3) + \Delta(x_2, x_4) = \Delta(x_1, x_4) + \Delta(x_2, x_3),$$

где аргументы указывают, для каких переменных определен соответствующий оператор на плоскости. Для каждого из трех представлений построим на ω_h разностный оператор Δ_{τ_i}, $i = 1,2,3$, являющийся суммой двух операторов типа Δ_τ из (3.2). Например

$$\Delta_{\tau_1} u = \Delta_\tau(x_1, x_2)u + \Delta_\tau(x_3, x_4)u.$$

Теорема 6. Пусть $u \in C^6(\bar{\Pi})$, $f_r = 0$ и решения $v_i(x)$, $i = 1,2,3$ уравнения

$$\Delta v_i = -\frac{1}{12}\left(\frac{1}{2}\sum_{k=1}^{4} \frac{d^4}{dx_k^4} + 3D_i\right)u; \quad v_{i,r} = 0, \quad i = 1,2,3$$

где

$$D_1 = \frac{d^4}{dx_1^2 dx_2^2} + \frac{d^4}{dx_3^2 dx_4^2}; \quad D_2 = \frac{d^4}{dx_1^2 dx_3^2} + \frac{d^4}{dx_2^2 dx_4^2}; \quad D_3 = \frac{d^4}{dx_1^2 dx_4^2} + \frac{d^4}{dx_2^2 dx_3^2},$$

имеют ограниченные четвертые производные, т.е. $v_i \in C^4(\bar{\Pi})$. Тогда для решений u_{τ_i} уравнений

$$\Delta_{\tau_i} u_{\tau_i} = P_h f, \quad u_{\tau_i, r} = \varphi(x), \quad i = 1,2,3$$

справедливо представление

$$\frac{1}{3}\sum_{i=1}^{3} u_{\tau_i}(x) + \frac{h^2}{4} f(x) = u(x) + O(h^4), \quad x \in \omega_h.$$

Рассмотрим далее бигармоническое уравнение

(3.6) $$\Delta^2 u = f(x,y), \quad u_\Gamma = 0, \quad \Delta u_\Gamma = 0$$

В прямоугольнике Π, покрытом сеткой $\omega_h = \{(x_i, y_j);$ $x_i = ih,\ i = 0,1,\dots,N;\ y_j = jh,\ j = 0,1,\dots,M\}$ с шагом h. Уравнение (3.6) аппроксимируем соотношением

(3.7) $$\Delta_h^2 u_h \equiv u_{h\bar{x}x\bar{x}x} + 2u_{h\bar{x}x\bar{y}y} + u_{h\bar{y}y\bar{y}y} = P_h f,$$

$$u_{h,\Gamma} = 0, \quad \Delta_h u_{h,\Gamma} = 0,$$

используя фиктивные точки $(x_i, b+h)$, $(x_i, -h)$, $(-h, y_j)$, $(a+h, y_j)$, $1 \le i \le N-1$, $1 \le j \le M-1$. После поворота системы координат на $\varphi = \pi/4$ рассмотрим на сетке ω_h решение u_τ уравнения

(3.8) $$\Delta_\tau^2 u_\tau = P_h f, \quad u_{\tau,\Gamma} = 0, \quad \Delta_\tau u_{\tau,\Gamma} = 0,$$

где u_τ определен в (3.2).

Теорема 7. Пусть $u(x,y) \in C^8(\bar{\Pi})$, $f(x,y)_\Gamma = 0$ и решение $v(x,y)$ уравнения

$$\Delta^2 v = -\frac{1}{6}\left(\Delta^3 - 2\Delta\frac{d^4}{dx^2dy^2}\right)u;$$

$$v_\Gamma = 0, \quad \Delta v_\Gamma = -\frac{1}{12}\left(\Delta^2 - 2\frac{d^4}{dx^2dy^2}\right)u,$$

таково, что $v \in C^6(\bar{\Pi})$. Тогда для решений u_h и u_τ уравнений (3.7),(3.8) справедливо представление

$$\frac{2u_h + u_\tau}{3} + \frac{h^2}{6}\Delta_h u_h = u(x,y) + O(h^4), \quad (x,y) \in \omega_h.$$

Доказательства теорем 4–7 имеются в [15]. Для иллюстрации теоремы 4. приведем результаты уточнения по формуле (3.4) решения $u = \sin\pi x \sin\pi y$ уравнения

$$\Delta u = -2\pi^2 \sin\pi x \sin\pi y, \quad u_\Gamma = 0$$

в квадрате Π для $a = b = 1$. Результаты расчетов при $h = 0.1$ приведены в таблице 3.

Таблица 3.

	$u_h - u$	$u_r - u$	$v_h - u$
$x=y=0.5$	$-8.27\cdot10^{-3}$	$-3.36\cdot10^{-2}$	$2.47\cdot10^{-4}$
$x=0.1$ $y=0.2$	$1.5\cdot10^{-3}$	$-6.1\cdot10^{-3}$	$-4.5\cdot10^{-5}$

Здесь сеточная функция v_h получена по формуле (3.4).

§ 4. Итерационные процессы на последовательности сеток

При ускорении сходимости итерационных процессов решения систем конечно-разностных уравнений (I.2), где $A: R^N \to R^N$, одной из плодотворных идей оказалось включение задачи (I.2) в семейство задач

(4.I) $$A_{n_i} u_{n_i} = f_{n_i} \; ; \; i = 1,2,\dots,l \; , \; n_e = N,$$

соответствующих различным дискретизациям исходного уравнения.

Число итераций для достижения заданной точности ε пропорционально двум сомножителям: $\ln \varepsilon^{-1}$, и некоторому множителю R^{-1}, определяемому спектральными свойствами оператора перехода итерационного процесса. Избавиться от $\ln \varepsilon^{-1}$, используя последовательность сеток, можно используя лишь общие характеристики алгоритмов, такие, как погрешность аппроксимации и скорость сходимости итераций [7,8]. Подавление множителя R^{-1} основано на более специальных подходах [24-27].

Пусть для решения уравнения (I.2) заданного на сеточной области Ω_h построен итерационный процесс ($u_h \equiv u_n$)

(4.2) $$\frac{u_h^{k+1} - u_h^k}{\tau} = -\Psi(u_h^k) \; ; \; \Psi(u_h) = 0,$$

сходящийся к решению u_h со скоростью

$$\| u_h^k - u_h \| \le [q(h)]^k \| u_h^0 - u_h \| \; , \; q < 1.$$

При построении итерационных процессов на последовательности сеток Ω_{h_i}, $i = 1,2,\dots,l$; $h_1 > h_2 > \dots > h_l = h$ решается задача (I.2) методом (4.2) на сетке Ω_{h_1} с точностью ε_1 : $\| u_{h_1}^k - u_{h_1} \| \le \varepsilon_1$, после чего сеточная функция $u_{h_1}^k$

интерполируется на сетку Ω_{h_2} , где используется как начальное приближение для итераций (4.2), проводимых до уменьшения по отрезкам до величины $\varepsilon_2 < \varepsilon_1$ и т.д. В работах [7,8] оценивается эффективность такой процедуры при использовании двух или нескольких вспомогательных сеток. Если $\varepsilon_i = O(h_i^\alpha)$, где h_i^α - погрешность приближенных решений u_{h_i} , то как правило число итерации уменьшается в $O(\ln \varepsilon_e^{-1})$ раз при $h_e \to 0$. Далее, согласно [19], рассмотрим предельный случай, когда $\ell \to \infty$, и вычислим асимптотический коэффициент экономии. для нелинейного случая. При этом полагаем, что размерность задачи возрастает непрерывно, что позволяет построить дифференциальное уравнение для числа арифметических действий.

В пространстве R^N рассмотрим эволюционный процесс

$$(4.3)\qquad \frac{du}{dt} = -\psi(u),\quad u(0)=u^0;\quad \psi(u_h)=0,$$

сходящийся со скоростью

$$\|u(t)-u_h\| \le \exp(-\sigma(N)t)\|u^0-u_h\| \equiv B\exp(-\sigma t)$$

и являющийся непрерывным аналогом (4.2). Положим для удобства $q^k = \exp(-\sigma(N)k)$. Предположим, что процесс (4.3) ведется до уменьшения начальной погрешности $B = \|u^0 - u_h\|$ в N^p раз, т.е. $\exp(-\sigma(N)t) = N^{-p}$, откуда $t = p\,\sigma(N)^{-1}\ln N$. Здесь t имеет смысл числа итераций для (4.3). Если на одну итерацию используется $\kappa_0 N^m$ арифметических действий, то их суммарное число выразится формулой

$$(4.4)\qquad Q_0(N) = p\kappa N^m \sigma^{-1}(N)\ln N \equiv \frac{p\kappa}{m}\int_0^N n^{-1}\sigma(N)^{-1}\ln N\,dn .$$

Рассмотрим процесс (4.3) на последовательности пространств R^n , $\kappa_0 \le n \le N$, $n \in \mathbb{Z}$, в каждом из которых существует решение u_n уравнения (I.2). Пусть имеется оператор интерполирования $P_{nm} : R^n \to R^m$, $n \le m$, $\|P_{nm}\| = 1$, такой что

$$(4.5)\qquad \|P_{nm}u_n - u_m\| \le c(m,n)\,n^{-p-q},\quad c \le B,\ q \ge 0 .$$

Полагая, что для $\forall n$ выполнено $t = p\sigma(n)^{-1}\ln n$, и используя элемент $P_{n,n+\Delta}\,u_n$ как начальное приближение в пространстве $R^{n+\Delta}$, легко посчитать приращение вычислительной работы /19/

$$Q(n+\Delta) - Q(n) = \tag{4.6}$$
$$= -\kappa(n+\Delta)^m \sigma^{-1}(n+\Delta)\ln\left[1+\beta^{-1}n^{-2}c(n,n+\Delta)\right]\left(\frac{n}{n+\Delta}\right)^p.$$

Устремляя $\Delta \to 0$ и полагая $c(n, n+\Delta) = O(\Delta)$ из (4.6) получаем дифференциальное уравнение для $Q(n)$

$$\frac{dQ}{dn} = p\kappa n^{-1+m}\sigma^{-1}(n), \quad n_0 \le n \le N,$$

интегрируя которое получаем

$$Q(N) = Q(n_0) + p\kappa\int_{n_0}^{N} n^{m-1}\sigma^{-1}(n)\,dn. \tag{4.7}$$

Формулу (4.7) можно считать непрерывным аналогом выражения (30) из /7/ для объема вычислений на последовательности сеток. Отношение величины $Q_0(N)$ из (4.4) к $Q(N)$ из (4.7) и дает коэффициент экономии $\varkappa$. В таблице 4 приводится коэффициент $\varkappa$ для случая разностных эллиптических уравнений, где N - размерность задачи по одной переменной, m -число переменных. Полагаем $Q(n_0) = 0$. Напомним, что $\sigma(N) = N^{-2}$ соответствует методу Зейделя, $\sigma(N) = N^{-1}$ - методу ПВР; $\sigma(N) = N^{-1/2}$ - попеременно-треугольный метод; $\sigma(N) = [\ln N]^{-1}$ - метод переменных направлений, σ не зависит от N для методов, оптимальных по порядку числа арифметических действий. Для столбца НК (метод Ньютона-Канторовича)) полагаем $\|u(t) - u_h\| \le \exp(-\sigma_0 2^t)\|u^0 - u_h\|$

Таблица 4.

m \ $\sigma(N)$	N^{-2}	N^{-1}	$N^{-1/2}$	$[\ln N]^{-1}$	const	НК
2	$4\ln N$	$3\ln N$	$\frac{5}{2}\ln N$	$\varkappa > 2\ln N$	$2\ln N$	$\varkappa > 2\ln\ln N$
3	$5\ln N$	$4\ln N$	$\frac{7}{2}\ln N$	$\varkappa > 3\ln N$	$3\ln N$	$\varkappa > 3\ln\ln N$

В работе /19/ показано, что если в формуле (4.5) $q \ge 1$, т.е. используется интерполяция более высокого порядка (в смысле (4.5)), чем точность аппроксимации h^{-p} уравнения (I.2), то величину $Q(N)$ в (4.7) можно существенно уменьшить. Такой подход исполь-

зован в работе [18], где повышение точности интерполяции (4.7) достигается за счет использования разложения (I.9) по степеням шага дискретизации. Пусть на сетках Ω_{h_i}, $i = 1, \dots, l-1$ получены решения u_{h_i} с одинаковой точностью $\varepsilon = O(h_e^{\alpha})$, где α - максимальная степень h в разложении (I.9). Тогда для вычисления u_{h_e} используем начальное приближение

$$u^{0}_{h_e} = \sum_{i=1}^{l-1} \gamma_i u_{h_i}, \quad x \in \Omega_{h_e}, \tag{4.8}$$

где u_{h_i} интерполируем на сетку Ω_{h_e} с точностью $O(h_e^{\alpha})$, а коэффициенты γ_i определяются из системы

$$\begin{cases} \gamma_1 + \gamma_2 + \dots + \gamma_{e-1} = 1 \\ \gamma_1 h_1 + \gamma_2 h_2 + \dots + \gamma_{e-1} h_{e-1} = h_e \\ \dots\dots\dots\dots \\ \gamma_1 h_1^{e-1} + \dots + \gamma_{e-1} h_{e-1}^{e-1} = h_e^{e-1}. \end{cases} \tag{4.9}$$

Таким образом, с помощью (4.8) производится экстраполяция к точному решению разностной задачи, т.к.

$$u_{h_e} = \sum_{i=1}^{l-1} \gamma_i u_{h_i} + O(h_e^{l-1}), \quad x \in \Omega_{h_e}.$$

Проиллюстрируем эффективность этого подхода (подробности см. в [18]) для разностного оператора Лапласа (3.3) при условии, что существует разложение

$$u_h = u(x,y) + c_1(x,y) h^2 + O(h^4), \quad (x,y) \in \Omega_h.$$

При этом из (4.8), (4.9) следует

$$u_{\frac{h}{4}} = \frac{5}{4} u_{\frac{h}{2}} - \frac{1}{4} u_h + O(h^4), \quad (x,y) \in \Omega_h. \tag{4.10}$$

Для расчета гармонической в области $\Pi = \{0 \le x, y \le 1\}$ функции $u = \exp \pi y \sin \pi x$ на сетке Ω_h размерности I29xI29, $h = 1/128$ использовались вспомогательные сетки: Ω_{2h}, Ω_{4h}, Ω_{8h}, Ω_{16h}, на каждой из которых проводились расчеты с одинаковой точностью ε (табл.5) либо методом ПВР, либо Зейделя (3). Через $\kappa(h)$ обозначено число итераций на сетке Ω_h, число $\kappa\Sigma$ - общее число итераций в пересчете на сетку Ω_h с учетом интерполяции $\kappa\Sigma = \sum_{i=0}^{4} \kappa(2^i h) 4^{-i} + 1$, число γ означает асимптотическую эффективность итерационного процесса $\exp(-\gamma \cdot \kappa\Sigma) = \varepsilon \cdot \varepsilon_0^{-1}$, где ε_0 - начальная погрешност

Таблица 5.

ε	$k(16h)$	$k(8h)$	$k(4h)$	$k(2h)$	$k(h)$	$k\Sigma$	γ
10^{-4}	21	24	28	2	1	3.7	2.00
10^{-5}	24	32	47	6	2	7.0	1.43
10^{-6}	27	37	64	44	5	20.6	0.67
10^{-7}	30	44	67	278	33	107.0	0.15
	ПВР	ПВР	ПВР	3	3		

При $\varepsilon = O(h^2) \simeq 10^{-4}$ число итераций практически монотонно убывает от сетки к сетке, то есть суммарная вычислительная работа оценивается величиной $O(h^{-2})$. Отметим, что наибольшая эффективность релаксационного метода /26/ для аналогичной задачи характеризуется величиной γ =0.42.

Другим примером эффективного использования последовательности сеток является интегро-разностный метод /20/ решения задачи Дирихле для уравнения Лапласа, оптимальный по порядку числа арифметических действий. Этот метод основан на поочередном использовании ГИУ (2.19) и разностного уравнения (3.3) в некоторой приграничной полосе.

§ 5. Заключение

Изложенные здесь приемы использования разложения (1.9) для уточнения приближенных решений и при организации расчетов на последовательности сгущающихся сеток представляют эффективный и единообразный подход при решении широкого круга задач. При этом реализация алгоритмов существенно не меняется при переходе к нелинейным проблемам и при увеличении числа пространственных переменных. В последнем случае наряду с (1.9) можно использовать многопараметрические формулы разложения погрешности. Рассмотрение исходного уравнения как предела последовательности разностных задач (1.2) позволяет эффективно использовать априорную информацию о решении без усложнения структуры разностной задачи, а в нелинейном случае иметь надежный способ для построения начальных приближений. Такая организация численных расчетов может служить одной из основ при создании комплексов программ для решения ряда задач математической физики.

Авторы выражают благодарность Э.А.Айряну, М.Нгуену, О.И.Юлдашеву за плодотворное сотрудничество в процессе работы над проблемой.

ЛИТЕРАТУРА

I. Марчук Г.И., Шайдуров В.В. Повышение точности решений разностных схем. "Наука", М., I979.

2. Штеттер Х. Анализ методов дискретизации для обыкновенных дифференциальных уравнений. "Мир", М., I978.

3. Pereyra V. Iterated deferred corrections for nonlinear operator equations, Numer. Math. 1967, vol. 10, No4, p. 316-323.

4. Joyce D.C. Surrey of extrapolation processes in numerical analysis., SIAM Revew.,vol. 13, No4, 1971, p. 435-490.

5. Урванцев А.Л., Шайдуров В.В. В сб.: Вариационно-разностные методы решения задач математической физики. Новосибирск, ВЦ СО АН СССР, I976, с.I37-I44.

6. Волков Е.А. Решение задачи Дирихле методом уточнений разностями высших порядков. -Дифф. уравнения, I965, т.I, I-№7, с.946-960; II-№8, с.I070-I084.

7. Ильин В.П., Свешников В.М. О разностных методах на последовательности сеток. Численные методы механики сплошной среды. ИБ (Новосибирск), I97I, т.2, № I, с.43-54.

8. Коновалов А.Н. Об одном способе построения итерационных процессов. Изв. СО АН СССР, сер. техн.наук, I967, № I3, вып.3, с.I05-I08.

9. Баатар Д., Пузынин И.В., Ракитский А.В. ОИЯИ, PII-I2908, Дубна, I979.

I0. Урванцев А.Л. В сб.: Численные методы решения задач электронной оптики. Новосибирск, ВЦ СО АН СССР, I979, с.77-88.

II. Жидков Е.П., Нгуен М., Хоромский Б.Н. ОИЯИ, P5-I2979, Дубна, I979.

I2. Жидков Е.П., Нгуен М., Хоромский Б.Н. ОИЯИ, P5-80-259, Дубна, I980.

I3. Жидков Е.П., Нгуен М., Хоромский Б.Н. Повышение точности приближенных решений нелинейного сингулярного интегрального уравнения типа Чу-Лоу. Ж. вычисл. матем. и матем. физ., I98I, т.2I, № 4, с.962-969.

I4. Нгуен М., Хоромский Б.Н., Ямалеев Р.М. Уточнение разностных решений задачи на собственные значения для интегро-дифференциального уравнения. Дифф.уравнения, I980, т.I6, № 7, с.I293-I302.

I5. Хоромский Б.Н. ОИЯИ, P5-80-736, Дубна, I980.

I6. Жидков Е.П., Хоромский Б.Н., Юлдашев О.И. ОИЯИ, II-8I-398, Дубна, I98I.

I7. Жидков Е.П., Хоромский Б.Н., Айрян Э.А. ОИЯИ, P5-80-6I7, Дубна, I980.

I8. Айрян Э.А., Жидков Е.П., Хоромский Б.Н. ОИЯИ, 5-8I-820, Дубна, I98I.

I9. Жидков Е.П., Хоромский Б.Н. ОИЯИ, 5-8I-783, Дубна, I98I.

20. Хоромский Б.Н. ОИЯИ, PII-8I-823, Дубна, I98I.

2I. Вайнберг М.М. Вариационный метод и метод монотонных операторов. "Наука", М., I972.

22. Гареев Ф.А. и др. Численное решение задач на собственные значения для интегро-дифференциальных уравнений теории ядра. ЖВМ и МФ, I977, т.I7, № 2, с.407-4I9.

23. Ладыженская О.А. Краевые задачи математической физики. М., Наука, I965.

24. Самарский А.А., Николаев Е.С. Методы решения сеточных уравнений. "Наука", М., I978.

25. Капорин И.Е., Николаев Е.С. Дифф. уравнения, I980, т.I6, № 7, с.I2II-I225.

26. Федоренко Р.П. УМН, I973, т.28, вып.2, с.I2I-I82.

27. Кузнецов Ю.А. В сб.: Вариационно-разностные методы в математической физике, Новосибирск, I978, с.I78-2I2.

ЧИСЛЕННЫЕ ЭКСПЕРИМЕНТЫ ПО ИССЛЕДОВАНИЮ ДИНАМИЧЕСКИХ СВОЙСТВ НЕОДНОМЕРНЫХ СОЛИТОНОВ

А.Б. ШВАЧКА

Объединенный институт ядерных исследований
Лаборатория вычислительной техники и автоматизации

АННОТАЦИЯ

Приведен краткий обзор результатов численных экспериментов по исследованию динамических свойств неодномерных квазисолитонов для ряда нелинейных моделей классической теории поля. Показано, что типы взаимодействия неодномерных квазисолитонов являются модельно-независимыми, по крайней мере, для исследованных моделей.

ABSTRACT

The brief review of the dynamical properties of many-dimensional quasi-solitons studied by means of the computer simulation in the framework of the non-linear classical field theory models is presented. It is shown that the types of soliton interactions are model independent for studied models.

I. Согласно теореме Деррика, в рамках релятивистски инвариантных моделей теории поля (исключая модели с градиентным взаимодействием) поверхность постоянной энергии в функциональном пространстве не может быть долиной. В лучшем случае это седловина, то есть поверхность, не обладающая абсолютным минимумом. Абсолютно устойчивых решений в таких моделях не существует. Одна из возможностей получения тем не менее устойчивых солитоноподобных решений (СПР) - введение некоторой изоптопической группы симметрии лагранжиана и связанных с ней законов сохранения.

Мы обсудим свойства моделей /I/ с наиболее простой $U(I)$ группой, ведущей к закону сохранения "изозаряда":

$$Q = \frac{i}{2} \int (\Phi_t^* \Phi - \Phi^* \Phi_t)\, d^D x, \quad \frac{dQ}{dt} = 0,$$

где Φ - полевая функция, D - размерность пространства. Сохранение "изозаряда" Q означает наличие ограничения на возможные виды возмущений, а именно $\delta Q[\Phi]=0$, что приводит к условию устойчивости СПР /2/ (то есть решения, обладающего "хорошими" свойствами в нуле и на бесконечности):

$$\frac{\omega}{Q} \frac{dQ}{d\omega} < 0. \qquad (I)$$

Очевидно, что действительные стационарные полевые конфигурации не могут удовлетворять этому условию и будут неустойчивы. Все эти выводы были подтверждены ранее в численных экспериментах различных групп.

В заключение этого пункта подчеркнем важность исследования неодномерных (пространственно) СПР, поскольку до сих пор, известна лишь пара эволюционных двумерных вполне интегрируемых моделей (уравнение Кадомцева-Петвиашвили (КП) и цилиндрическое уравнение КдВ), в рамках которых солитонные решения обладают в одном из направлений неудобными свойствами на бесконечности (слабое степенное убывание). Более того, распад начального состояния уже не представляется столь удивительным, как в одномерном случае. Поэтому особый интерес представляет изучение с помощью ЭВМ динамических свойств двумерных, а затем и трехмерных, хорошо локализованных решений для различных моделей теории поля. Исследование качественных

свойств этих решений с помощью ЭВМ может подсказать пути к их дальнейшему изучению аналитическими (возможно приближенными) методами.

2. Рассмотрим две модели классической теории поля с потенциалом взаимодействия в лагранжиане ж):

$$U \simeq \ln(1+|\Phi|^2)\,\Phi^2 \qquad (2)$$

и

$$U \simeq \ln(|\Phi|^2)\,\Phi^2. \qquad (3)$$

Легко видеть, что эти модели принципиально различны в следующем смысле: первая в пределе $\Phi \to 0$ с точностью до $O(|\Phi|^2)$ переходит в обычную свободную теорию, поскольку $\Phi^2 \ln(1+|\Phi|^2) \simeq \Phi^2$, вторая модель содержит в себе конституенты с бесконечной массой, так как $\ln(|\Phi|^2) \to -\infty$ при $\Phi \to 0$. Это означает, что в рамках первой модели возможен при определенных условиях распад нелинейного решения на конституенты с излучением линейных плоских волн. Во втором случае такой распад невозможен (запрещен законом сохранения), и все возможные конфигурации полей состоят только из нелинейных решений; модели второго типа иногда называют "конфайнинг" моделями. В результате в рамках первой модели СПР могут распадаться, в то время как в рамках второй неустойчивость СПР проявляется в виде их коллапса. Вторая модель интересна еще и тем, что в ней СПР могут быть найдены в явном виде для любой размерности D. Более того, из Q-теоремы следует, что независимо от D устойчивые $U(1)$ симметричные СПР вида

$$\Phi = \psi(r)\, e^{-i\omega t}$$

существуют при $\omega > \omega_{cr} = 2^{-1/2}$. В этом смысле модель $\ln(|\Phi|^2)$ размерно инвариантна и качественно отличается от модели (2), где ω_{cr} существенно зависит от D. Графики зависимости Q от ω при $D = 2$ и $D = 3$ представлены на рис.I.

ж) В случае наиболее простой Φ^4_D теории поля при $D > 2$ устойчивых СПР не существует даже в системах с изогруппой.

Покажем, что характер взаимодействия солитонов в столкновениях определяется дисперсионной зависимостью $Q(\omega)$, а не типом модели (независимо от характера неустойчивости - распад или коллапс).

Это предположение было проверено в серии численных экспериментов /3,4/. В расчетах варьировались два параметра: скорость относительного движения квазисолитонов v и величина их заряда Q . В обоих случаях выявлены четыре вида взаимодействия:

I) упругое и квазиупругое взаимодействие квазисолитонов;

2) распад (коллапс) провзаимодействовавших квазисолитонов;

3) распад (коллапс) через короткоживущее связанное состояние (резонанс);

4) долгоживущее связанное состояние двух квазисолитонов-бион, что указывает в действительности на модельно-независимый характер взаимодействия солитонов (во всяком случае, в рамках рассмотренных моделей). Последние два типа взаимодействия возможно лишь в области $dQ/d\omega \simeq 0$. Это дает основание предполагать, что подобные виды взаимодействия будут присутствовать и в моделях, лагранжианы которых допускают более высокие группы симметрии, в случае, если зависимость соответствующего "изозаряда" ("изоспина" и т.д.) будет аналогична $Q(\omega)$ на рис.I.

Более детальное исследование процесса взаимодействия квазисолитонов показывает, что он зависит также от величины прицельного параметра ρ (или, что то же самое углового момента $\ell = \rho m v$) и начальной разности фаз $\Delta\theta$. Численные эксперименты показали, что:

а) существует некоторая резонансная область по величине момента ℓ , в которой неупругость взаимодействия, квазисолитонов резко возрастает (см.также /5/);

б) чисто антисимметричная начальная полевая конфигурация приводит к упругому расталкиванию квазисолитонов.

3. Как мы уже отмечали выше, стационарные конфигурации действительных полей не могут быть устойчивы, то есть действительных квазисолитонов не существует. Более того, не во всех системах с внутренней изосимметрией и не всегда существуют устойчивые квазисолитоны. Такие решения могут возникать в системах,в которых поверхность постоянной энергии в функциональном пространстве может иметь условные (или локальные) минимумы. Естественно возникает вопрос: существуют ли в подобных системах нестационарные устойчивые конфигурации действительных полей? При этом нестационарность играет роль стабилизирующего фактора, аналогичного зависимости $exp(-i\omega t)$ в случае $U(1)$ группы *).

Это предположение было проверено в серии численных экспериментов в рамках модели (2), проведенных в Дубне /6/. Результаты, полученные в этих экспериментах, выглядели, на первый взгляд, парадоксально. Помещая неустойчивые солитоноподобные объекты достаточно близко друг к другу - так, чтобы кинематическое время их взаимодействия было меньше времени распада каждого из них (рис.2), (рис.2), мы наблюдали при достаточно малых скоростях встречного движения квазисолитонов возникновение их связанного состояния - двумерного биона. Амплитуда в центре биона регулярно осциллировала, лишь весьма незначительно уменьшаясь в течение счета (несколько периодов колебаний). Дальнейшее изучение показало, что аналогичные объекты могут возникать из недостаточно тяжелого односолитонного начального состояния. Поведение во времени и вид обнаруженных бионов качественно совпадают с пульсонами, открытыми ранее в работах /7/. Тем самым показано, что существование пульсонов не является привилегией систем с вырожденным вакуумом типа уравнений поля Хиггса и синус-Гордона, где полевая функция осциллирует между двумя смежными вакуумами. Заметим, что аналогичные пульсоны должны естественно возникать и в рамках системы (3). Возможное объяснение устойчивости обнаруженных пульсонов с помощью некоторого адиабатического инварианта можно найти в работах /8,9/.

---

*) Вспомним также задачу П.Л.Капицы о маятнике с качающейся точкой подвеса.

Отметим, что устойчивые связанные состояния из неустойчивых конституентов уже давно известны в ядерной физике (дейтрон). Так же как и в нашем случае, это состояние мало похоже на связанное состояние двух классических объектов типа Луна-Земля, двойные звезды и т.д. При формировании связанного состояния входящие в состав системы конституенты теряют свою индивидуальность фактически полностью. В этой связи уместно вспомнить широко распространенную в серидине нашего века поговорку "...природа сложна и нелинейные уравнения сложны, поэтому следует моделировать природу с помощью нелинейных уравнений", которую можно найти в книге Уилера /IO/.

Динамические свойства неодномерных солитонов обсуждаются также в работах /II, I2/. В /II/ исследовано взаимодействие цилиндрически-симметричных и сферически-симметричных ионоакустических солитонов с плоским и изогнутым фронтами. Показано, что взаимодействие таких солитонов, являющихся решением модифицированного уравнения Буссинеска, носит резонансный характер. В результате взаимодействия в области пересечения двух сталкивающихся солитонов возникает солитон большой амплитуды. Исследованию адронных свойств цилиндрически-симметричных солитонов в рамках моделей с полиномиальным потенциалом, а также полуцелым показателем нелинейности посвящена работа /I2/. Авторы указывают на возможное рождение пульсона ("брезера") в результате столкновения солитонов как с нулевым, так и с отличным от нуля прицельным параметром в рамках модели с дробным показателем нелинейности.

4. Ниже исследуется предложенная в /I3/ релятивистски-инвариантная система уравнений для скалярного комплексного поля Φ и вещественного поля η :

$$\Phi_{tt} - \Phi_{xx} - \Phi(1+\eta) = 0,$$
$$\eta_{tt} - \eta_{xx} + \nu\eta - g\eta^2 - |\Phi|^2 = 0, \qquad (4)$$

частные решения которой могут иметь как нулевые, так и ненулевые асимптотические значения при $x \to \infty$.

Система уравнений (4) может быть получена путем варьирования плотности лагранжиана

$$\mathcal{L} = |\Phi_t|^2 - |\Phi_x|^2 - |\Phi|^2 + \frac{1}{2}(\eta_t^2 - \eta_x^2 - \nu\eta^2 + \frac{2}{3}g\eta^3 + 2|\Phi|^2\eta). \qquad (5)$$

Плотность гамильтониана для этой системы имеет вид

$$\mathcal{H} = |\Phi_t|^2 + |\Phi_x|^2 + (1-\eta)|\Phi|^2 + \frac{1}{2}(\eta_t^2 + \eta_x^2 + \nu\eta^2 - \frac{2}{3}g\eta^3). \qquad (6)$$

Частные решения (4) в системе отсчета, связанной с солитоном, ищем в виде

$$\Phi(x,\tau) = \frac{1}{\sqrt{2}}\varphi(x)\,e^{-i\omega_0\tau}, \qquad (7)$$

$$\eta(x,\tau) = \eta(x).$$

В силу $U(1)$ симметрии модели (4) имеет место закон сохранения заряда для комплексного поля Φ :

$$Q = \omega_0 \int_{-\infty}^{\infty} \varphi^2 dx = const. \qquad (8)$$

Частные решения (4) имеют следующий вид:

а) при $\omega_0 = \sqrt{1-\nu}$

$$\Phi_1 = \frac{3\nu}{2}\sqrt{1-g}\,ch^{-2}\frac{\sqrt{\nu}}{2}x\,e^{-i\omega_0\tau},$$

$$\eta_1 = \frac{3\nu}{2}\,ch^{-2}\frac{\sqrt{\nu}}{2}x; \qquad (9)$$

б) при $\omega_0 = \sqrt{1-\frac{\nu}{6}}$, $g=3$

$$\Phi_2 = 2(1-\omega_0^2)\,ch^{-1}\sqrt{\omega_0^2-1}\,x\,e^{-i\omega_0\tau},$$

$$\eta_2 = 2(1-\omega_0^2)\,th^2\sqrt{\omega_0^2-1}\,x. \qquad (10)$$

5. Для нахождения области зарядовой устойчивости решения (9) используем теорему Маханькова /I4/ , согласно которой условие устойчивости имеет вид

$$\frac{\omega_0}{Q}\frac{dQ}{d\omega_0} < 0. \qquad (II)$$

С учетом (8) определим область устойчивости солитоноподобного решения (I0):

$$\frac{1}{2} < |\omega_0| < 1 \quad \text{или} \quad 0 < \nu < \frac{1}{2}\sqrt{3} \simeq 0{,}866.$$

Область устойчивости решения (9) может быть найдена также из условия минимума гамильтонина вида

$$\mathcal{H} = |\Phi_t|^2 + |\Phi_x|^2 + (1-\eta+\eta_{vac})|\Phi|^2 + \qquad (I2)$$

$$+\frac{1}{2}\left[\eta_t^2+\eta_x^2+\nu(\eta^2-\eta_{vac}^2)-\frac{2}{3}g(\eta^3-\eta_{vac}^3)\right],$$

где

$$\eta_{vac} = \lim_{x\to\infty}\eta_z = \frac{\nu}{g}.$$

Полагая $(\delta\eta, \delta\varphi) \propto ch^{-2}\frac{\sqrt{\nu}}{2}x$, найдем область устойчивости для φ $(0 < \nu < \frac{7}{18}(2-1/(1-g))$ и η $(0 < \nu < \frac{3}{4})$. Область устойчивости решения системы (4) определяется неравенствами

$$0 < \nu < 10/(10+3\sqrt{2(1-g)}), \quad g < (1-\sqrt{5})/4. \qquad (I3)$$

Найденная выше область устойчивости решения (9) показана пунктиром на рисунке.

Как видно из рис.3, область устойчивости решения (9), найденная из условия минимума гамильтониана, уже области устойчивости, полученной на основании Q -теоремы /I4/.

6. Методом численного эксперимента исследованы динамические свойства квазисолитонов (9) в лобовых столкновениях. Для перехода в лабораторную систему отсчета, движущуюся вдоль оси x со скоростью v (в единицах c), воспользуемся преобразованием Лоренца:

$$\varphi_1 = \frac{3\nu}{2}\sqrt{1-g}\;\mathrm{ch}^{-2}\left(\frac{\sqrt{\nu}}{g}\gamma(x-vt)\right)e^{i(\omega\gamma(vx-t)+\theta_1)}, \quad (I4)$$

$$\eta_1 = \frac{3\nu}{2}\mathrm{ch}^{-2}\left(\frac{\sqrt{\nu}}{g}\gamma(x-vt)\right).$$

Здесь $\omega = \sqrt{1-\nu}$, γ -релятивистский фактор.

В зависимости от частоты ν и постоянной g при нелинейном члене в (4) можно выделить три вида взаимодействий квазисолитонов (см.рис.3):

I) квазиупругое взаимодействие;
2) образование сингулярности поля в центре тяжести системы в момент перекрытия квазисолитонов;
3) образование сингулярности поля до взаимодействия квазисолитонов.

Отметим, что квазиупругое взаимодействие имеет место внутри области устойчивости отдельного квазисолитона, в то время как коллапс до взаимодействия (сингулярность поля) возникает на границе области устойчивости. В результате лобового столкновения квазисолитонов в области IУ (см.рис.3) квазисолитоны коллапсируют лишь в момент их перекрытия, когда возмущение каждого из них максимально. При сдвиге фаз $\Delta\theta = \pi$ упруго взаимодействующие квазисолитоны расталкиваются, в остальных случаях при $\Delta\theta \neq 0$ возникают пульсации амплитуд квазисолитонов, вследствие чего картина после взаимодействия становится асимметричной. При $\omega \to 1 \quad (v \to 0)$ область упругого взаимодействия квазисолитонов расширяется (область У на рис.3). Таким образом, "тяжелые" (с большой амплитудой) квазисолитоны (9) коллапсируют, а "легкие"- взаимодействуют упруго. Численный эксперимент показал, что картина взаимодействия квазисолитонов практически не зависит от скорости их встречного движения ($0{,}09 \leq v \leq 0{,}9$).

Виды взаимодействий квазисолитонов в модели (4) практически совпадают с обнаруженными ранее при исследовании взаимодействия пульсонов в рамках уравнения Клейна-Гордона с кубической нелинейностью [15]. Следовательно, квазисолитоны (10) являются объектами пульсонного типа. Они устойчивы лишь в ограниченной области значений параметров v и g [16].

Завершаются работы по исследованию свойств неодномерных (x, y, t)-солитонов в рамках двухполевой модели теории поля.

Автор считает своим приятным долгом выразить благодарность профессору В.Г.Маханькову за плодотворное сотрудничество и стимулирующие дискуссии в процессе выполнения этой работы.

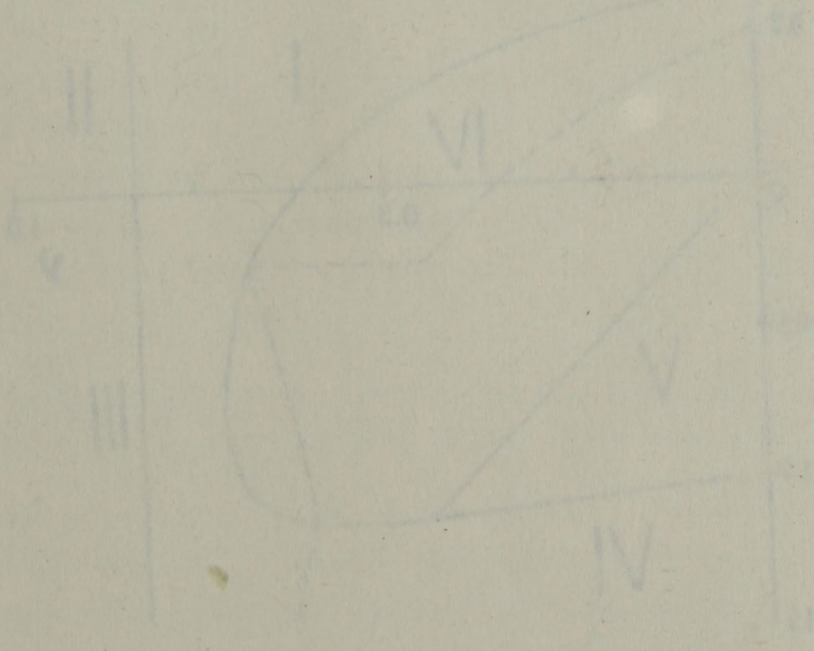
1. Маханьков В.Г.,Швачка А.Б. ОИЯИ, Р2-I304I, Дубна,I980;
Physica, 3D (1981) 1&2, 396.
2. Makhankov V.G. Phys.Reports, 35C (1978) 1-128.
3. Makhankov V.G., Kummer G., Shvachka A.B. Phys. Scripta, 20 (1979) 454;
4. Маханьков В.Г. и др.ОИЯИ, Р2-80-367, Дубна,I980;
Phys. Scripta, 23 (1981) 767.
5. Devi S., Strayer M., Irvine J. J.Phys.G: Nucl. Phys., 5 (1979) 281.
6. Маханьков В.Г.,Г.Куммер, А.Б.Швачка ОИЯИ, Р2-I042, Дубна, I980;
Physica 3D (1981) 1&2, 344.
7. Боголюбский И.Л.,Маханьков В.Г. Письма в ЖЭТФ, 24 (I976) I2;
8. Makhankov V.G. Phys. Scripta, 20 (1979) 558;
9. Манаков С.В. Письма в ЖЭТФ, 25 (I977) 589.
10. Уиллер Д.А. В кн.:Гравитация, нейтрино и вселенная. ИЛ, М., I962.
11. Kako F. and Yajima N. Preprint IPPJ-535,Inst. of Plasma Phys., Nagoya Univ., Nagoya, 1981.
12. Drohm J.K., Kok L.P., Simonov Yu.A., Tjon J.A and Veselov A.I. Phys. Lett., 101B (1981) 204.
13. Катышев Ю.В., Маханьков В.Г. ОИЯИ, ДIO,II-II264, Дубна,I978.
14. Маханьков В.Г. ОИЯИ, Р2-I0362, Дубна,I977.
15. Bogolubsky I.L., Makhankov V.G. and Shvachka A.B. Phys. Lett., 63A (1977) 225.
16. Sautbekov S.S. and Shvachka A.B. JINR, E2-82-413, Dubna 1982.

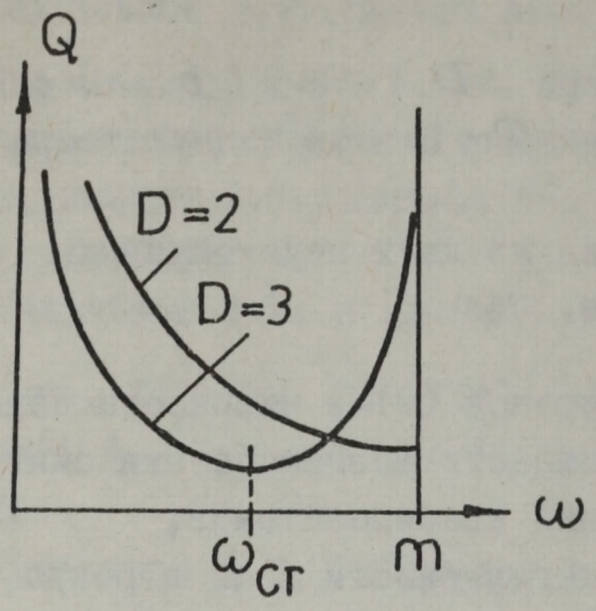
ПОДПИСИ К РИСУНКАМ

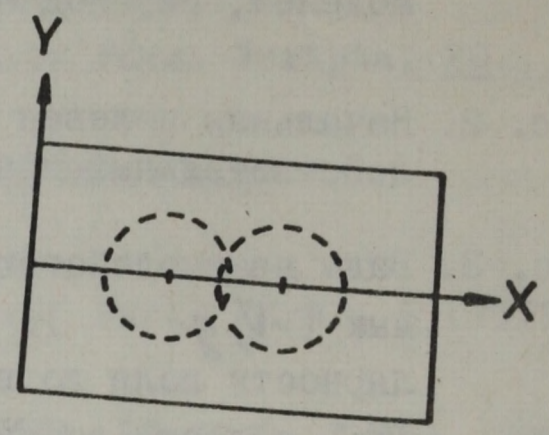
Рис. I. Графики зависимости $Q(\omega)$ при $D = 2$ и $D = 3$ для моделей, переходящих в пределе $\varphi \to 0$ в свободную теорию.

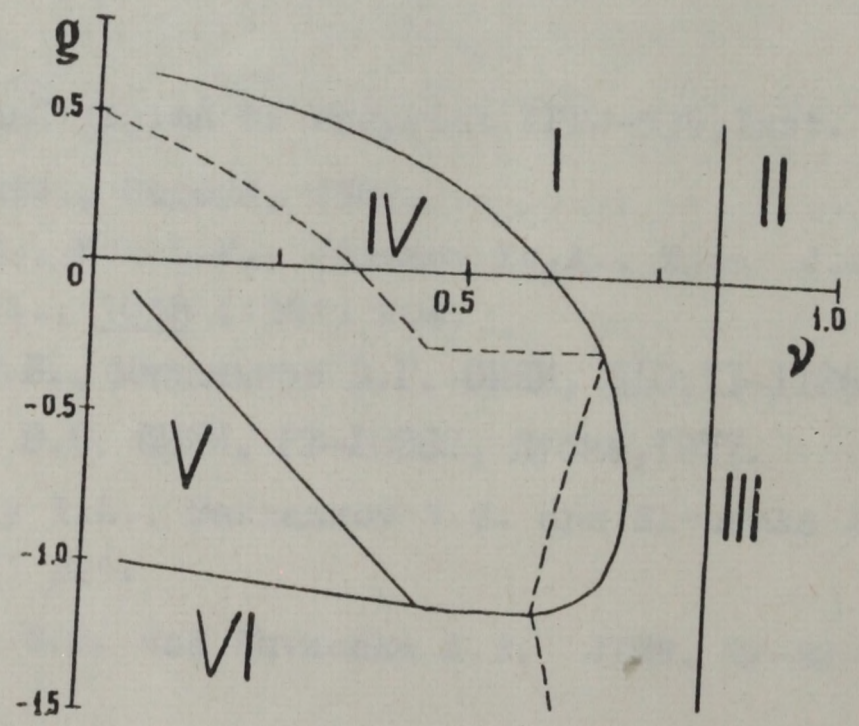
Рис. 2. Начальная полевая конфигурация из двух неустойчивых действительных квазисолитонов.

Рис. 3. Виды взаимодействий квазисолитонов (9) в плоскости переменных (ν, g). I–Ш, УI – области возникновения сингулярности поля до взаимодействия квазисолитонов;
IУ – область возникновения сингулярности поля в результате взаимодействия квазисолитонов;
У – область квазиупругого взаимодействия.




Q
D=2
D=3
ω
ω_{cr}
m


Y
X


g
0.5
0
-0.5
-1.0
-1.5
0.5
1.0
ν
I
II
III
IV
V
VI

GEOMETRIC CONVERGENCE OF SOME TWO-POINT PADÉ APPROXIMATIONS

G. NÉMETH

Central Research Institute for Physics
H-1525 Budapest 114, P.O.Box 49, Hungary

ABSTRACT

In this paper the geometric convergence of some two-point Padé approximations on certain infinite sets of the complex plane is considered.

АННОТАЦИЯ

В данной работе исследуется вопрос геометрической сходимости специальных приближений Паде на неограниченных областях комплексной плоскости.

I. INTRODUCTION

The main aims of this paper are to investigate the convergence of some two-point Padé approximations on certain infinite sets of the complex plane. The convergence of Padé approximations has received much interest, both for its application in numerical computations and for approximation theory problems.

In particular, we consider the function

$$F(x) = \mu \int_0^1 (1-u)^{\mu-1} e^{-ux} du \quad , \qquad \mu > 0 \quad . \tag{1}$$

For $\mu = \frac{1}{2}$ this function is the subject of numerical calculations connected with the plasma dispersion function [1]-[3]

$$Z(s) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{+\infty} \frac{e^{-t^2}}{t-s} dt = i\sqrt{\pi} e^{-s^2} - 2e^{-s^2} \int_0^s e^{u^2} du \quad . \tag{2}$$

The case $\mu = 0$ (the exponential function) was considered by Saff et al. in their excellent papers [4]-[6].

After the development of some preliminary considerations in Section 2, we consider, in Section 3, the convergence of two-point Padé approximations to the function F(x) on the real positive axis. We shall prove that these rational approximants of $R_k(x)$ type

$$R_k(x) = \frac{p_o + p_1 x + \ldots + p_{k-1} x^{k-1}}{1 + q_1 x + \ldots + q_k x^k} \quad , \tag{3}$$

have a geometric convergence rate as of at least $\frac{1}{2^k}$: Theorem 2. In Theorem 3 we establish that the best generalized two-point Padé approximations have a geometric convergence rate like $\frac{1}{3^k}$. In Section 4 we consider some infinite parabolic-type domains of the complex plane in which the geometric convergence of two-point Padé approximations also holds. Our results, Theorem 4, is an application of the results of Saff and Varga [11]. In Theorem 5 we present infinite sectors of the complex plane in which the special generalized two-point Padé approximations converge in geometric order.

II. DEFINITIONS AND PRELIMINARY RESULTS

When a function f(x) satisfies the conditions

$$f(x) \sim \sum_{k=0}^{\infty} c_k x^k \quad , \quad x \to 0 \quad ,$$
$$f(x) \sim \sum_{k=0}^{\infty} d_k x^{-k-1} \quad , \quad x \to \infty \quad , \tag{4}$$

we can determine rational fractions $R_k(x)$, (3), for which the following relations hold

$$f(x) - R_k(x) = O(x^k) \quad , \quad x \to 0 \quad ,$$
$$f(x) - R_k(x) = O(x^{-k-1}) \quad , \quad x \to \infty \quad . \tag{5}$$

Definition 1. The rationals $R_k(x)$ satisfying both previous conditions are called two-point Padé approximations to the function f(x). There exists a more general conception of this definition.

Definition 2. The rationals $R_k^{(m)}(x)$

$$R_k^{(m)}(x) = \frac{p_o^{(m)} + p_1^{(m)} x + \ldots + p_{k-1}^{(m)} x^{k-1}}{1 + q_1^{(m)} x + \ldots + q_k^{(m)} x^k} \tag{6}$$

satisfying the conditions

$$f(x) - R_k^{(m)}(x) = O(x^{k+m}) \quad , \quad x \to 0 \quad ,$$
$$f(x) - R_k^{(m)}(x) = O(x^{-k+m-1}) \quad , \quad x \to \infty \quad , \tag{7}$$

where m is a positive integer $m = 0,1,2,\ldots,k$, we call generalized two-point Padé approximations to the function f(x). The reason for this generalization is obvious: we take k+m terms from the series near t = 0, and k-m terms from the series near $t = \infty$ to calculate the coefficients of the rational $R_k^{(m)}(x)$. Let us mention that the case m=0 corresponds to Definition 1 and that m=k is the classic (one-point) Padé approximation. For our function F(x) we can solve exactly the problem of generalized two-point Padé approximation in closed form.

Theorem 1. For the generalized two-point Padé approximations to the function F(x) the following results hold:

(i) the denominator of the rationals

$$R_k^{(m)}(x) = \frac{P_k^{(m)}(x)}{Q_k^{(m)}(x)} \quad ,$$

$Q_k^{(m)}(x)$, in hypergeometric notation, is

$$Q_k^{(m)}(x) = {}_1\mathcal{F}_1(-k; 1-\mu-m-k; x) \quad ,$$

(ii) the numerator of the error term

$$E_k^{(m)}(x) = F(x) - R_k^{(m)}(x) = \frac{S_k^{(m)}(x)}{Q_k^{(m)}(x)} ,$$

$S_k^{(m)}(x)$, in integral form, is

$$S_k^{(m)}(x) = (-1)^m \frac{\Gamma(1+\mu)}{\Gamma(m+k+\mu)} x^{k+m} \int_0^1 e^{-xu} u^k (1-u)^{m+\mu-1} du ,$$

(iii) the functions $P_k^{(m)}(x)$, $Q_k^{(m)}(x)$ (and $S_k^{(m)}(x)$ too) satisfy the second order difference equations with respect to k

$$(k+m+\mu-1)(k+m+\mu) y_{k+1} = (k+m+\mu-1)(k+m+\mu+x) y_k - kxy_{k-1} , \quad k=1,2,\ldots ,$$

(iv) the error function has a more economic representation:

$$E_k^{(m)}(x) = (-1)^m \Gamma(1+\mu)\Gamma(m+\mu) \sum_{j=k}^{\infty} \frac{j! t^{j+m}}{\Gamma(j+m+\mu)} \frac{1}{\Gamma(j+m+\mu+1) Q_j^{(m)}(x) Q_{j+1}^{(m)}(x)} .$$

Proofs. First we mention that the function F(x) has the series representations

$$F(x) = \sum_{k=0}^{\infty} \frac{(-x)^k}{(1+\mu)_k} , \quad x \to 0 ,$$

$$\mu > 0 ,$$

$$F(x) \sim \mu \sum_{k=0}^{\infty} (1-\mu)_k x^{-k-1} , \quad x \to \infty .$$

The coefficients of the rationals $R_k^{(m)}(x)$ are determined (corresponding to Definition 2) from the equations

$$\sum_{j=0}^{\ell} \frac{(-1)^j}{(1+\mu)_j} q_{\ell-j} = p_\ell , \quad \ell=0,1,\ldots,k-1,$$

$$\sum_{j=0}^{k} q_j \frac{(-1)^{i-j}}{(1+\mu)_{i-j}} = 0 , \quad i=k,k+1,\ldots,k+m-1,$$

$$\mu \sum_{j=0}^{i} (1-\mu)_{i-j} q_{k-j} = p_{k-i-1} , \quad i=0,1,\ldots,k-m-1 .$$

This is a system of 2k simultaneous equations in 2k unknowns: $p_0, p_1, \ldots, p_{k-1}$, $q_1, q_2, \ldots, q_k$. We determine explicitly the q_k numbers only. When we eliminate the numbers p_k we get the system:

$$\sum_{j=0}^{\ell} \frac{(-1)^j}{(1+\mu)_j} q_{\ell-j} = \mu \sum_{j=0}^{k-1-\ell} (1-\mu)_{k-1-\ell-j} q_{k-j} , \quad \ell=m,m+1,\ldots,k-1,$$

$$\sum_{j=0}^{k} q_j \frac{(-1)^{\ell-j}}{(1+\mu)_{\ell-j}} = 0 , \quad \ell=k,k+1,\ldots,k+m-1 ,$$

or in simpler form

$$\sum_{j=0}^{k} q_j \frac{(-1)^{\ell-j}}{(1+\mu)_{\ell-j}} = 0 , \quad \ell=m,m+1,\ldots,m+k-1 .$$

By elementary manipulations we can transform this system to

$$\sum_{j=0}^{k} (j+1)_{\ell}\Gamma(j-k-m-\mu+1)q_j = 0 \quad , \qquad \ell=0,1,\ldots k-1 \quad .$$

This system we solve by the orthogonal polynomial method:

$$\int_0^{\infty} e^{-u} u^{\ell} \sum_{j=0}^{k} q_j u^j \frac{\Gamma(j-k-m-\mu+1)}{j!} du = 0 \quad , \qquad \ell=0,1,2,\ldots,k-1 \quad ,$$

and

$$\sum_{j=0}^{k} q_j u^j \frac{\Gamma(j-k-m-\mu+1)}{j!} = \frac{\Gamma(-k-m-\mu+1)}{k!} e^u \frac{d^k}{du^k}(e^{-u}u^k) \quad ,$$

therefore

$$q_j = \frac{(-k)_j}{j!(1-\mu-m-k)_j} \quad , \qquad j=0,1,\ldots,k \quad .$$

Now (i) is proved. Next, to get $S_k^{(m)}(x)$ we must compute the series

$$S_k^{(m)}(x) = \sum_{j=k+m}^{\infty} x^j \sum_{\ell=0}^{k} q_\ell \frac{(-1)^{j-\ell}}{(1+\mu)_{j-\ell}} =$$

$$= (-1)^m \Gamma(1+\mu) \sum_{j=0}^{\infty} (-1)^j x^{j+m+k} \frac{\Gamma(m+\mu)}{\Gamma(k+m+\mu)\Gamma(j+m+\mu+1)} \sum_{\ell=0}^{k} \frac{(-k)_\ell}{\ell} \cdot$$

$$\cdot \frac{(m+\mu)_\ell}{(j+m+\mu+1)_\ell} = (-1)^m k! x^{k+m} \frac{\Gamma(1+\mu)\Gamma(m+\mu)}{\Gamma(k+m+\mu)\Gamma(k+m+\mu+1)} \cdot$$

$$\cdot \sum_{j=0}^{\infty} \frac{(k+1)_j}{j!(k+m+\mu+1)_j} (-x)^j \quad .$$

This is a hypergeometric function (${}_1\mathcal{F}_1$ type). It is not difficult to see [7] that is satisfies the difference equation (iii). The function $Q_k^{(m)}(x)$ also satisfies (iii) and therefore $P_k^{(m)}(x)$ is the solution of the same equation. Thus (iii) is proved. When we apply the usual integral representation [8] to this function we get (ii). Finally we prove (iv). Let us consider the difference

$$\frac{S_{k+1}^{(m)}(x)}{Q_{k+1}^{(m)}(x)} - \frac{S_k^{(m)}(x)}{Q_k^{(m)}(x)} = \frac{H_k(x)}{Q_k^{(m)}(x) Q_{k+1}^{(m)}(x)} \quad .$$

Applying (iii) we arrive at the difference equation for H_k:

$$H_k(x) = \frac{kx}{(k+m+\mu)(k+m+\mu-1)} H_{k-1}(x)$$

from which

$$H_k(x) = (-1)^m \frac{\Gamma(1+\mu)\Gamma(m+\mu)}{\Gamma(k+m+\mu)\Gamma(k+m+\mu+1)} x^{k+m} \quad .$$

Summing the previous difference relation we get formula (iv).

III. NEW RESULTS ON GEOMETRIC CONVERGENCE

With the aid of the results of the previous sections, we now establish the convergence of generalized two-point Padé approximations to the function F(x). First we deal with the parameter m having a bounded, fixed value.

Theorem 2. For the maximum value of the error function

$$E_k^{(m)} = \max_{0 \le x < \infty} |E_k^{(m)}(x)|$$

for $m+\mu > 1$ the following estimation holds

$$E_k^{(m)} < \frac{1}{2} E_{k-1}^{(m)} \quad , \qquad k=m,m+1,\ldots, \tag{8}$$

Proof. From (ii) it is not difficult to check that $E_k^{(m)}(0) = E_k^{(m)}(\infty) = 0$, and thus undoubtedly there exists a positive value x_k where

$$E_k^{(m)} = |E_k^{(m)}(x_k)| \quad ,$$

and it naturally holds that $\frac{d}{dx} E_k^{(m)}(x_k) = 0$. When we differentiate the integral form of the error function we get

$$0 = -\frac{\mu}{x_k} E_k^{(m)}(x_k) - \frac{k}{k+m+\mu-1} \frac{Q_{k-1}^{(m)}(x_k)}{Q_k^{(m)}(x_k)} E_k^{(m)}(x_k) - E_k^{(m)}(x_k) +$$

$$+ \frac{k}{k+m+\mu-1} \frac{Q_{k-1}^{(m)}(x_k)}{Q_k^{(m)}(x_k)} E_{k-1}^{(m)}(x_k) \quad .$$

In a more compact form this is

$$E_k^{(m)}(x_k) = \frac{kx_k Q_{k-1}^{(m)}(x_k)}{kx_k Q_{k-1}^{(m)}(x_k) + (k+m+\mu-1)(x_k+\mu) Q_k^{(m)}(x_k)} E_{k-1}^{(m)}(x_k)$$

Next we show that for $0 \le x \le \infty$ and $m+\mu > 1$

$$Q_k^{(m)}(x) \ge Q_{k-1}^{(m)}(x) \quad , \qquad k=1,2,\ldots \quad .$$

A short comparison of the coefficients of the same powers in the polynomials shows

$$\frac{(-k)_j}{j!(1-\mu-m-k)_j} \ge \frac{(-k+1)_j}{j!(2-\mu-m-k)_j} \quad , \qquad j=0,1,\ldots,k-1 \quad ,$$

because of trivial inequality

$$\frac{k}{k+m+\mu-1} \geq \frac{k-j}{k+m+\mu-1-j} \quad .$$

Now applying the inequality and the fact that $E_{k-1}^{(m)}(x_k) \leq E_{k-1}^{(m)}$ we get

$$E_k^{(m)} \leq \frac{kx_k Q_{k-1}^{(m)}(x_k)}{kx_k Q_{k-1}^{(m)}(x_k) + (k+m+\mu-1)(x_k+\mu) Q_k^{(m)}(x_k)} E_{k-1}^{(m)} \leq$$

$$\leq \frac{kx_k}{kx_k + (k+m+\mu-1)(x_k+\mu)} E_{k-1}^{(m)} \leq \frac{k}{2k+m+\mu-1} E_{k-1}^{(m)} < \frac{1}{2} E_{k-1}^{(m)} \quad .$$

Theorem 2 is now proved.

Next we shall show that there exists an optimal choice of parameter m in connection with the convergence rate of the generalized two-point Padé approximations to the function F(x). We treat the case when the parameter $m \to \infty$, if $k \to \infty$, in a suitable manner.

Theorem 3. Let us suppose $m \to \infty$,

$$\lim_{k\to\infty} \frac{m}{k} = \beta \quad , \quad 0 \leq \beta \leq 1 \quad , \tag{9}$$

and $\mu > 0$, then the generalized two-point Padé approximations to the function F(x) have geometric convergence rate

$$\lim_{k\to\infty} \{E_k^{(m)}\}^{\frac{1}{k}} = \varphi(\beta) = \beta^{\beta}(1-\beta)^{1-\beta} 2^{\beta-1} < 1 \quad . \tag{10}$$

Before proving this we would comment on our result. From the form of the function $\varphi(\beta)$ one can see that for $\beta = 0$ (two-point Padé approximation), the geometric convergence rate is $\frac{1}{2}$; for $\beta = 1$ (classic Padé approximation), $\varphi(1)=1$: geometric convergence does not exist; for $\beta = \frac{1}{3}$ (this is the minimal position of the function $\varphi(\beta)$), the geometric convergence rate is $\frac{1}{3}$. In view of this, we can state that the optimal choice of parameter m is $m = [\frac{k}{3}]$ with regard to the convergence rate of the generalized two-point Padé approximations to the function F(x).

Proof. We shall apply formula (iv) and Lemma 1 to investigate the function $L_\ell(x)$ in the error function

$$E_k^{(m)}(x) = (-1)^m \Gamma(1+\mu) \sum_{\ell=k}^{\infty} L_\ell(x) \quad ;$$

$$L_\ell(x) = \frac{\Gamma(m+\mu)\, \ell!\, x^{\ell+m}}{\Gamma(\ell+m+\mu)\Gamma(\ell+m+\mu+1) Q_\ell^{(m)}(x) Q_{\ell+1}^{(m)}(x)} \quad .$$

Lemma 1. Let us suppose that $\mu > 0$, $\ell \to \infty$, $m \to \infty$ and

$$\lim_{\ell\to\infty} \frac{m}{\ell} = \beta \quad , \quad 0 \leq \beta \leq 1 \quad ,$$

then

$$\lim_{\ell\to\infty}\{\max_{0\le x\le\infty}|L_\ell(x)|\}^{\frac{1}{\ell}}=\varphi(\beta)\quad . \tag{11}$$

Proof. First we determine the asymptotic approximation of the denominator polynomial. We apply its integral representation:

$$\Gamma(\ell+m+\mu)Q_\ell^{(m)}(x)=\int_0^\infty u^{m+\mu-1}(x+u)^\ell e^{-u}du\quad .$$

Taking $m=\beta\ell$, $x=\alpha\ell$ we use Laplace's method to obtain the main term of the integral for $\ell\to\infty$. The main contribution comes from the neighbourhood of the point $u_o=\ell s$ where s is the root of equation

$$-1+\frac{1}{s+\alpha}+\frac{\beta}{s}=0\quad .$$

In the usual manner of doing the calculations the main term is

$$\int_0^\infty u^{\beta\ell+\mu-1}(\alpha\ell+u)^\ell e^{-u}du\sim a\cdot\ell^b\cdot\exp\{\ell(-s+\ln(s+\alpha)+\beta\ln s)+(1+\beta)\ell\ln\ell\}\ ,$$

$$\ell\to\infty\ ,$$

where a and b are constans independent of ℓ. With the aid of this result we get an asymptotic representation of $L_\ell(x)$ for $\ell\to\infty$:

$$L_\ell(x)\sim a^*\ell^{b^*}\cdot\exp\{\ell(-1-\beta+(1+\beta)\ln\alpha+\beta\ln\beta+2s-2\ln(s+\alpha)-2\beta\ln\beta)\}\quad .$$

Because $L_\ell(0)=L_\ell(\infty)=0$ the function $L_\ell(x)$ has its maximum value where $\frac{d}{dx}L_\ell(x)=0$ or $\frac{d}{d\alpha}L_\ell(x)=0$. By differentiating the main term of $L_\ell(x)$ we obtain the equation

$$\frac{1+\beta}{\alpha}-\frac{2}{s+\alpha}=0$$

and therefore we can solve the equations for α and s explicitly:

$$\alpha=\frac{(1+\beta)^2}{2(1-\beta)}\ ,\qquad s=\frac{1}{2}(1+\beta)\quad .$$

Eliminating α and s in the asymptotic expression of $L_\ell(x)$ we get the required result.

Returning to the proof of Theorem 3 the following estimations are obvious

$$\Gamma(1+\mu)L_k(x)<|E_k^{(m)}(x)|\le\Gamma(1+\mu)(L_k(x)+L_{k+1}(x)+\ldots)\quad .$$

Here, when we raise this inequality to the k^{-1}-th power, then on letting $k\to\infty$ we obtain the statement of Theorem 3.

IV. CONVERGENCE ON THE COMPLEX PLANE

In the previous section we considered convergence on the real positive axis. Here we shall be concerned with the convergence on unbounded domains of the complex plane that are symmetric with respect to the real positive axis. Such an extension of the convergence to larger domains of the complex plane "overconvergence problem" very much depends on the knowledge of the location of the poles of the two-point Padé approximations to $F(z)$. It is clear from formula (ii) that the poles of these approximants are the zeros of polynomial $Q_k^{(m)}(z)$. Our next results come from investigations of the location of the zeros for the polynomial $Q_k^{(m)}(z)$.

First of all we show that the convergence of the generalized two-point Padé approximation holds for any bounded domain of the complex plane.

From representation (i), by the Theorem of Tannery [9], it follows that

$$Q_k^{(m)}(z) \to \exp\{z/(1+\beta)\} \ , \quad k \to \infty \ , \quad |z| < K = \text{const.} \ , \tag{12}$$

and

$$\lim_{k\to\infty} \frac{m}{k} = \beta \quad (\text{where } \beta = 0, \text{ when } m \text{ is bounded}).$$

As a consequence of this result all zeros of the polynomial $Q_k^{(m)}(z)$ tend to infinity when $k \to \infty$.

Another consequence is that the rationals $R_k^{(m)}(z)$ converge to $F(z)$ faster than geometrically

$$\lim_{k\to\infty} \{ \max_{|z| < K = \text{const.}} |F(z) - R_k^{(m)}(z)| \}^{\frac{1}{k}} = 0 \ . \tag{13}$$

This follows easily from the integral representation (ii) of the error.

Next we shall consider the convergence problem in parabolic type unbounded domain of the complex plane. We state another result on the location of the poles of the rational $R_k^{(m)}(z)$.

Lemma 2. The polynomials $Q_k^{(m)}(z)$ have no zeros in the parabolic domain

$$S = \{z=x+iy \in C; \ y^2 < 4(m+\mu)(x+m+\mu)\} \ . \tag{14}$$

Proof. This statement immediately follows from a Theorem of Henrici [10], when we use the identifications $z_k = z$, $\beta_k = k+m+\mu-1$, $\varepsilon_k = k-1$, $q_k = \Gamma(k+m+\mu-1) Q_k^{(m)}(z)$, $k=1,2,\ldots$, $\alpha = m+\mu$.

Now we define the parabolic type unbouded domains

$$S_r = \{z=x+iy \in C; \ y^2 < 4r(m+\mu)(x+m+\mu)\} \ . \tag{15}$$

The following theorem gives the estimation of the convergence rate to the $R_k^{(m)}(z)$ in S_r.

Theorem 4. Let us suppose, for the number r, that

$$0 < r < 3-2\sqrt{2} \quad , \tag{16}$$

holds; then the rationals $R_k^{(m)}(z)$ converge to $F(z)$ in the domain S_r with the geometric convergence rate

$$\lim_{k\to\infty} \{\max_{z\in S_r} |F(z)-R_k^{(m)}(z)|\}^{\frac{1}{k}} < \frac{1}{2}\left(\frac{1+r}{1-r}\right)^2 < 1 \quad . \tag{17}$$

Proof. We can apply a general Theorem of Saff and Varga [11]. For our special result we need the identifications $q = 2$, $r_k = Q_k^{(m)}(z)$. Their exceptional bounded subset K_d is missing here, i.e. we proved in previous considerations that on every bounded set stronger than geometric convergence holds.
Finally we consider the problem of convergence on unbounded sectors

$$W = \{z=x+iy\in C,\ |\arg z| < \theta\} \quad . \tag{18}$$

When m has a finite value there exists no infinite sector of this type which is devoid of zeros of $Q_k^{(m)}(z)$, $k=1,2,\ldots$; consequently, there is no infinite sector in which the geometric convergence of the rationals $R_k^{(m)}(z)$ can hold. But when we consider the polynomial $Q_k^{[\beta k]}(z)$, $k=1,2,\ldots$ such a sector does exist nevertheless.

Lemma 3. For $k=1,2,\ldots$ the polynomial $Q_k^{[\beta k]}(z)$ has no zeros in the infinite sector

$$W_\beta=\{z=x+iy\in C,\ |\arg z|<\arccos\frac{1-\beta}{1+\beta}\} \quad , \quad 0 < \beta < 1 \quad . \tag{19}$$

Proof. We can apply a Theorem of Saff and Varga [12]. Instead of their ν (which is an integer value) must take $\beta k+\mu-1$. In this case the (rather long) proof is easy, therefore we omit it for the sake of brevity.
The following result gives the estimation of the convergence rate of $R_k^{[\beta k]}(z)$ in the infinite sector W.

Theorem 5. Let us suppose that for $\theta_o = \arccos\frac{1-\beta}{1+\beta}$, $0<\beta<1$ the sector W_β contains no poles of $R_k^{[\beta k]}(z)$, then for every θ satisfying the inequality

$$0 < \theta < 4\arctan\left[\frac{1-\sqrt{\varphi(\beta)}}{1+\sqrt{\varphi(\beta)}}\cdot\tan\frac{\theta_o}{4}\right] \quad , \tag{20}$$

the rationals $R_k^{[\beta k]}(z)$ converge to $F(z)$ in the infinite sector W with the geometric convergence rate

$$\lim_{k\to\infty}\{\max_{z\in W}|F(z)-R_k^{[\beta k]}(z)|\}^{\frac{1}{k}} < \varphi(\beta)\left\{\frac{\sin\frac{1}{4}(\theta_o+\theta)}{\sin\frac{1}{4}(\theta_o-\theta)}\right\}^2 < 1 \quad . \tag{21}$$

Proof. We can apply a general Theorem of Saff and Varga [11]. For our special result we need the identifications $q = \frac{1}{\varphi(\beta)}$, $r_k = R_k^{[\beta k]}(z)$, $\theta_o = \arccos\frac{1-\beta}{1+\beta}$. The cited author's exceptional part $|z| \leq \mu$ of the sector is missing here

because of its boundedness (on the bounded domain the stronger than geometric convergence holds).

REFERENCES

[1] P. Martin, G. Donoso, I. Zamaudi-Cristi: A Modified Asymptotic Padé Method. Application to Multipole Approximation for the Plasma Dispersion Function Z
J. Math. of Phys., 21 (1980) p.280

[2] B.D. Fried, C.L. Hedrick, I. McCune: Two-Pole Approximation for the Plasma Dispersion Function
Phys. Fluids, 11 (1968) p.247

[3] G. Németh, Á. Ág, Gy. Páris: Two-Sided Padé Approximations for the Plasma Dispersion Function
J. Math. of Phys. (to be published)

[4] E.B. Saff, R.S. Varga: Convergence of Padé Approximants to e^{-z} on Unbounded Sets
J. Approx. Theory, 13 (1975) p.470

[5] E.B. Saff, R.S. Varga: On the Zeros and Poles of Padé Approximants to e^{-z} I,II,III
Numer. Math., 25 (1975) p.1; Conf. on Rat. Appr., Tampa, USA (1976) p.195; Numer. Math., 30 (1978) p.241

[6] E.B. Saff, R.S. Varga, W.C. Ni: Geometric Convergence of Rational Approximations to e^{-z} in Infinite Sectors
Numer. Math., 26 (1976) p.211

[7] M. Abramovitz, I.A. Stegun: Handbook of Mathematical Functions
Nat. Bur. of Stand. Applied
Math. Series 55, Washington (1967) p.506

[8] A. Erdélyi et al.: Higher Transcendental Functions I.
McGraw-Hill, New York, N.Y. (1953)

[9] P. Szász: Introduction to the Analysis, II. (In Hungarian)

[10] P. Henrici: Note on a Theorem of Saff and Varga
Conference on Rational Approximations
(Eds. E.B. Saff, R.S. Varga) Tampa, USA (1976) p.157

[11] E.B. Saff, R.S. Varga: Geometric Overconvergence of Rational Functions in Unbounded Domain
Pacific J. of Math., 62 (1976) p.523

[12] See [5] Theorem 2.1, p.3

UNIFIRM - A COST-EFFECTIVE UNIVERSAL MICROPROGRAM DEVELOPMENT SYSTEM BASED ON PDP-11/LSI-11 COMPUTERS

M. SALAMON, G. BALATONI, G. LŐRINCZE, M. NAGY, J. TIBOR

Central Research Institute for Physics
H-1525 Budapest 114, P.O.Box 49, Hungary

ABSTRACT

Even bulky microprogram development software can be run on small LSI-11 systems, moderately equipped with memory and backup store, if the LSI-11's are connected on-line to a mid-range PDP-11 - this is the basis of the UNIFIRM architecture. The computers in the UNIFIRM system are arranged in star network configuration. The single-user LSI-11's are located at the development sites and connected to the central multi-user PDP-11 system via telephone lines.

The increasing usage of microprocessors requires the help of intelligent hardware - firmware - software systems, which can effectively assist the work of the engineers and programmers. The great microprocessor manufacturers try to offer such systems, but these are not suitable or attainable in every case. We must take into account the fact that microprogramming is a fairly sophisticated job and it postulates and demands the knowledge of the hardware.

A microprogram translator is always needed for the work. Generally it is a cross-assembler, which can be run on a relatively big computer configuration. After the compilation the microprogram object code is loaded into the memory of the device being developed, then follows the debugging of the microprogram. Either the loading or the debugging do not require the usage of a big computer; a simpler, intelligent system is quite sufficent. It is not negligible that sometimes the hardware does not still operate perfectly in this phase of the development, and this fact increases the difficulties. Therefore it is advisable to use such a simple system, which aids the hardware trouble-shooting as well. A prerequisite of the effective work is that all the software components of the support system should be on-line accessible at any time /e.g. if the cross-assembler can only be used in batch environment, the whole work will be far complicated/.

We have developed the UNIFIRM microprogram development system, which helps in solving the problems described above.

The UNIFIRM supports the developing of the hardware and firmware of devices built of different types of microprocessors /bit-slice, Z80, 8080/. A number of working stations can be connected into the system. /A working station is a room or laboratory, where hardware and firmware developments are done./ One of our main purposes was the economical usage of the UNIFIRM components, i.e. we do not want to occupy unnecessarily any component of the system.

The cross-assemblers require large memory and backup store. They run under the RSX-11M operating system on a mid-range PDP-11 minicomputer. This mid-range computer can be simultaneously used for any other purposes, since the microprogram translation does not intensively load the machine. Each working station is built around an LSI-11 microcomputer which is connected to the central mid-range PDP-11 computer via telephon lines. The RT-11 operating system runs on the LSI-11 configuration, and the RSX-11M system runs on the mid-range PDP-11 computer. The connection between the PDP-11 and the LSI-11 processors is controlled by the Terminal Communication Program /TCP/. By the help of the TCP the users can access all the facilities provided by the RSX-11M system and control the file transfers between the backup stores of the configurations as well. The LSI-11 processor must support the last phases of the microprogram development Figure 1. illustrates a possible UNIFIRM system.

The hardware and software components of the UNIFIRM, and their tasks are as follows:

Component	Task
PDP-11	The cross-assemblers run on this computer under the RSX-11M operating system.

Component	Task
LSI-11	Software support at the working station. Loading and debugging the microprogram. Access to the PDP-11 computer by the help of the TCP program.
MCA CPX ZPX	Cross-assemblers which run under the RSX-11M operating system. The MCA program can be used for bit-slice based devices. The CPX generates code for the I 8080, the ZPX generates code for the Z80 microprocessor.

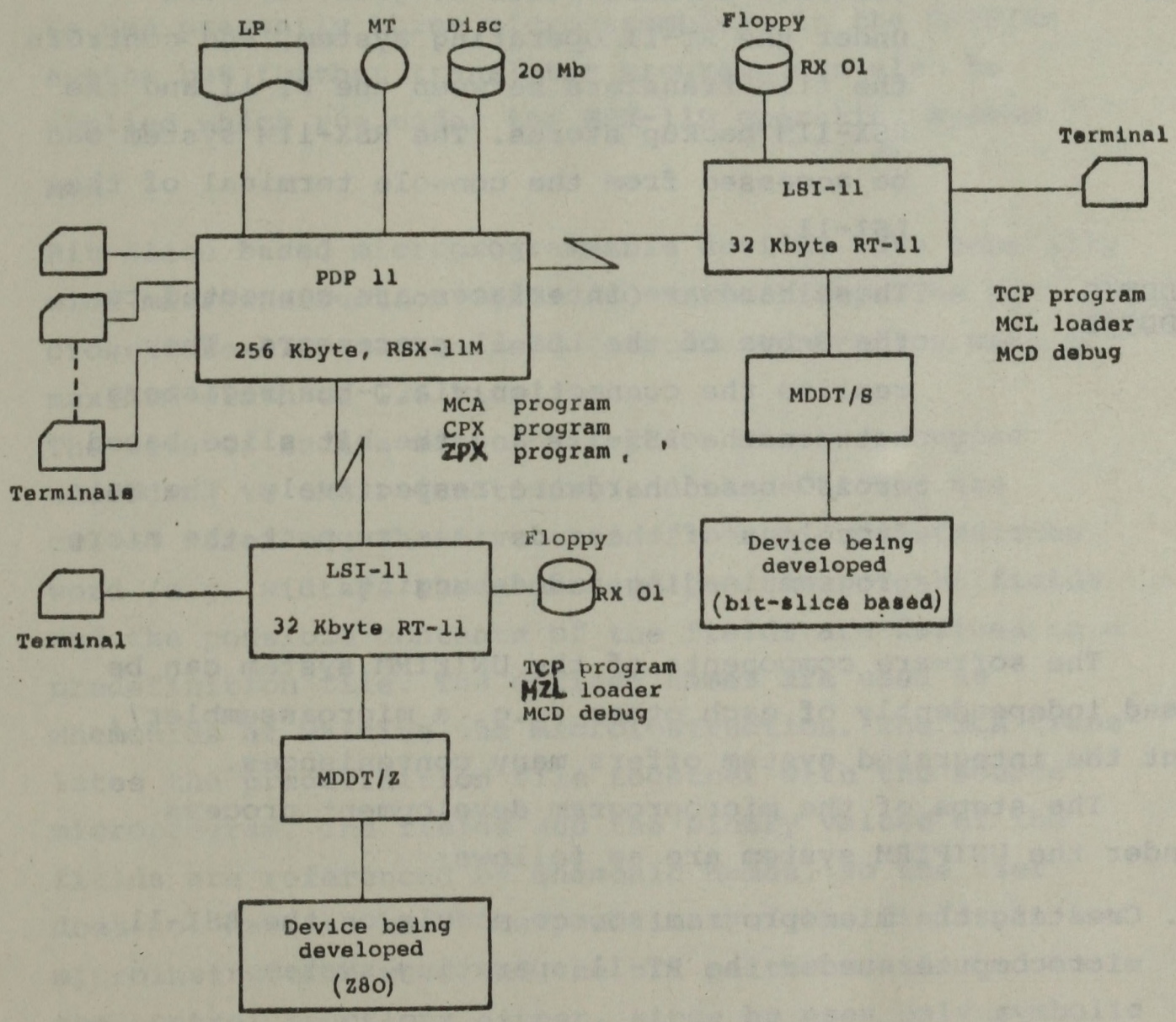


Figure 1. UNIFIRM system

Component	Task
MCL	Loader programs under the RT-11 operating system.
MZL	These programs can load the microprogram object code into the control memory from the backup store of the LSI-11s.
MCD	Debug programs under the RT-11 operating system.
MZD	The usual debug functions are provided by them /examine and deposit micromemory locations, breakpoints, start, microstep, etc./
TCP	Terminal Communication Program. It runs under the RT-11 operating system, and controls the file transfers between the RT-11 and the RSX-11M backup stores. The RSX-11M system can be accessed from the console terminal of the LSI-11.
MDDT/S MDDT/Z	These hardware interfaces are connected to the Q-bus of the LSI-11 processors. They realize the connection via Q-bus registers between the LSI-11s and the bit-slice based or Z80 based hardware respectively. The main functions of these devices support the microprogram loading and debugging.

The software components of the UNIFIRM system can be used independently of each other /e.g. a microassembler/, but the integrated system offers many conveniences.

The steps of the microprogram development process under the UNIFIRM system are as follows:

1. Creating the microprogram source module on the LSI-11 microcomputer under the RT-11 operating system.

2. Building up the connection between the LSI-11 and PDP-11 computers by means of the TCP program. Login into the RSX-11M system.
3. Sending the source microprogram file to the backup store of the RSX-11M system.
4. Translating the program under the RSX-11M system by one of the cross-assemblers.
5. Receiving the microprogram object module from the PDP-11 disk.
6. Loading and debugging the microprogram.

Cross-assemblers

We use presently three microassemblers in the UNIFIRM system but further translator programs can also be applied which run under the RSX-11M operating system.

MCA

Bit-slice based microprogrammable devices have generally wide microinstruction words /40 - 100 bits/. The MCA cross-assembler can be used for code generation with a maximum width of 128 bits.
The bits of such a microinstruction can be groupped according to their functions, and these groups are called "fields". The structure of the microinstruction word /e.g. width/, the names and position of the fields and the possible contents of the fields are defined in a predefinition file. The defined names are used as mnemonics at writing the microinstruction. The MCA translates the predefinition file together with the source microprogram. The fields and the binary values of the fields are referenced by mnemonic names, so the user does not have to know the positions of the fields in the microinstruction word and the bit patterns attached to the control functions either, since he uses only symbolic

names. Thus, the writing of the microprogram is simplified to the assembly level programming. The listing and the object file are written into disk files.

The MCA cross-assembler is written in FORTRAN language, and it is a simple RSX-11M task.

CPX

The instructions of the CPX cross-assembler are identical with the instructions of the original Intel assembly language. The macro facilities provided by the DOS-80 Macro Assembler can also be used. A new advanced feature of this assembler is, that in addition to absolute program, relocatable code can also be generated. The usage of a library file is also supported.

The features of the CPX are as follows:

1. Macro definitions /redefinitions, nested macro definitions and calls/.
2. Macro library
3. Conditional assembly directives
4. Absolute and relocatable code generation
5. Global symbols in the relocatable modules
6. Symbol table and cross reference table generation.

The object code is linked by the CPY crosslinker program, which generates the executable code. The CPY handles the library files.

ZPX

The ZPX runs under the RSX-11M system. The language of the ZPX is indentical with the CPX language. The extra instructions of the Z80 microprocessor are implemented by macro definitions. The referenced Z80 instructions are linked into the program from the system macro library.

MDDT

The MDDT/S and the MDDT/Z are those interfaces which connect the currently developed devices /bit-slice based and Z80 based hardware/ to the LSI-11 microcomputer. The two versions of the MDDT are slightly different in implementation, but their principle and purpose are the same. One port of both MDDT/S and MDDT/Z is connected to the Q-bus and the other port serves for connecting microprocessors. We use a synchronous bus here which is rather simple, so the bus connections are simple too. The MDDT/Z uses the standard Z80 bus protocol. In the case of bit--slice based system some additional hardware logic must be used in the device being developed, which supports the access to the microprogram memory and to the internal registers. This extra hardware must be present only during the development. The above mentioned bus /Service bus/ consists of

16 data lines
16 address lines
6 control lines
1 clock line

Each line is tri-state and the transmitter receiver logic meets the Z80 requiements. The MDDT/S and MDDT/Z interfaces support the following activities:

- microprogram loading
- microprogram modification
- microprogram start/executation/halt
- single step execution
- dynamic microprogram monitoring
- trace mode
- post-mortem examinations
- interrupt request

Figure 2 illustrates the LSI-11 configuration at the development sites.

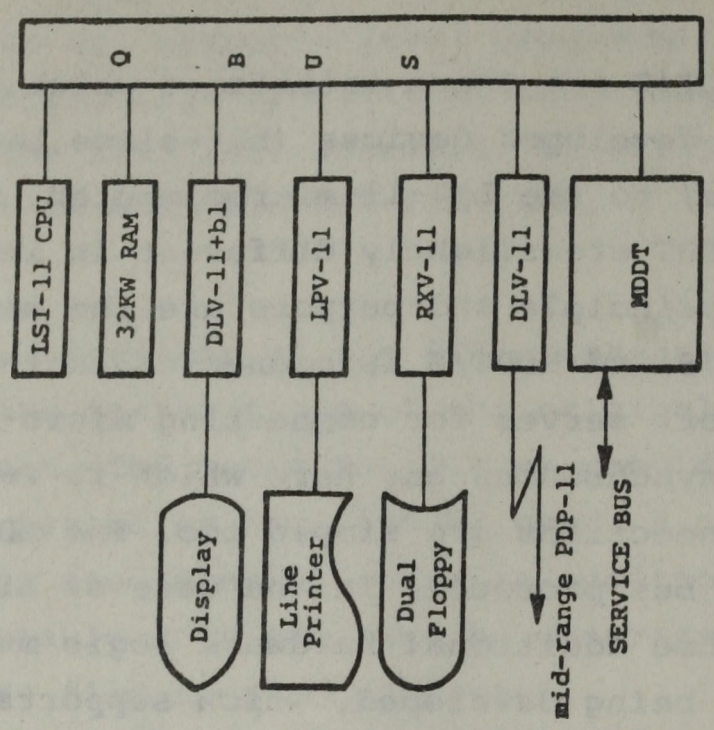


Figure 2. Working station configuration

MCL, MZL

These programs use the two versions of the MDDT hardware interface. The MCL and the MZL programs load the executable code into the micromemory from the backup store of the LSI-11.

The MCL asks for the width of the microinstruction word. The MCL program has an assembly language module, which must be rewritten for every new device according to the bit definitions of the Q-bus registers. This subroutine handles the Q-bus registers of the MDDT/S. The specification of the routine is simple, and the routine consists of 10-15 assembly language instructions.

The MZL is a modified version of the MCL program. It can be used for the Z80 and the Intel 8080 microprocessors. The function of the program is the same as that of the MCL but this uses the MDDT/Z interface.

MCD

The MCD debug program uses all the possibilities provided by the MDDT/S and MDDT/Z hardwares.It helps the de-bugging of the microprogram. The MCD program reads and writes microprogram memory locations, handles 8 breakpoints in the microprogram, and start or restart address can be typed in interactively.

The modification of the microprogram memory is sometimes rather difficult because of the wide micro-instruction word, but the MCD can handle the fields separately so the user may examine and modify only one field at one time.

Of course, the MCD program must be considerably modified in the case of a new hardware. The user has to decide about the modification. The modification of MCD is worth doing in the case of a few hundreds of microinstructions.

TCP

The TCP program under the RT-11 operating system provides the possibility of the online connection between the RT-11 and the RSX-11M operating systems.

It means two features:

1. The console terminal of the RT-11 system can be used as the console terminal of a remote RSX-11M system.
2. File transfers are possible between the backup stores of the two computer configurations.

The TCP program, in cooperation with an RSX-11M task called SRV, sends and receives files to and from the RSX-11M system. The records /lines/ of the files are checked, and in the case of checksum error the transfer is repeated. ASCII and binary /object/ files can be handled.

Summary

The UNIFIRM system provides us with a convenient environment for firmware development. The system utilizes the advantages of a mid-range PDP-11 computer. A lot of useful features are integrated into the system. The resources are used in a costeffective way, which was a strong requirement in our work. This is an "open system", which means that new features and new devices can be added to the system.

ON THE BIFURCATION AND NON-UNIQUENESS OF MHD-EQUILIBRIUM AND TOKAMAK TRANSPORT

G. PÁRIS, Á. ÁG, A. MONTVAI, G. NÉMETH

Central Research Institute for Physics
H-1525 Budapest 114, P.O.Box 49, Hungary

ABSTRACT

Non-linear MHD-equilibrium equation is investigated. By the help of the Ritz-method in a simplified form calculations are given for the Astron and for the non-linear heat-transport problems.

АННОТАЦИЯ

Исследуется нелинейное уравнение МГД-равновесия. Применяется метод Ритца в упрощенной форме. Пример дается для проблемы Астрона, затем рассматривается нелинейное уравнение теплопроводности.

Introduction

The nonlinear character of the models describing relevant problems in plasma physics, especially in the field of controlled nuclear fusion research where one is not allowed to neglect boundary conditions, often leads to the non-existence or non-uniqueness of the solution. This problem is mentioned and treated in some depth by Marder and Weitzner /1970/, and Strauss /1974/. A generalization is given by Field and Papaloizou /1977/. In spite of the amount of work done, a number of open questions remain that have to be investigated.

In what follows we consider two, methodically interesting problems, our aim being to reveal the structure of the resulting non-uniqueness, and to give a useful, practical procedure to unfold it. The given method of approximation is applied to solve an analytically tractable case and to demonstrate its utility.

The importance of such procedures is clear. The global knowledge of the behaviour of the solutions allows us to organize very effectively the actual numerical work, and enables us to show in detail those features which are not always accessible with the exclusive use of numerical computing.

The two investigated problems pertain to magneto-hydrodynamical equilibrium in Astron configuration and to the heat transport in tokamak systems. Both of them easily generalize to other situations and as such may be of general interest.

Bifurcation phenomena in Astron configuration

Let us start from the well known equation describing the Astron configuration:

$$\frac{d^2\psi}{dr^2} - \frac{1}{r}\frac{d\psi}{dr} + \lambda r^2 e^{\psi} = 0 \qquad (1)$$

Here ψ is the system function, λ a constant value depending the actual geometry and other physical parameters of the system. The derivation and the necessary assumption can be found in Weitzner /1970/.

For the cited equation there are two fundamentally differing sets of boundary conditions: one of them is the case of fixed plasma boundary which is characterized by the fact that there is an ideally conducting region around the plasma. In the second case there is a finite vacuum region between the plasma and the conducting wall. Let us investigate the two cases separately.

In the case of the fixed plasma boundary the appropriate boundary conditions are:

$$\left.\frac{d\psi}{dr}\right|_{r=0} = 0 \qquad \left.\psi\right|_{r=a} = 0 \tag{2}$$

a being the inner radius of the limiting cylinder. When the solution is given by

$$e^{\psi} = \frac{c^2}{ch^2 \frac{kc}{2\sqrt{2}}(r^2 - c_1)} \tag{3}$$

where c and c_1 are integration constants, $\lambda^2 = k^2$.

By differentiating and making the necessary substitutions one can easily show that with the following values of the integration constants:

$$c_1 = 0 \,, \qquad c^2 = ch^2 \frac{kc}{2\sqrt{2}} a^2 \tag{4}$$

the boundary conditions are fulfilled.
The second requirement cannot be met with the values of k and a giving $k\,a^2/2\sqrt{2} > 0.6627436$. When the given combination of k and a is equal to 0.6627436 there is a single value of c, say c_0, which satisfies the equation. When $k\,a^2/2\sqrt{2} < 0.6627436$ holds there are two

values of c, viz. c_{01} and c_{02}, leading to different solutions. There is a constant, $c_0 = 1.81017$, with the property that one of the two values is always greater than the other, see fig. 1.

The second interesting case is when a con-conducting region encircles the plasma with finite width. This means that the solution given above is valid only with r lying between zero and r_1, the radius of the plasma cylinder. In the non-conducting layer /vacuum/ no current is flowing, thus:

$$\frac{d^2\psi}{dr^2} - \frac{1}{r}\frac{d\psi}{dr} = 0, \qquad r_1 \leq r \leq a \tag{5}$$

Here r_1 denotes the radius of the plasma column and the other notations correspond to the former case.

The boundary conditions can be formulated as follows:

$$\left.\frac{d\psi_p}{dr}\right|_{r=0} = 0; \quad \psi_p|_{r=r_1} = \psi_v|_{r=r_1} = 0 \tag{6a}$$

$$\left.\frac{d\psi_p}{dr}\right|_{r=r_1} = \left.\frac{d\psi_v}{dr}\right|_{r=r_1}; \quad \psi_v|_{r=a} = b \tag{6b}$$

Here the subscripts p and v refer to the plasma and vacuum regions respectively.

The solution of (1) differs from that given above only in taking into account the appropriate boundary conditions for the $r = r_1$ value. It is easy to see, that $c_1 = 0$, and

$$c^2 = ch^2 \frac{kc}{2\sqrt{2}} r_1^2 \tag{7}$$

The vacuum region has the solution:

$$\psi_v = \frac{b}{a^2 - r_1^2}\left(r^2 - r_1^2\right) \tag{8}$$

which automatically satisfies the second requirements of (6a) and (6b). In (6b) the first equation holds when:

$$\frac{2br_1}{a^2 - r_1^2} = \frac{kc}{\sqrt{2}} \operatorname{th} \frac{kc}{2\sqrt{2}} r_1^2 \qquad (9)$$

Equation (9) can be simplified further. Multiplying it by r_1^2 and assuming the right hand side of (9) to be known we get

$$\frac{br_1^3}{a^2 - r_1^2} = \frac{kc}{2\sqrt{2}} r_1^2 \operatorname{th} \frac{kc}{2\sqrt{2}} r_1^2 \qquad (10)$$

which in turn gives a modified form for (7):

$$\frac{kc}{2\sqrt{2}} r_1^2 = \operatorname{arch} c \qquad (7')$$

Rescaling the problem with $r_1/a = \xi$ we get the final equation:

$$\frac{1+\xi^2}{(1-\xi^2)^2} = \frac{ak^2}{4b} \xi = \nu \xi \qquad (11)$$

In the region $0 \leq \xi \leq 1$ this equation has no solution, one solution, or two solutions depending upon the parameter value ν. The value leading to bifurcation is:

$$\nu_0 = 4.09 \qquad (12)$$

One can see that the bifurcation phenomenon in Astron configuration, first mentioned by Strauss /1974/, Weitzner /1970/ and Lackner /1976/, has a fairly complicated structure depending on the boundary conditions used. This is reflected in the physical state of the machine as well, because the parameter values controlling the scheme and the equilibrium state are connected by the obvious relation

$$B_z = \frac{1}{r}\frac{\partial \psi}{\partial r}\frac{1}{\omega}; \qquad B_r = -\frac{1}{r}\frac{\partial \psi}{\partial z}\frac{1}{\omega}$$

where ω is the angular velocity of the plasma rotation.

The situation in the case of Astron was a relatively simple one because it could be treated analytically.

Our second example shows a situation where this is not the case. In order to reach a fairly complete understanding of the phenomena we have to use an appropriate approximation procedure.

Heat transport in tokamaks

A simple model of the heat flow in a stationary or quasi-stationary energy producing toroidal fusion reactor can be given as

$$3\frac{\partial (nT)}{\partial t} = \frac{1}{x}\frac{\partial}{\partial x}\left[x\left(\chi\frac{\partial T}{\partial x}\right) + 5TD\frac{\partial n}{\partial x}\right] + Q \qquad (13)$$

Here χ is the heat conductivity, n the density and

$$Q \cong \left(3\cdot 10^{-18}T^2 - 10^{-13}\sqrt{T}\right)n^2$$

while T is the plasma temperature. At this moment D, the diffusion coefficient, has not to be specified. The only restriction is that it is a decreasing function of the density. This statement is based on the results of numerous theoretical and experimental works. Because it has recently been shown that for the density, n, self-similar solutions are available, we consider only such density distributions which depend solely on the space coordinate x,

$$n = \frac{n(0)}{(1 + x^{2+\alpha})^2}$$

Neglecting the time derivative one can write (13) in the form:

$$\frac{d}{dx}\left\{x(1+x^{2+\alpha})^2\frac{d}{dx}\left[\frac{T}{(1+x^{2+\alpha})^2}\right]\right\}+B\frac{x}{(1+x^{2+\alpha})^4}\left[T^2-C\sqrt{T}\right]=0 \quad (14)$$

Here, B and C are constants.
Naturally this equation has to be completed with the boundary conditions:

$$\left.\frac{dT}{dx}\right|_{x=0}=0; \qquad T(a)=T_a=\text{const.} \quad (15)$$

It is clear that (14) hardly can be solved analytically. Nevertheless it is possible to consider it as a necessary condition for a variational problem, i.e. the solving of the equation can be transformed into seeking the extremum of

$$I(y)=\int_0^a\left\{\left[-\frac{1}{2}x(1+x^2)^2y'^2+Bx\left[\frac{1}{3}y^3-\frac{2}{3}C\frac{y^{3/2}}{(1+x^2)^3}\right]\right\}dx \quad (16)$$

which is a functional of y. The prime means derivation with respect to the argument. The form given above can be reached after derivation and by introducing the notation $y=T/(1+x^{2+\alpha})^2$. For simplicity we assume $\alpha=0$ to hold, and denote x/a by ξ.

The desired extremum can be reached using the well known Ritz method. This consists of choosing a set of functions containing an arbitrary parameter. The choice is arbitrary. The only requirement to be satisfied is that the functions must be compatible with the stated boundary conditions. In our case a natural choice is

$$y=T(0)(1-\xi^2)^\nu \quad (17)$$

Substituting (17) into (16) and integrating between the specified limits we arrive at a functional depending not only on B and C, but on $T(0)$ and ν as well. Differentiating with respect to these parameters and using the

extremum condition we get three equations. All of these happen to be algebraic in form, and are cubic in $K = \sqrt{\tau(0)}$:

$$K_1^3 - f_1(a,\nu)\,K_1/B - C\,g_1(a,\nu) = 0$$

$$K_2^3 - f_2(a,\nu)\,K_2/B - C\,g_2(a,\nu) = 0 \qquad (18)$$

$$K_3^3 - f_3(a,\nu)\,K_3/B - C\,g_3(a,\nu) = 0$$

where the f_i-s and g_i-s are given functions of a and ν . The exact form of f_i-s and g_i-s shows that each of the three equations has exactly one positive root and two roots which are either negative real or complex conjugates of each other. The physically interesting case is obviously the positive root. From physical arguments it follows that in our case they have to be of the same value. This means that

$$g_1(f_2 - f_3) + g_2(f_3 - f_1) + g_3(f_1 - f_2) = 0 \qquad (19)$$

This equation contains only a and ν , as was stated before.

Taking different values for a we can determine those νs which satisfy (19). Naturally they depend on B and C as parameters. With specific C values /in our case $C = 1.3 \times 10^5$/ it is possible to calculate those ν values which satisfy (19). Furthermore it is possible to estimate the B value which corresponds to the intersection points of the curves. In this way the third equation of (18) may serve as a control: substituting the derived B value one has to get a positive root with the same value as that corresponding to the intersection point. The numerical investigation of the equations gives the following results /table 1/.

$a^2=2$	$\nu_1=1.21,$	$B=0.104$	$\nu_2=4.60,$	$B=0.007875$
$a^2=3$	$\nu_1=1.61,$	$B=0.085$	$\nu_2=6.96,$	$B=0.00646$
$a^2=4$	$\nu_1=1.95,$	$B=0.0827$	$\nu_2=9.35$	$B=0.005784$

table 1. The ν and B values at $a^2=2$, 3 and 4. /see text/

The values shown allow two stationary distributions: one for the smaller value of B and one for the greater. Physically this means the following: if the stationary state of the system is perturbed this may become unstable, and the system may get into another stationary state. Such a phenomenon occurs for example when puffing an inert gas into a quasi-stationary tokamak: the density distribution can then be changed from a "bell-like" shape into a "hollow-type" one, as is shown in fig. 2.

It is of interest to investigate with the devised method the Astron problem as well. Comparison of the results allows us to judge the effectivity of the approximation because it has an analytic solution.

Making the $(r/a)^2=\xi$ substitution we arrive at the functional needed for investigating the extremum:

$$I=\int_0^1\left(-\frac{1}{2}\psi'^2+\lambda e^{\psi}\right)d\xi \qquad (20)$$

Here $\lambda=k^2a^4/4$. It is easy to see that the

$$\psi=\psi(0)(1-\xi^2)$$

choice as base function is an appropriate one. Here $\psi(0)$ is the parameter to be determined by searching the extremum of (20). After integration we get:

$$I=-\frac{2}{3}\psi^2(0)+\lambda e^{\psi(0)}\frac{\sqrt{\pi}}{2\sqrt{-\psi(0)}}\,\Phi\left(\sqrt{-\psi(0)}\right) \qquad (21)$$

where Φ is the error function. Using its known series expansion we get

$$I = -\frac{2}{3}\psi^2(0) + \lambda \sum_{k=0}^{\infty} \left[2\psi(0)\right]^k \Big/ (2k+1)!! \qquad (22)$$

The extremum condition gives

$$\lambda = \frac{1}{3}\left[2\psi(0)\right]^2 \Big/ \sum_{k=1}^{\infty} \frac{k\left[2\psi(0)\right]^k}{(2k+1)!!} \qquad (23)$$

that is, the expression to be numerically investigated. Doing so it can be seen that for every $\lambda \leq \lambda_{max}$ there are two $\psi(0)$ values satisfying the extremum requirement. The λ_{max}=0.892271511 value gives:

$$k \cdot a^2/2\sqrt{2} \leq 0.667933946$$

in complete agreement with the value determined by the analytic method. The $\psi(0)$ value corresponding to λ_{max} is $\psi(0)$ =1.81933 which is again within 1 % of the analytic value. The situation does not deteriorate throughout the whole physically interesting range. Furthermore, there are indications that the correspondence of the analytic and numerically determined solution remains very good even on the negative half axis.

Summing up, we can conclude that the possible very complicated behaviour of the nonlinear differential equations frequently met in plasma physics requires not only the extensive use of numerical methods, but makes necessary the use and development of analytical or semi-analytical procedures which are capable of showing up the occasional complications. Our examples prove this.

Finally we investigate the instability behaviour stemming from the source term in eq. (14).

Fig. 3. shows if we do not use the optimal parameter combination in the beginning, the temperature decreases more quickly as in the previous case. For example changing the optimal value $\nu = 4.6$ /solid curves/ to $\nu = 5$ the dashed curves exhibit a more powerful decrease.

In these calculations we applied a program /KCP 8. T 5 RANS/ written by A.K. Kukushkine /Institute Kurchatov, Moscow/; for this help we acknowledge to him.

References:

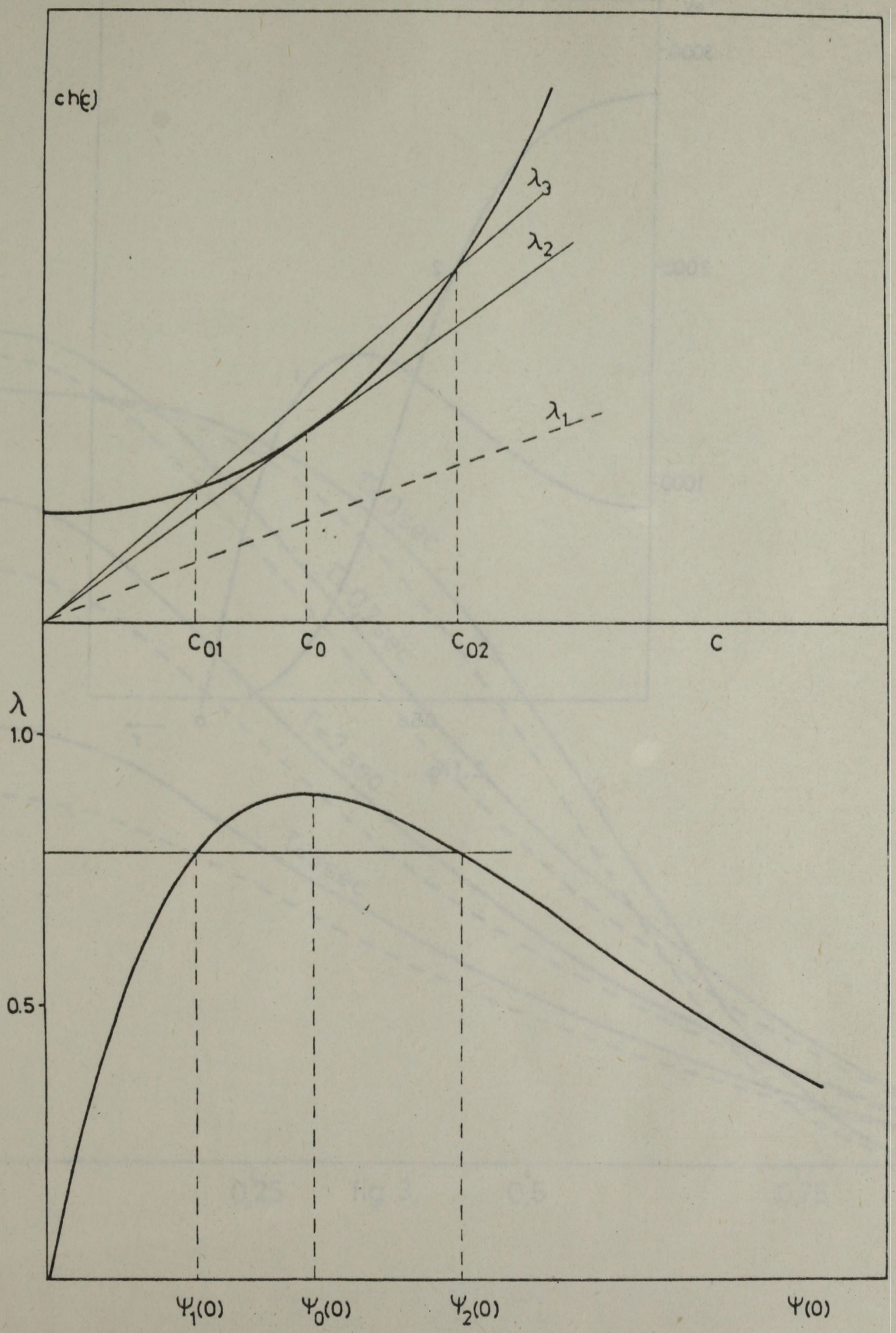
J.J. Field, J.C.B. Papaloizou, (1977), J. Plasma Phys. 18, 347.
K. Lackner, (1976), Comp. Plasma Phys, Comm. 12 33.
M.B. Marder, W. Weitzner, (1970), Plasma Phys. 12 435.
H.R. Strauss, (1974), Phys. Fluids. 17 1040.

Figure captions

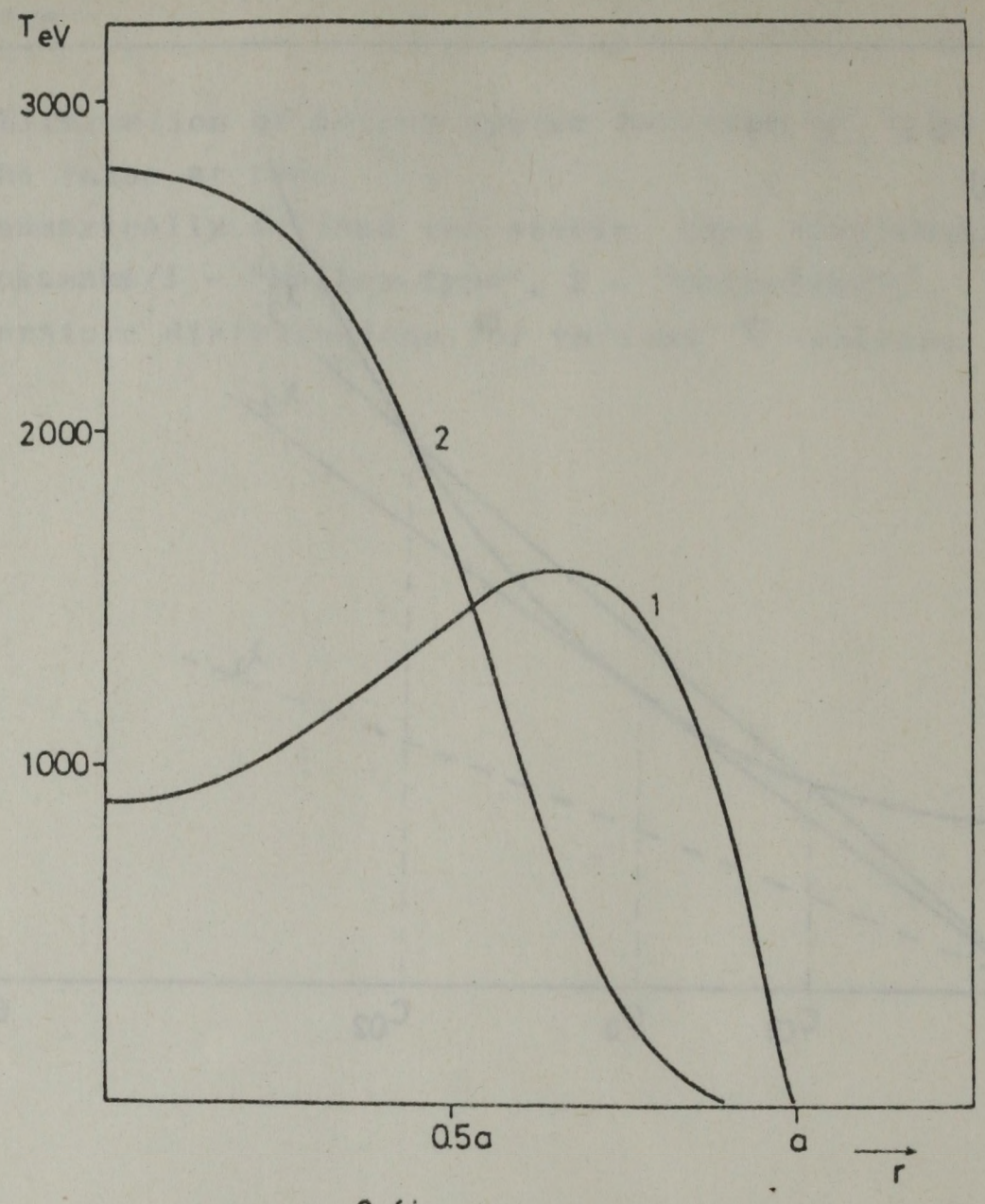
Fig. 1. The bifurcation of Astron system function ψ. $\psi(0)$ is the value at r=0.

Fig. 2. The numerically devised two stable heat distributions in tokamaks /1 - "hollow-type", 2 - "bell-like"/

Fig. 3. Temperature distributions for various τ -values.



1.fig.



2.fig.

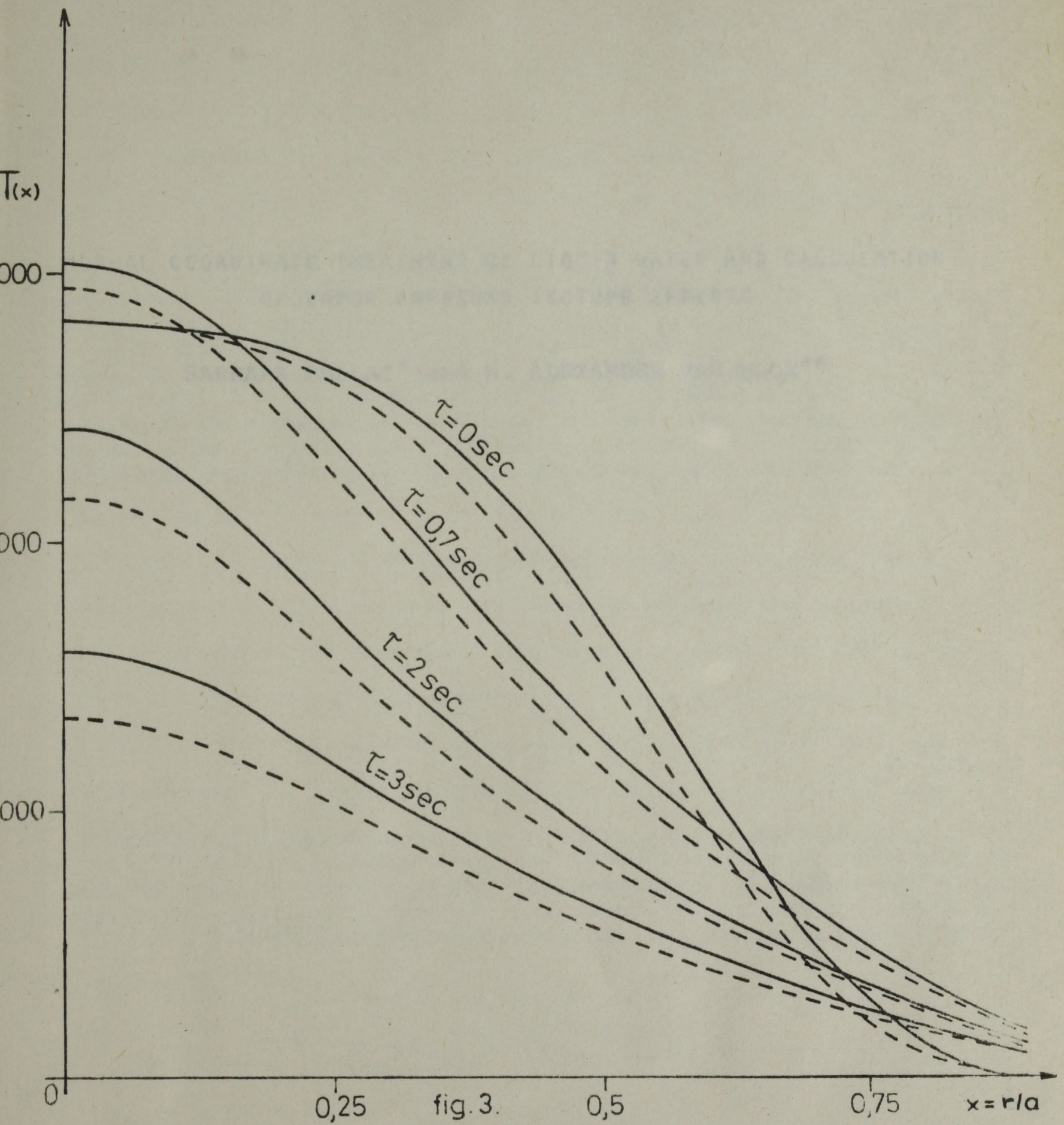
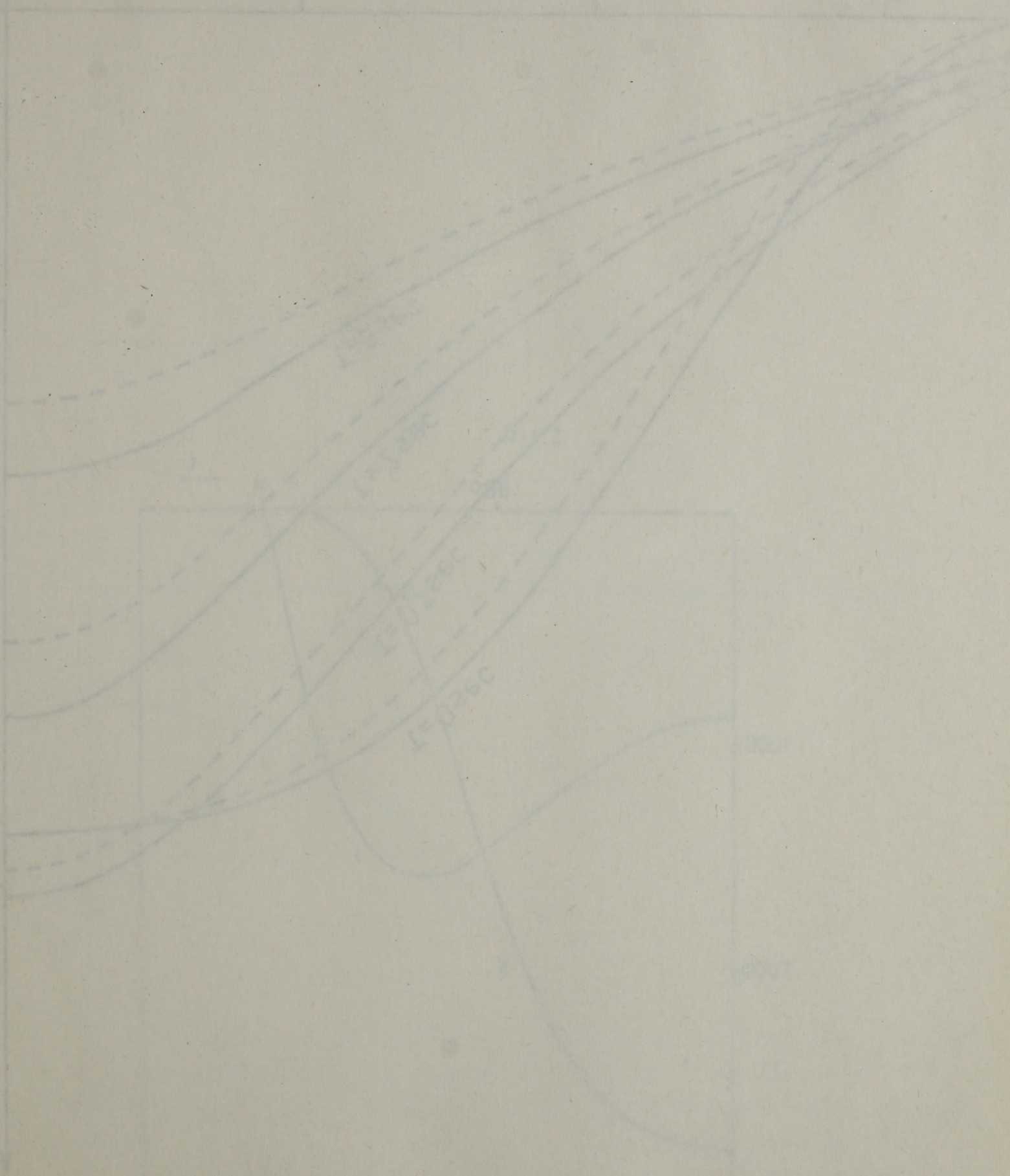


fig. 3.



NORMAL COORDINATE TREATMENT OF LIQUID WATER AND CALCULATION OF VAPOR PRESSURE ISOTOPE EFFECTS

BARBARA GELLAI* and W. ALEXANDER VAN HOOK**

*Central Research Institute for Physics,
H-1525 Budapest 114, P.O.Box 49, Hungary

**Chemistry Department, University of Tennessee,
Knoxville, TN 37996 USA

ABSTRACT

A vibrational analysis of liquid water is reported, assuming a completely hydrogen-bonded network with continuously varying strengths of the hydrogen bonds. Frequency distribution calculations are made for intramolecular stretching and bending modes and for the intermolecular frequency region. The calculated distributions are compared with the experimental spectroscopic ones. As another test vapor pressure isotope effects are calculated from the theoretical distributions for some isotopic water molecules. Results are compared with ones of other authors obtained from a mixture model.

АННОТАЦИЯ

Распределение фундаментальных частот колебания по состояниям молекулы H_2O с изотопами T,D,O^{17},O^{18} были определены в жидкости. Мы сравнивали теоретические распределения частот колебания с экспериментальными спектрами. Термодинамические изотопные эффекты были вычислены из теоретических распределений.

I. Introduction

In spite of the relentless efforts of scientists since the days of Roentgen [1], there is still no generally accepted theory of the structure of liquid water, to say nothing of a general theory of the liquid state. In its absence various models have been proposed that can be classified under two broad headings; viz. mixture models, continuum models.

Mixture models [2-4] describe liquid water as an equilibrium mixture of several molecular species with a different number of hydrogen bonds per molecule, whereas continuum mopdels [5-10] consider liquid water to be a completely hodrogen-bonded network. If a vibrational analysis of liquid water is performed assuming one of the above-mentioned models the frequency distribution of the intra- and intermolecular vibrations can be calculated. Comparison between the theoretical and the experimental spectroscopic distributions can give information on the validity of the assumed model. In addition to the spectroscopic comparison the calculation of certain thermodynamic properties which depend on the frequencies of liquid water, together with a comparison with experimental data can also be useful. One such thermodynamic property is the vapor pressure ratio of isotopically substituted molecules. This ratio is often discussed in the framework of the oversimplified cell model [11], using the harmonic approximation. The condensed phase is modelled in terms of a single and assumed average molecule which executes its motions in a harmonic and isotropic cell imposed by the intermolecular field of its neighbors [12]:

$$\frac{P'}{P} = \prod_{\substack{\text{internal}\\ \text{frequencies}}}^{3n-6} \left[\frac{(u_i/u_i')_c}{(u_i/u_i')_g}\right]\left[\frac{\exp(u_i'-u_i)_{c/2}}{\exp(u_i'-u_i)_{g/2}}\right] x \qquad (1)$$

$$\left[\frac{(1-\exp(-u_i'))_c/(1-\exp(-u_i))_c}{(1-\exp(-u_i'))_g/(1-\exp(-u_i))_g}\right] x \prod_{\substack{\text{external}\\ \text{condensed}\\ \text{frequencies}}}^{6} \frac{u_i}{u_i'}\left[\exp\left(\frac{u_i'-u_i}{2}\right)\right] x \left[\frac{1-\exp(-u_i')}{1-\exp(-u_i)}\right]$$

where $u_i = \frac{h\nu_i}{kT}$, ν_i is the ith normal frequency and n is the number of atoms per molecule. Using Eqn. (1), vapor pressure isotope effects (VPIE) prove to be a very sensitive probe for measuring the net frequency shift on condensation. For example, consider OH substitution at 300K where one readily calculates that a 1 cm^{-1} change in the shift of an OH stretching motion is equivalent to $\Delta \ln(P_H/P_D) = \pm 2.4 \times 10^{-3}$.

In the past ten years there have been several computations of general spectroscopic features for water using both mixture [13-16] and continuum models [17-20]. Using the continuum model Curnutte and Bandekar calculated the intramolecular frequency distributions of the molecule HDO as well as intermolecular frequency distributions for $H_2{}^{16}O$ and $D_2{}^{16}O$ [19-20]. O'Ferrall et al. [16] calculated liquid phase intra- and intermolecular frequencies and vapor pressures for the molecules $H_2{}^{16}O$ and $D_2{}^{16}O$ using a nine-atom mixture model.

In this paper we report a normal coordinate treatment of liquid water considered as a completely hydrogen bonded network. It is considered that the frequency shifts on condensation are mainly due to the changes in the OH stretching force constant on condensation. Frequency distributions are calculated for intramolecular stretching and bending modes as well as for translation and librational modes in the intermolecular frequency region. The theoretical distributions are compared with the experimental spectroscopic ones and, in addition, are used to calculate the isotopic vapor pressure ratios of isotopic water molecules.

II. Normal coordinate treatment of liquid water

For computing economy and mathematical simplicity, the inter- and intramolecular portions of the problem were solved independently using the appropriate continuum models [19,20].

1. Computation of intramolecular frequency distribution

We used the following symmetry coordinates [21]:

$$u_1 = x_1' - [(m_2 x_2' + m_3 x_3')/(m_2 + m_3)]$$
$$u_2 = y_1' - [(m_2 y_2' + m_3 y_3')/(m_2 + m_3)] \qquad (2)$$
$$u_3 = x_2' - x_3'$$

where the primed quantities are rectangular components of displacement from the equilibrium position and $Z_i' = 0$. Identification of the internal coordinates of a general bent XYZ molecule is given in Fig 1. For the molecules $H_2{}^{16}O$ and $D_2{}^{16}O$ Eqns. (2) simplify appropriately.

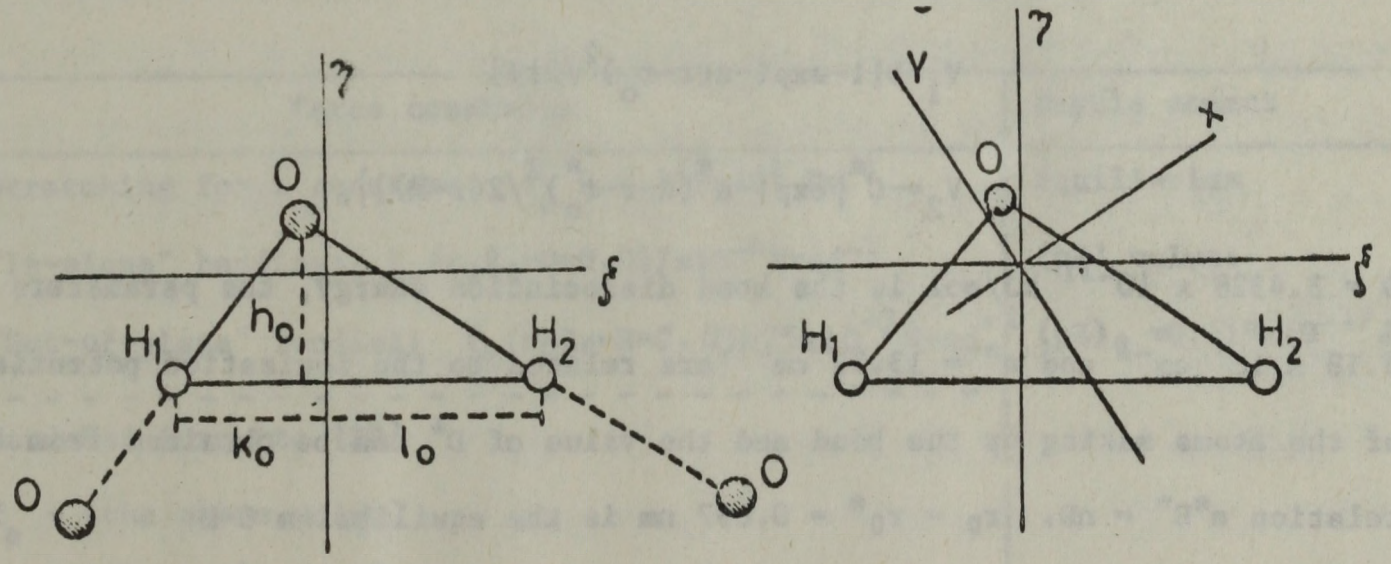


Fig 1. Identification of internal coordinates of a bent XYZ molecule [21].

The kinetic and potential energy in terms of the symmetry coordinates (2) follow:

$$T=(1/2)[\mu_{11}\dot{u}_1^2+\mu_{22}\dot{u}_2^2+\mu_{33}\dot{U}_3^2+2\mu_{13}\dot{u}_1\dot{u}_3 + 2\mu_{12}\dot{u}_1\dot{u}_2+2\mu_{23}\dot{u}_2\dot{u}_3]$$

$$V=(1/2)[f_{11}u_1^2+f_{22}u_2^2+f_{33}u_3^2+2f_{12}u_1u_2 + 2f_{13}u_1u_3+2f_{23}u_2u_3]. \quad (3)$$

Coefficients $\mu_{ij}(i_{ij}=1, 2, 3)$ are the elements of the kinetic energy matrix and can be calculated from the mass and geometric parameters of the molecule. The numerical values of these parameters are given in Table 1. The quantities f_{ij} (i, j = 1, 2, 3) are generalized force constants defined by Eqn. (4).

$$f_{ij}=[(A_iA_j)+(B_iB_j)]f_r+C_iC_jf_\beta + [(A_i+B_i)C_j+(A_j+B_j)C_i]f_{r\beta}+(A_iB_j+A_jB_i)f_{rr}. \quad (4)$$

The coefficients A_i, B_i, C_i are the elements of the matrix of the transformation between the internal and the symmetry coordinates. The diagonal force constants f_r and f_β are functions of the oxygen-oxygen distance. The formula for f_r can be derived by using the linear model of the Lippincott-Schroeder potential function $V=V_1 + V_2$ [22], where

$$V_1=D[1-\exp(-n(r-r_o)^2/2r)] \quad (5)$$

$$V_2=-C^*\{\exp[-n^*(R-r-r_o^*)^2/2(r-R)]\}.$$

$D = 3.4328 \times 10^{-11}$ kJ/mol is the bond dissociation energy, the parameters $n = 9.18 \times 10^8$ cm^{-1} and $n^* = 13.32$ cm^{-1} are related to the ionization potentials of the atoms making up the bond and the value of D^* can be obtained from the relation $n^*D^* = nD$. $r_0 = r_0^* = 0.097$ nm is the equilibrium O-H distance, r and R are a given O-H and a given O...O distance respectively.

From the conditions

$$\left(\frac{\partial V}{\partial r}\right)_{eq}=0 \quad \frac{\partial^2 V}{\partial r^2} = f_r \quad (6)$$

Table 1. Masses and moments of inertia of some isotopic water molecules

Atom	Atomic masses (AWU) [24]	Molecule	Moments for inertia in equilibrium geometry (kgm^2)		
			I_{xx}	I_{yy}	I_{zz}
H	1.00783	$H_2{}^{16}O$	2.18671×10^{-47}	3.15787×10^{-47}	0.971157×10^{-47}
$^2H(D)$	2.01410	$D_2{}^{16}O$	4.37006×10^{-47}	6.11581×10^{-47}	1.74575×10^{-47}
$^3H(T)$	3.01605	$T_2{}^{16}O$	6.54402×10^{-47}	8.92040×10^{-47}	2.37637×10^{-47}
^{16}O	15.99491	$H_2{}^{17}O$	2.18671×10^{-47}	3.17012×10^{-47}	0.983411×10^{-47}
^{17}O	16.99914	$H_2{}^{18}O$	2.18638×10^{-47}	3.16367×10^{-47}	0.97729×10^{-47}
^{18}O	17.99916				
HOH $\measuredangle$ = 109°28' R_0 = 0.285nm					r_0 = 0.099nm

Table 2. Force field and dipole moment for liquid water in equilibrium geometry (25°C)

Force constants	Dipole moment
Stretching force constant: $f_\rho = 0.192 \times 10^2$ Nm^{-1}	Equilibrium
"In-plane" bending: $f_\sigma / r_0 R_0 = G = 0.037 \times 10^{-8} Nrad^{-1}$	(pE) value:
"Out-of-plane" bending: $f_\tau / r_0 R_0 = H = 0.03675 \times 10^{-1} Nrad^{-1}$	$(pE)_0 = 0.119 \times 10^{-17} J$
Bandekar-Curnutte [20]	
f_ρ = (the above value)	
$G = 0.024 \times 10^{-8}$ $Nrad^{-1}$	
$H = 0.03275 \times 10^{-8}$ $Nrad^{-1}$	

characteristic of the motion of the hydrogen atom along the axis of the hydrogen bond we obtain for the intramolecular stretching force constant

$$f_r=\left(\frac{nD}{r^3}\right)x\left[\exp\left(\frac{-n\Delta r^2}{2r}\right)\right]x\left[r_0^2-\frac{n\Delta r^2}{4r}(r+r_0)^2\right]+$$
$$\left[\frac{n^*D^*}{(R-r)^3}\right]x\left[\exp\left(-\frac{n^*(R-r-r_0^*)^2}{2(R-r)}\right)\right]x\left[r_0^{*2}-n^*(R-r-r_0^*)^2\frac{(R-r+r_0^{*2})}{4(R-r)}\right]. \quad (7)$$

Because we were not able to obtain an expression relating the O-O distance to the bending force constant, the following empirical relation was substituted:

$$F_\beta=f_\beta^\circ\exp\{a[(R_0/R)-1]\}. \quad (8)$$

Here R_0=0.285 nm is the equilibrium O-O distance, R is a given O...O distance and f_β^0 = 2.019 Ncm^{-1} is the bending force constant at equilibrium. The coefficient a = 0.3465 was chosen to reproduce the experimental bandwidth and maximum (298K) of the bending mode of H_2O. After marking off the range of the accessible O...O distances 0.318 nm $\geq R_i \geq$ 0.256 nm (the upper bound is arbitrary), several thousand molecular configurations were selected by picking O...O distance pairs from the allowable range for the O...O distances according to the method of Metropolis et al. [23]. The main diagonal stretching and bending constants were changed according to randomly picked O...O distances using Eqns. (7) and (8); the numerical value of the off-diagonal force constants was chosen to reproduce the water vapor fundamental frequencies with R = ∞ and can be seen in Table 2. The intramolecular frequencies of a given molecule can be obtained as the solution of the following eigenvalue problem:

$$\begin{bmatrix} \mu_{11}\lambda-f_{11} & \mu_{12}\lambda-f_{12} & \mu_{13}\lambda-f_{13} \\ & \mu_{22}\lambda-f_{22} & \mu_{23}\lambda-f_{23} \\ \text{symm.} & & \mu_{33}\lambda-f_{33} \end{bmatrix} \quad (9)$$

where $\lambda_k = 4\pi^2\nu_k^2$ and ν_k / k=1,2,3 / is the k-th normal frequency of the molecule. The ordinate values of the observed O...O pair correlation function [25-27] /Fig.2./ represent the probability that an O-H...O band has the corresponding O...O length, so the probability of a random water molecule configuration is the product of the ordinates corresponding to the randomly chosen O...O distances. These probability values were stored in subintervals or "boxes" having an appropriate length in a given range of frequencies. The stored probability values are defined as

$$a_i(\nu) = \text{Prob}[\nu_{i-1} < \nu \leq \nu_i] = \frac{N_i}{N} \qquad (10)$$

where ν is a given frequency in a given frequency distribution; N_i is the number of ν such that $\nu_{i-1} < \nu \leq \nu_i$; ; and N is the total number of frequencies in the given frequency range [28]. Vibrational distributions are converted into spectra by taking the intensity of the band at each frequency as being determined by the number of oscillators having that particular frequency at any instant. The intramolecular stretching frequency distribution for $H_2{}^{16}O$ and $D_2{}^{16}O$ are presented in Figs. 3 and 4 respectively; Fig. 5 reviews the calculated spectra in the bending region. The histograms are drawn over 20 cm^{-1} intervals for the stretching region and over 3 cm^{-1} intervals for the bending region. The observed spectra are shown as insets in the figures. It can be seen that the calculated shapes and bandwidths $(\nu_1+\nu_3)$ correlate nicely with the experimental spectroscopic distribution. Assignments of the band maxima and a comparison with observed spectra and the results of O'Ferrall et al [16] will be provided in the last section of the paper.

2. Computation of intermolecular frequency distribution

The local structure model for our computation was the same as that used

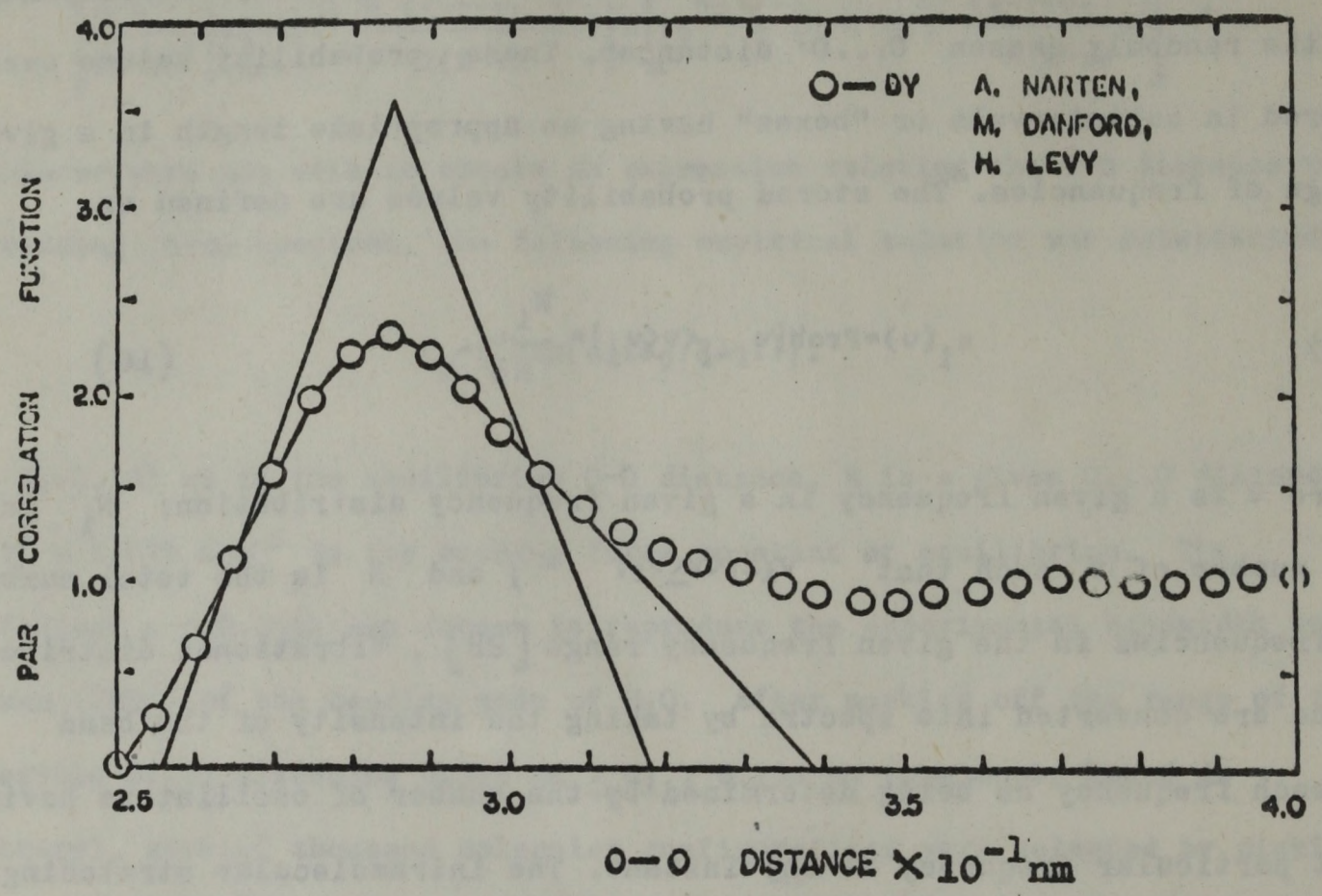


Fig. 2. The pair correlation function for liquid water at 25°C vs. the oxygen-oxygen distance is denoted by circles [25].

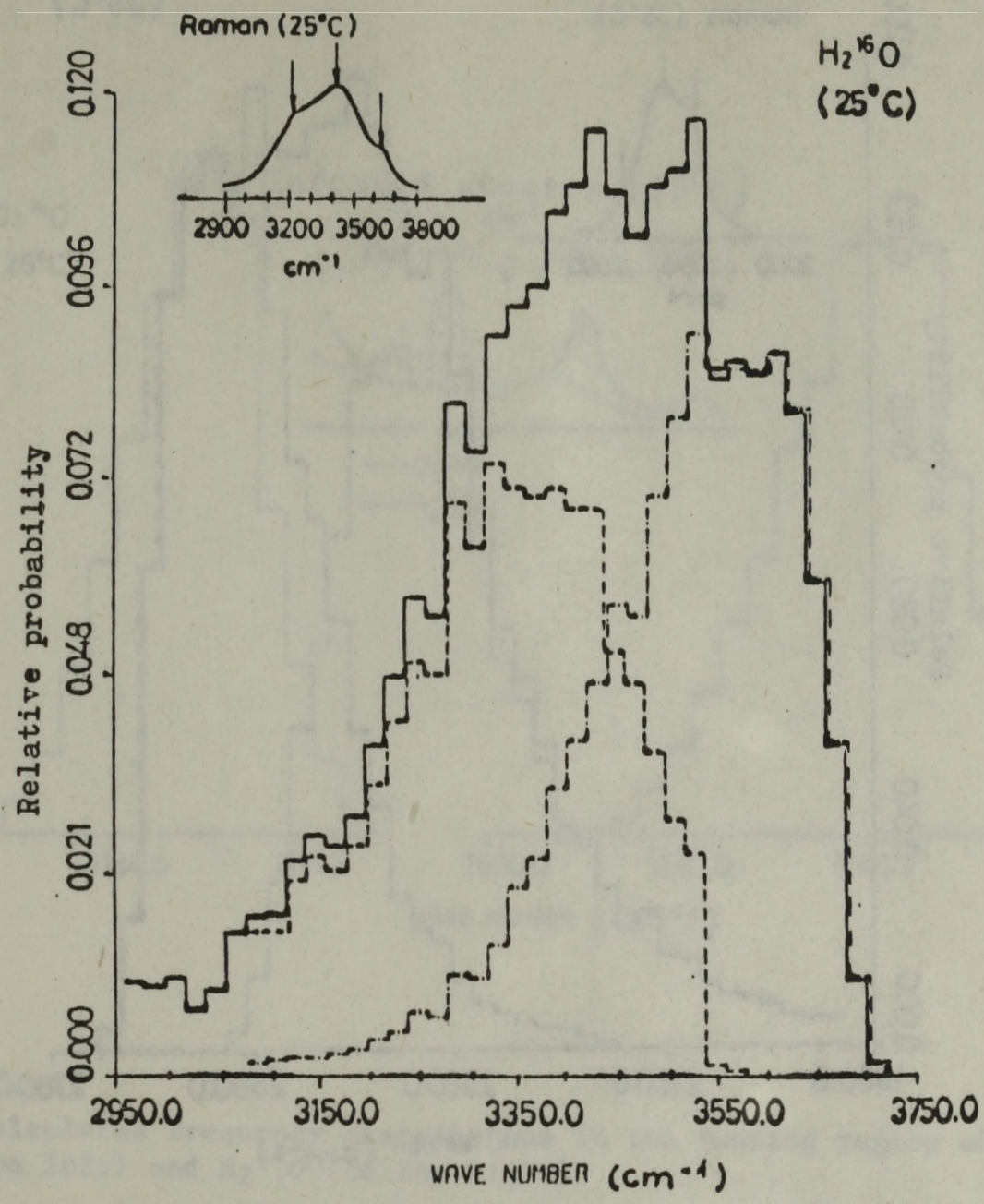


Fig. 3. Calculated frequency distributionof $H_2{}^{16}O$ in the OH stretching region. ν_1 (------) is the left; ν_3 (•—•—•—) is to the right; $\nu_1+\nu_3$ = solid line.

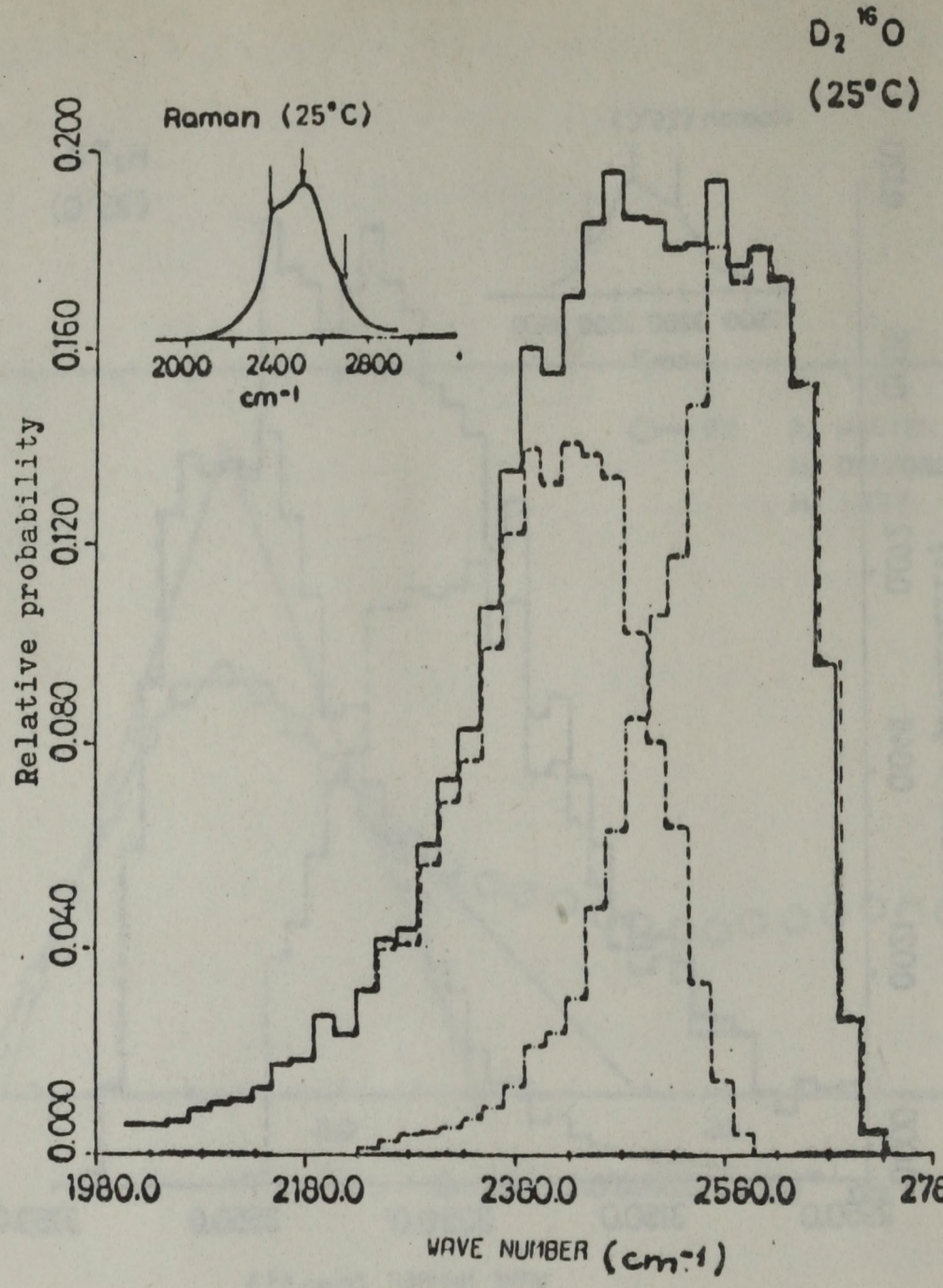


Fig. 4. Calculated frequency distribution of $D_2{}^{16}O$ in the OD stretching region. ν_1 (------) is the left; ν_3 (•—•—•—•—) is to the right; $\nu_1+\nu_3$ = solid line.

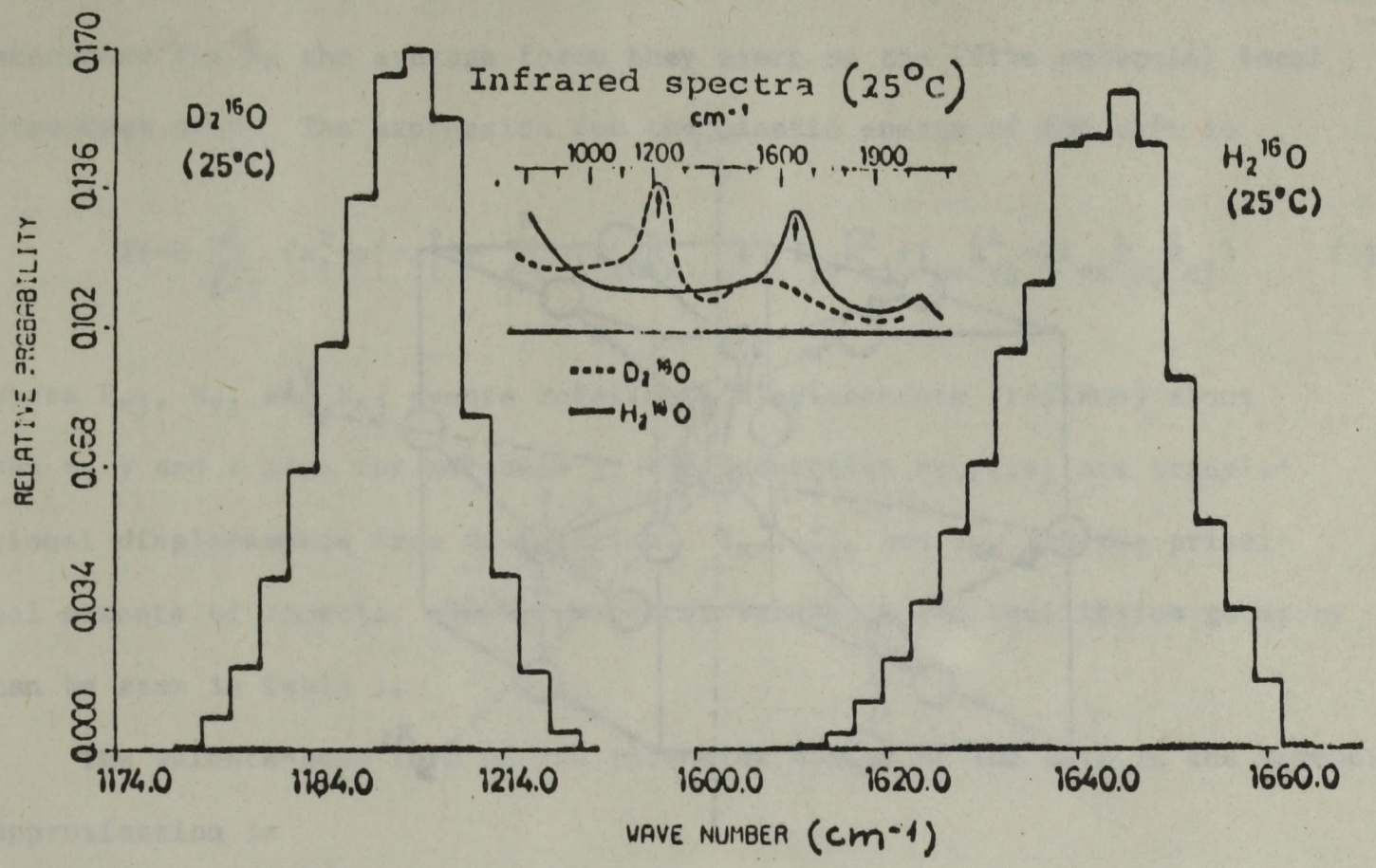


Fig. 5. Calculated frequency distribution in the bending region of $D_2{}^{16}O$ (to the left) and $H_2{}^{16}O$ (to the right).

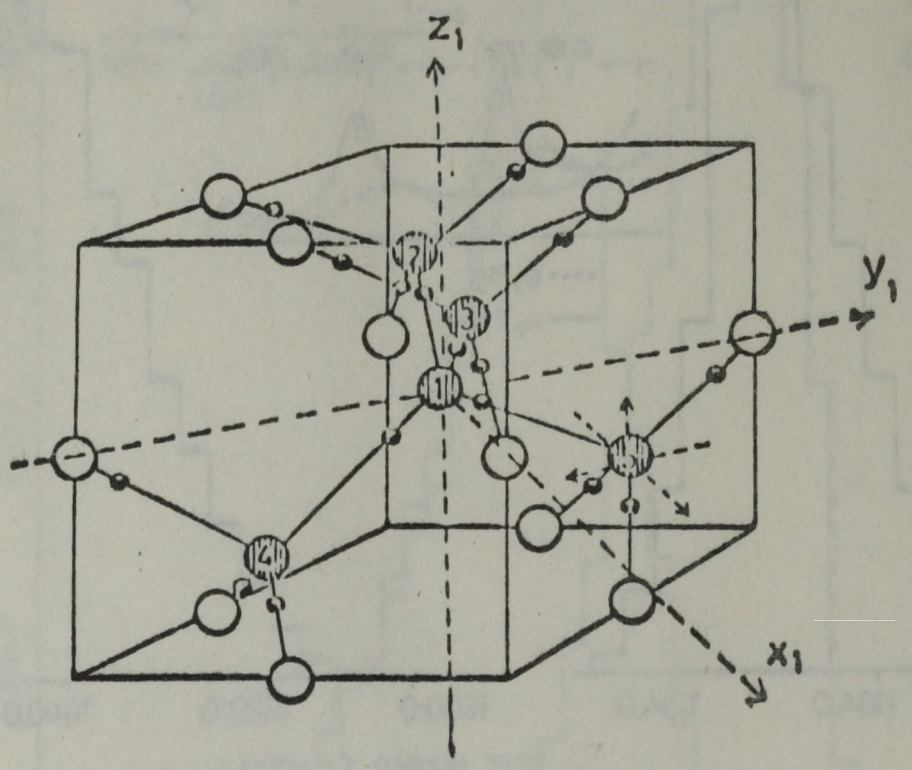


Fig. 6. Local strucutre unit $(H_2O)_5$ used for computing the intermolecular frequency distribution of liqiud water.

by Bandekar and Curnutte [20]. It consists of a central rigid water molecule tetrahedrally coordinated to four nearest neighbors surrounded by a rigid case of next nearest neighbors (Fig. 6). The influence of these last neighbors is accounted for by the average force they exert on the (five molecule) local structure unit. The expression for the kinetic energy of the unit is

$$2T=M\sum_{i=1}^{5}(x_i^2+y_i^2+z_i^2)+\sum_{j=1}^{5}(I_{xx}\dot{R}_{xj}^2 + I_{yy}\dot{R}_{zj}^2+I_{zz}\dot{R}_{yj}^2-2I_{yz}\dot{R}_{yj}\dot{R}_{zj}) \qquad (11)$$

where R_{xj}, R_{yj} and R_{zj} denote rotational displacements (radians) about the x, y and z axes for molecule j; the quantities x_i, y_i, z_i are translational displacements from equilibrium. I_{xx}, I_{yy} and I_{zz} are the principal moments of inertia. Their numerical values in the equilibrium geometry can be seen in Table 1.

The valence-bond form of the potential energy of the unit in the harmonic approximation is

$$2V=\sum_{i=1}^{16}(f_{\rho i}\rho_i^2+f_{\sigma_i}\sigma_i^2+f_{\tau_i}\tau_i^2)+\sum_{j=1}^{5}(pE)_j(R_{xj}^2+R_{yi}^2) \qquad (12)$$

where ρ, σ and τ represent the stretching, in-plane bending, and out-of-plane bending displacement of hydrogen bonds from their position of equilibrium (Fig. 7).

Algebraic expressions for f_ρ, f_σ and f_τ, the corresponding force constants, were obtained from the Lippincott-Schroeder potential function which is valid for a bent hydrogen bond [29]; they were scaled by multiplication with a constant to yield the numerical value given in Table 2. These values were chosen as a means of

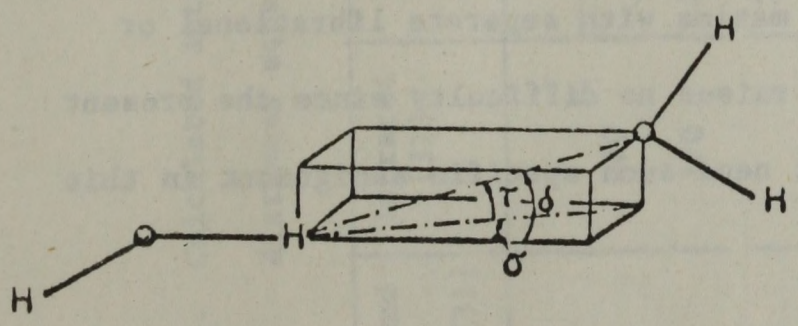


Fig. 7. Hydrogen bond coordinates

producing the convergence of the calculated VPIE (see next section). The last term in Eqn.(12) is the interaction energy of the electric dipole moment, p, of a molecule in the liquid state with the local electric field, E, due to its neighbors. The product pE was varied according to the average value of the reciprocal of the O...O distance cubed, where the average is taken over the four hydrogen bonds of the molecule of interest. The potential energy expression was transformed from valence bond coordinates to cartesian displacement coordinates using the geometrical relationships between the two sets of coordinates [30], and Euler-Lagrange equations were constructed for the local structure unit $(H_2O)_5$. Frequencies obtained as the solution of a 30x30 eigenvalue problem first constructed for the symmetric configuration (Fig. 6) with all O...O distances set to 0.285 nm. The resulting frequencies for this symmetric configuration are compared with the assignments from observed spectra [31] and the frequencies calculated by Bandekar and Curnutte [20], for $H_2{}^{16}O$ in Table 3, and for $D_2{}^{16}O$ in Table 4. The intermolecular frequency distributions of water in the liquid phase as obtained by various experimental techniques can be seen in Fig. 8. The calculated frequency distribution of 250 random $(H_2{}^{16}O)_5$ configurations obtained by assigning the 16 hydrogen bonds randomly for each configuration was obtained by accumulating the probability just as was done in the case of intramolecular vibrations. It can be seen for $H_2{}^{16}O$ in Fig. 9 and for $D_2{}^{16}O$ in Fig. 10. It is clear from the figure that it is almost impossible to identify band maxima with separate librational or hindered translational classes. This raises no difficulty since the present method for computing the VPIE does not need such specific assignment in this region.

Table 3. Calculated intermolecular frequencies for the symmetric unit $(H_2{}^{16}O)_5$ with all O...O distances equal to 0.285 nm are compared with the experimental findings (25°C; cm^{-1})

Raman and hyper Raman [31]	Infrared [31]	Present calculation	Bandekar-Curnutte [20]
		Frequencies of hindered translations	
~60	~70	70,73,75	71,77,75
166	170	160,162,162,163,163,163,164,164,164	164,164,160,163,164,162,165,165,165
		217,218,218	217,219,219
		Frequencies of hindered rotations	
425-450		394,394,394,394,394	450,450,450,452,452
550		577,577,577,578,578	553,553,554,554,553
	685	611,611,611,612,612	
720-740			722,722,722,723,723

Table 4. Calculated intermolecular frequencies for the symmetric unit $(D_2{}^{16}O)_5$ with all O...O distances equal to 0.285 nm are compared with the experimental findings (25°C; cm^{-1})

Raman and hyper Raman [31]	Infrared [31]	Present calculation	Bandekar-Curnutte [20]
		Frequencies of hindered translations	
~60		66,69,71	68,74,72
~175	165	152,153,154,155,155,155,155,155,155	157,157,152,155,157,154,157,157,157
		205,206,207	206,208,207
		Frequencies of hindered rotations	
305,350		282,282,282,283,284	319,320,318,320,319
415		427,427,431,431,431	397,399,397,398,397
		432,432,432,436,437	
500,570	505		538,539,538,539,538

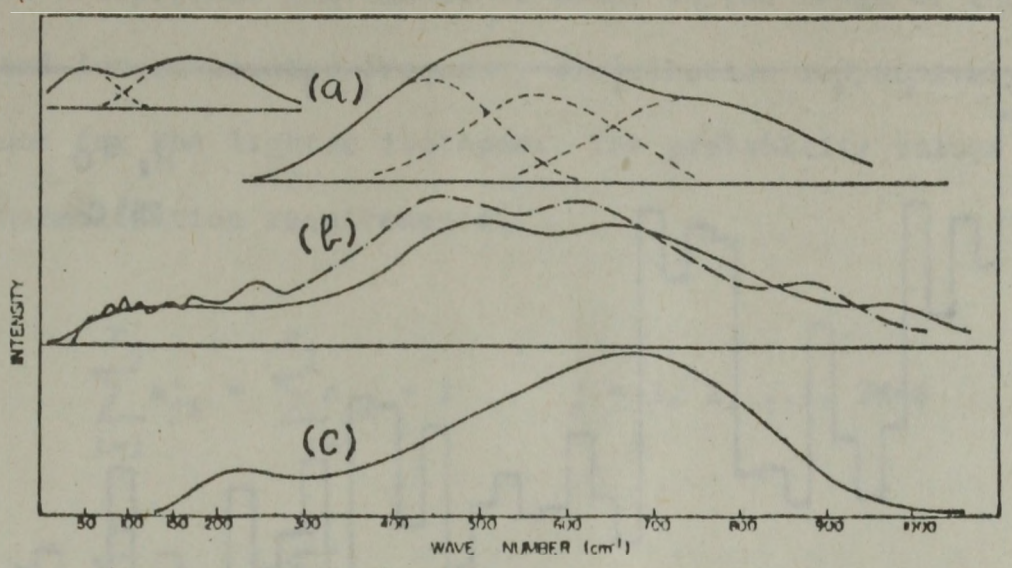


Fig. 8. Experimental Intermolecular Frequency Distributions in Water.
(a) Raman scattering at 25°C [31,32]
(b) Inelastic neutron scattering at 25°C [33]
(c) Far infrared results at 25°C [34]

III. Computation of the vapor pressure isotope effect

Taking the logarithm of Eq . (1) and using the stored probabilities as calculated in the previous section we obtain

$$\ln\frac{P'}{P} = \sum_{j=1}^{3N-6}\left\{\sum_{i=1}^{nj} a_{ji}\ln\left[\frac{u_{ji}\exp(-u_{ji}/2)}{1-\exp(-u_{ji})}\right] - \right.$$

$$\left. \cdot \sum_{i=1}^{j} a'_{ji}\ln\left[\frac{u'_{ji}\exp(-u'_{ji}/2)}{1-\exp(-u'_{ji})}\right]\right\} - \ln(s/s'\ f)_{gas} + \qquad (13)$$

$$+6\left\{\sum_{i=1}^{m} a_i\ln\left[\frac{u_i\exp(-u_{i/2})}{1-\exp(-u'_i)}\right] - \right.$$

$$\left. - \sum_{i=1}^{n'} a'_i\ln\left[\frac{u'_i\exp(-u'_i/2)}{1-\exp(-u'_i)}\right]\right\},$$

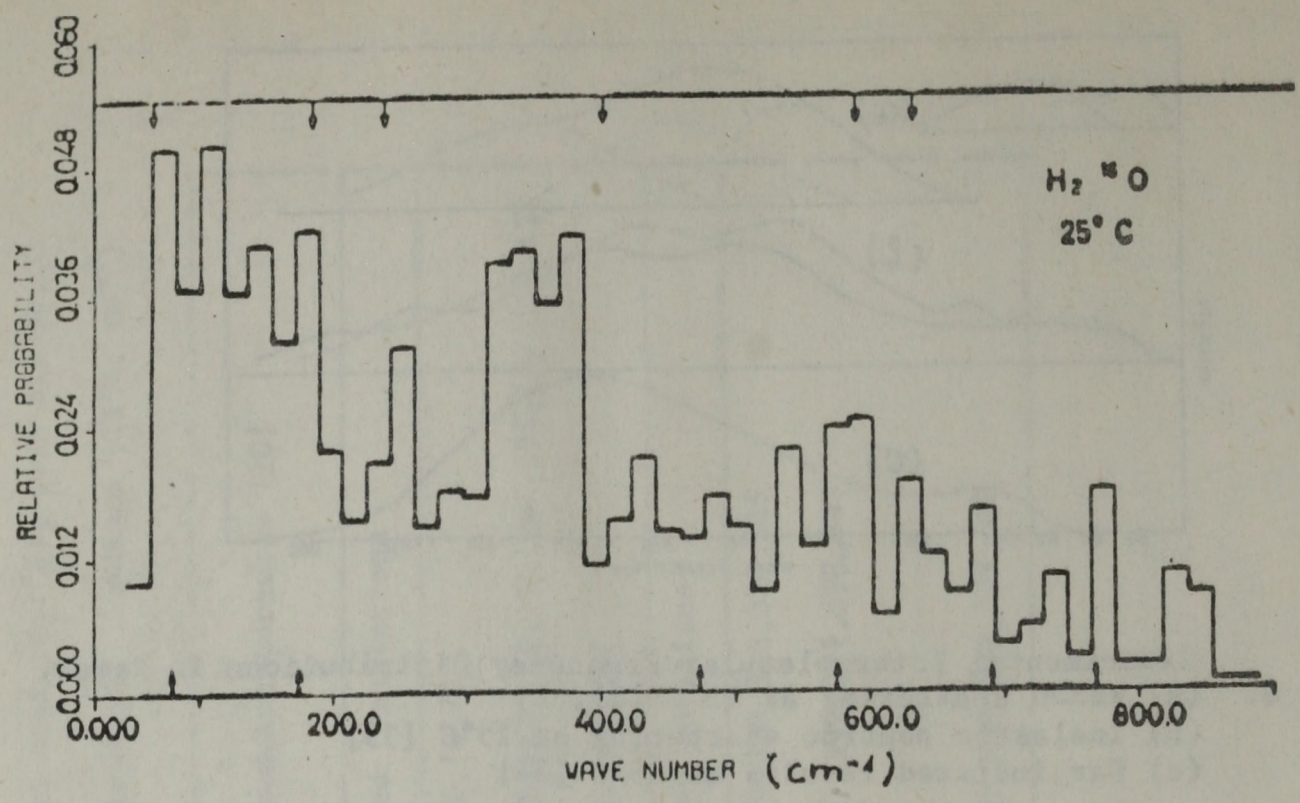


Fig. 9. Calculated frequency distribution of $H_2{}^{16}O$ in the intermolecular region.

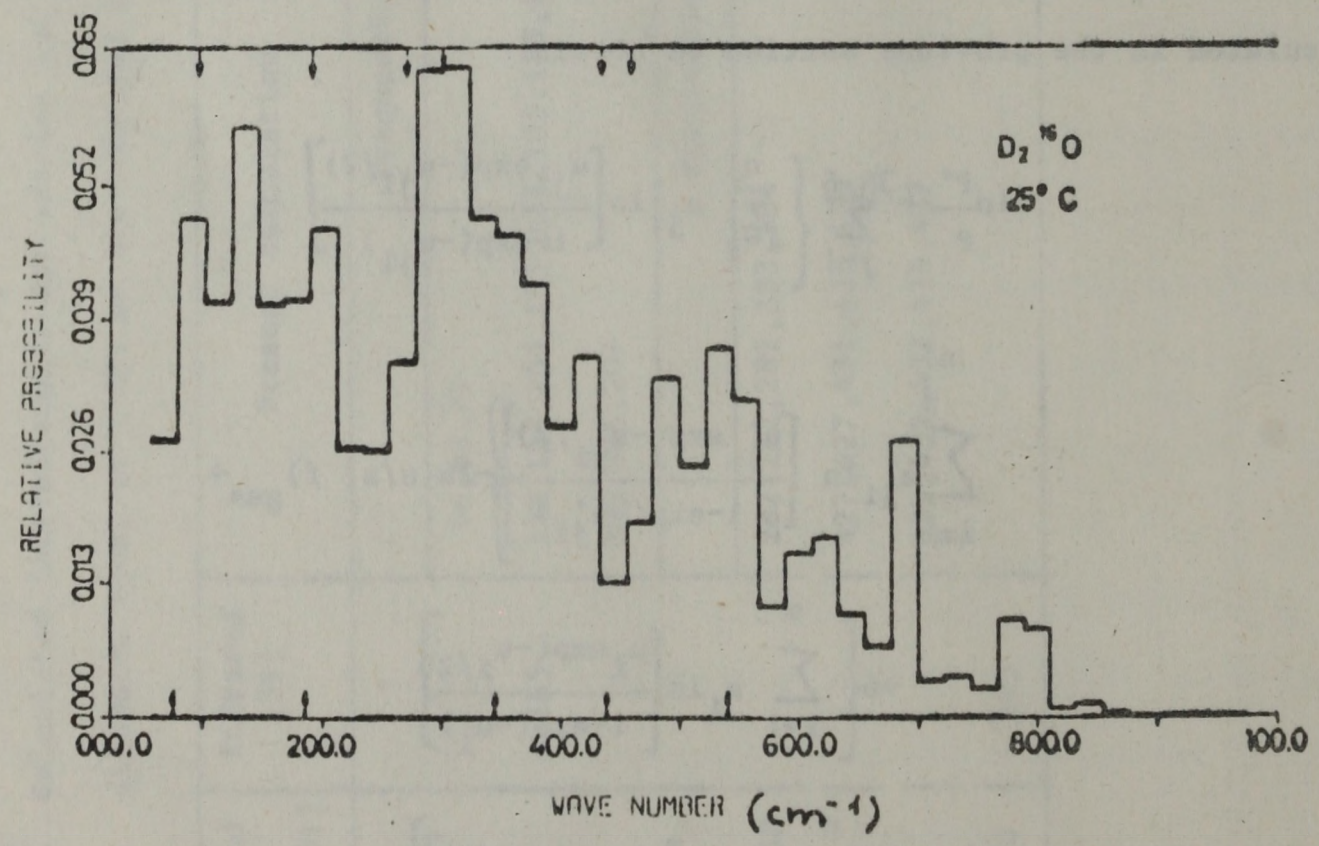


Fig.10. Calculated frequency distribution of $D_2{}^{16}O$ in the intermolecular region.

where n_j and n represent the number of boxes in the range of the j-th intramolecular and intermolecular frequency distribution respectively. Primes denote values for the lighter isotopes. The probability values fulfil the following normalization requirements:

$$\sum_{i=1}^{n'_j} a'_{ji} = \sum_{i=1}^{n_j} a_{ji} = 1 \qquad j = 1, 2, \ldots, 3N-6$$

$$\sum_{i=1}^{n'} a'_i = \sum_{i=1}^{n} a_i = 1 \quad . \tag{14}$$

In Eqn. (13) the term $\ln (s/s'f)_{gas}$ contains the frequencies of the gas phase. The main diagonal elements of the force constant matrix, F_{gas}, necessary for calculating these frequencies were obtained by substituting the maximum allowable O...O distance, R_{max} = 0.318 nm into the L.S. potential function (5) and into Eqn. (8). The off diagonal elements of the matrix were taken from Curnutte and Bandekar (Table 2.). It was assumed for such separated water molecules the intermolecular forces can be neglected. Thus the model is an approximation of the gas phase and it is more appropriate to label the term $\ln(s/s'f)_{gas}$ "isolated" instead of "gas." The frequencies of some "isolated" isotopic water molecules are compared with the real gas phase frequencies in Table 5.

Table 5. Comparison of "isolated" and gas phase frequencies

Phase	$H_2{}^{16}O$[35]	$D_2{}^{16}O$[35]	$T_2{}^{16}O$[36]	$H_2{}^{18}O$ [37]	$H_2{}^{17}O$
"Isolated"	3712.87 3614.02 1612.52	2720.38 2605.26 1180.16	2296.99 2169.23 992.74	3697.74 2606.39 1605.83	3704.85 3609.98 1608.98
Gas	3755.8 3656.7 1594.6	2788.1 2671.5 1178.3	2370.0 - 995.5	3741.6 3649.7 1588.2	- - -

The calculated frequencies differ from the exact measured gas phase frequencies but we emphasize that the point which interests us is the calculation of properly weighted gas-liquid frequency shifts, and not the duplication of the exact gas phase frequencies. In addition, it should be noted that the numerical value of R_{max} (0.318 nm in this paper) is arbitrary; if a larger value were to be chosen the gas phase frequencies might be better approximated.

Another important feature of the present calculation of the VPIE is that the whole intermolecular frequency region is treated as a group in order to avoid difficulties of assignment. This can be achieved by constructing the last term in Eqn. (13) in an appropriate manner.

In terms of the above approximations the computation of the VPIE was performed as follows. First $\ln(s/s'f)_{gas}$ was calculated by using the frequencies of the "isolated" molecules (Table 5.). Then, assuming that 20000 intramolecular configurations give reliable statistics, we computed the first term in Eqn. (13). Later the last ("intermolecular") term was calculated by generating 50 intermolecular configurations. We then tested the statistical convergence by increasing the number of configurations by units of 50, calculating the VPIE at each stage from Eqn. (13) with the a_{ji} and a_i values consistent with Eqn. (14), until we arrived at 250 configurations. Intermolecular bending force constants used by Bandekar and Curnutte were adjusted (see Table 2.) to obtain convergence of the VPIE values by increasing the number of intermolecular configurations. The converged values are given in Table 6 and compared with experiment [38] and the earlier cell model calculations [39]. The convergence of the process is presented graphically in Fig. 10. The convergence is apparently good to ± 0.002 unit at or above 150 configurations.

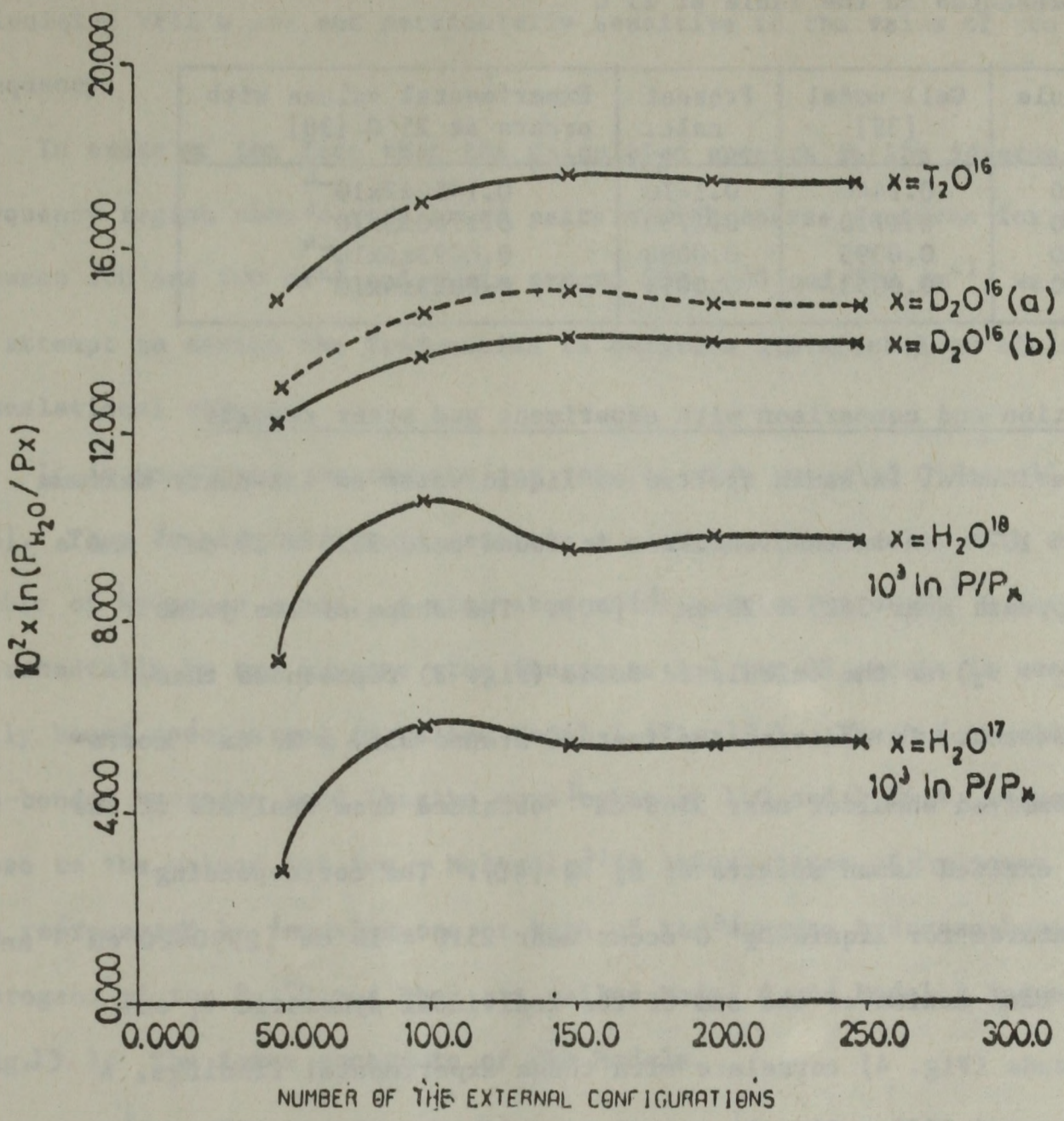


Fig. 10. Calculated VPIE values as a function of the number of intermolecular configurations (25°C)

(a) Intermolecular force constants: $f_\rho = 0.192 \times 10^2 Nm^{-1}$, $f_\sigma/r_0R_0 = 0.037 \times 10^{-8} Nrad^{-1}$, $f_\tau/r_0R_0 = 0.03675 \times 10^{-8} Nrad^{-1}$ (see Table 2).

(b) Intermolecular force constants: $f_\rho = 0.211 \times 10^2 Nm^{-1}$ $f_\sigma/r_0R_0 = 0.03816 \times 10^{-8} Nrad^{-1}$, $f_\tau/r_0R_0 = 0.03775 \times 10^{-8} Nrad^{-1}$

Table 6. Calculated and measured VPIE values. The values of $\ln(P_{H_2O}/P_x)$ are presented in the Table at 25°C

Molecule (X)	Cell model [39]	Present calc.	Experimental values with errors at 25°C [38]
$D_2{}^{16}O$	0.1440	0.1410	$0.1450\pm2\times10^{-4}$
$T_2{}^{16}O$	0.1770	0.1750	$0.1750\pm5\times10^{-4}$
$H_2{}^{18}O$	0.0095	0.0098	$0.0093\pm6\times10^{-4}$
$H_2{}^{17}O$	0.0053	0.0054	$0.0053\pm4\times10^{-4}$

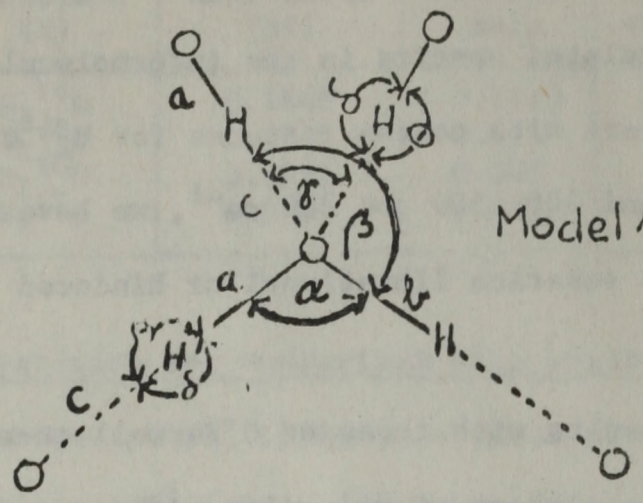
IV. Interpretation and comparison with experiment and other results

In the experimental IR/Raman spectra of liquid water an intensity maximum occurs at 3410 ± 10^{-1}, an intense shoulder is found near 3225 ± 20 cm^{-1} and a weak shoulder appears near 3625 ± 20 cm^{-1} [40]. The shape of the joint distribution ($\nu_1 + \nu_3$) of the calculated bands (Fig. 3) reproduces these experimental features. In addition the feature around 3550 ± 20 cm^{-1} correlates with an observed shoulder near 3565 cm^{-1} obtained from analysis of the argon-ion-laser excited Raman spectra of $H_2{}^{16}O$ [40]. The corresponding experimental features for liquid $D_2{}^{16}O$ occur near 2510 ± 10 cm^{-1}, 2390 ± 20 cm^{-1} and 2675 ±20 cm^{-1}[40].The band maxima of the sum of the individual symmetric ν_1 and asymmetric ν_3 bands (Fig. 4) correlate with these experimental findings. A fourth feature around 2580 ± 20 cm^{-1} correlates with the maxima 2590 cm^{-1} obtained by computer analysis of laser-Rman spectra in the OD stretching region from liquid $D_2{}^{16}O$ [40]. It should be emphasized that the above assignments are approximate. It is impossible to factor the experimental frequency distributions of HOH and DOD in the stretching region into individual ν_1 and ν_3 contributions. Fig. 5 reviews the calculated spectra in the bending region. Using Eqn. (8) the ν_2 band maxima are calculated and are found to be in agreement with experiment for $H_2{}^{16}O$ and $D_2{}^{16}O$, though the peak maximum for $D_2{}^{16}O$ is in less satisfactory agreement with the measurement [32]. This

indicates a short-coming in Eqn. (8). Even so it is to be noted that the calculated VPIE's are not particularly sensitive to the value of the bending frequency.

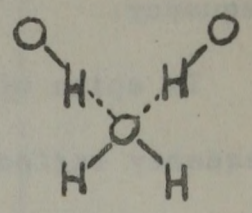
In spite of the fact that the calculated spectra in the intermolecular frequency region show a very broad pattern with coarse features for $H_2{}^{16}O$ between 100 and 200 cm^{-1} and again around 300, 500 and 800 cm^{-1}, we have made no attempt to assign the frequencies to separate librational or hindered translational classes.

It is worthwhile to compare our results with those of O'Ferrall et al. [16]. They considered liquid water as a mixture of molecules with a different number of hydrogen bonds. A nine-atom model consisting of $H_2{}^{16}O$ surrounded tetrahedrally by two oxygens atom fragments and two OH groups is used for fully bound species and is called Model 1 (Fig.12.). The O-H covalent and non-bonded hydrogen bond lengths were taken as 1.0 and 1.76 Å, respectively, close to the values for ice. Molecules in lower states of hydrogen bonding are represented by removing one or both of the oxygens hydrogen-bonded to the hydrogens of the $H_2{}^{16}O$ and they are called Model 4 and Model 5 respectively (Fig.13.). The force constants of the Models

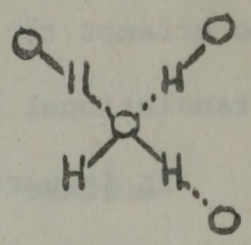


Model 1 [16]

Fig. 12. Fully hydrogen-bonded water



Model 4 [16]



Model 5 [16]

Fig. 13. Molecules in lower states of hydrogen bonding

are compared with ones used by us in Table 7 and the calculated internal and external frequencies for Models 1, 4 and 5 are presented in Table 8. The logarithms of the vapor pressure ratio, $P_{H_2}{}^{16}O/P_{D_2}{}^{16}O$, the cited authors obtained for Models 1 and 5 at 25°C are $\ln(P_{H_2}{}^{16}O/P_{D_2}{}^{16}O) = 0.241$ and $\ln(P_{H_2}{}^{16}O/P_{D_2}{}^{16}O) = 0.176$ respectively; the experimental value is 0.142 [41,42]. Table 6 clearly shows the consistency of our calculation as far as the effects in $D_2{}^{16}O$, $T_2{}^{16}O$, $H_2{}^{18}O$ and $H_2{}^{17}O$ are concerned. The calculated value of $\ln(P_{H_2}{}^{16}O/P_{D_2}{}^{16}O)$ obtained by us is more reasonable than the one calculated by O'Ferrall et al. This can be attributed to certain advantages of the present continuum model. The first is the range of constants describing coupling to neighbors. This is explicitly included by making the force constant a function of the nearest neighbor distances with the L.S. potential. Another advantage of the present continuum model is that the entire

Table 7. Force constants used in this paper compared with those used by O'Ferrall et al. [16]

<table>
<tr><td colspan="3">Force Constants used in this paper (R = 2.86Å)</td></tr>
<tr><td colspan="2">$f_r = 6.55\times10^5$ dynes/cm
$f_\alpha = 0.2019\times10^6$ dynes/cm</td><td>Stretching: $f_\rho = 1.92\times10^4$ dynes/cm
"In-plane"
bending: $G = 3.700\times10^3$ dynes/cm</td></tr>
<tr><td colspan="2">$f_{rr} = -0.0794\times10^5$ dynes/cm
$f_{r\alpha} = 0.3877\times10^5$ dynes /cm</td><td>"Out-of-plane"
bending: $H = 3.675\times10^3$ dynes/cm</td></tr>
<tr><td colspan="3">Force constants used by O'Ferrall et. al. [16]</td></tr>
<tr><td colspan="3">(The subscripts for the force constants follow the notation in Model 1.)</td></tr>
<tr><td>Model 1</td><td colspan="2">$f_a = f_b = 6.55\times10^5$ dynes/cm
$f_c = 0.12\times10^5$ dynes/cm
$f_{ab} = -0.25\times10^5$ dynes/cm
$f_\alpha = 0.65\times10^5$ dynes/cm
$f_\beta = 0.02163\times10^5$ dynes/cm
$f_\gamma = 0.09685\times10^4$ dynes/cm
$f_\phi = f_\delta = 0.02163\times10^5$ dynes/cm</td></tr>
<tr><td>Model 4</td><td colspan="2">$f_a = f_b = 7.35\times10^5$ dynes/cm
$f_\alpha = 0.73\times10^5$ dynes/cm
The other force constants were as used in Model 1.</td></tr>
<tr><td>Model 5</td><td colspan="2">$f_a = 7.35\times10^5$ dynes/cm
$f_b = 6.55\times10^5$ dynes/cm
$f_\alpha = 0.69\times 10^5$ dynes/cm
The other force constants were as used in Model 1.</td></tr>
</table>

Table 8. Calculated vibration frequencies for water with different number of hydrogen bonded hydroxyl group obtained by O'Ferrall et al [16]

Frequencies cm^{-1}	Model 1		Model 4		Model 5	
	$H_2{}^{16}O$	$D_2{}^{16}O$	$H_2{}^{16}O$	$D_2{}^{16}O$	$H_2{}^{16}O$	$D_2{}^{16}O$
ν_1	3355	2425	3585	2590	3405	2470
ν_3	3550	2605	3670	2695	3675	2690
ν_2	1645	1200	1645	1205	1650	1200
ν_{lib}	715	530	410	307	710	515
	710	500	355	254	420	300
	645	470	230	173	380	280
ν_{def}	530	380	620	440	610	445
	530	380	600	435	600	435
	530	380	580	425	580	420
	(580)	(420)	580	420	530	380
ν_{trans}	168	161			162	155
	168	161	158	148	161	154
	167	161	137	132	126	122
	105	104				
ν_{bend}	54-35	53-55	41	41	59-39	58-38

external frequency region can be treated as a whole, thus avoiding the separation of the librations which, as the above authors mention, "is largely artificial because there is a strong coupling between the modes"[16]. Finally it should be mentioned that in our method the computation of VPIEs was mainly achieved by adjusting the value of the in-plane and out-of--plane bending force constants. Apart from the relative contributions of the intramolecular frequencies, the calculated VPIEs particularly depend on the librational modes. Their magnitude was determined principally by the in-plane and out-of-plane bending constants. The librational modes lie at the top of the intermolecular band and display more isotope dependence than do the hindered translations.

It would seem that the present continuum model is better suited to describing the main spectroscopic features of liquid water and to maintaining agreement between calculated and experimental VPIEs than the simple mixture model as used by O'Ferrall et al. It is of course vastly superior to the simple cell model [39] which makes no pretense of spectroscopic accuracy.

References

[1] W.K. Roentgen, Ann.Phys.Chim. 45,92/1892/

[2] G.Nemethy and H.A. Scheraga, J.Chem.Phys. 36,3382/1962/

[3] A.Ben-Naim, J.Phys.Chem. 69,1922,3240/1965/

[4] R.P. Marchi and H. Eyring, J.Chem.Phys. 38,221/1964/

[5] J. Lennard-Jones and J.A. Pople, Proc.Roy.Soc. A205,155/1951/

[6] J.A. Pople, Proc.Roy.Soc. A205,163/1951/

[7] D.N. Glew, J.Phys.Chem. 66, 605/1962/

[8] M. Falk and T.A. Ford, Can.J.Chem. 44,1699/1966/

[9] M. Falk and H.R. Wyss J.Chem.Phys. 51,5727/1969/

[10] H.R. Wyss and M. Falk, Can.J.Chem. 48, 607/1970/

[11] J. Bigeleisen, J.Chem.Phys. 34,1435/1961/

[12] M.J. Stern, W.A. Van Hook and M. Wolfsberg, J.Chem.Phys. 39,3179/1963/

[13] H.S. Frank and W.Y. Wen, Discussions, Faraday Soc. 24,133/1957/

[14] H.S. Frank, Proc.Roy.Soc. London Ser. A 247,481/1958/

[15] F. Hajdu, Acta Cryst. A31,157/1975/

[16] R.A. More O'Ferrall, G.W. Koeppl and A.J. Kresge, J.Amer.Chem.Soc., 93, 1/1971/

[17] J.B. Bryan, "A Normal Coordinate Analysis of the Local Structure of Liquid Water for Interpretation of Far Infrared Spectra," Ph.D. Dissertation, Kansas State University, Manhattan, 1969

[18] J.A. Barker, Ann.Rev.Phys.Chem. 14,2145/1965/

[19] B. Curnutte and J. Bandekar, J.Mol.Spectrosc., 41,500/1972/

[20] J. Bandekar and B. Curnutte, J.Mol.Spectrosc., 58,169/1975

[21] W.H. Schaffer and R.P. Schumann, J.Chem.Phys., 12,504/1944/

[22] E.R. Lippincott and R. Schroeder, J.Chem.Phys., 23,1099/1955/

[23] N. Metropolis, A.W. Rosenbluth, M. H. Rosenbluth, A.H. Teller, E. Teller, J. Chem. Phys. 21, 1087 /1953/

[24] G.S. Kell, Water a Comprehensive Treatise /Ed.: Felix Franks/Plenum Press /1975/p. 377

[25] A.H. Narten, M.D. Danford and H.H. Levey, X-Ray Diffraction Data on Liquid Water in the Temperature Range 4° to 200°C. Oak Ridge National Laboratory, Oak Ridge, TN, ORNL-3997, UC-4./1966/

[26] A.H. Narten and H.H. Levey, J.Chem.Phys., 55,2263/1971/

[27] A.H. Narten, J.Chem.Phys., 56,5681/1972/

[28] I.S. Bendat and A.G. Piersol, Measurement and Analysis of Random Data. John Wiley, New York, pp. 309-310/1966/

[29] R. Schroeder and E.R. Lippincott, J.Chem.Phys. 61,921/1957/

[30] J.N. Bandekar, A Monte Carlo normal coordinate analysis treatment of intermolecular vibrations in liquid water, Ph.D. Dissertation, Kansas Sate University, Manhattan/1973/

[31] G.E. Walrafen in: "Hydrogen-bonded Solvent Systems" /Eds.: Covington A.K. and Jones P./ Taylor and Francis, London, 1968, pp. 9-30.

[32] G.E. Walrafen, J.Chem.Phys., 40,3249/1964/; 47,114/1967/

[33] G.J. Safford, P.S. Leung, A.W. Naumann and P.C. Schaffer, J.Chem.Phys., 50,4444/1969/

[34] C. Robertson, B. Curnutte, D. Williams, Mol.Phys., 26,183/1973/

[35] T. Shimanouchi and I. Suzuki, J.Chem.Phys., 42,296/1964/

[36] S. Pinchas and I. Laulicht, Infrared Spectra of Labelled Compounds, Academic Press, London, 1971, pp. 44-45

[37] M. Majoube, J.Chim.Phys., 68,1423/1971/

[38] G. Jancso and W.A. Van Hook, Chem.Rev., 74,689/1974/

[39] W.A. Van Hook, J.Phys.Chem., 72,1234/1968/; J.Phys.Chem., 76,3040/1972/

[40] G.E. Walrafen, Water. A comprehensive Treatise /Ed: Felix Franks/ Plenum Press 1975, pp. 197-199

[41] W.M. Jones, J.Chem.Phys., 48,207,/1968/

[42] L. Merlivat, R. Botter, G. Nief, J.Chim.Phys. Physicochim.Biol., 60,56/1963/; F.D. Rossini, J.W. Knowlton, H.L. Johnston, J.Res.Nat.Bur.Stand., 24,369/1940/

ON THE TRIANGLE PROPERTY AND REPRESENTING THE INVERSE OF TRIDIAGONAL MATRICES

CSABA J. HEGEDÜS
Central Research Institute for Physics
H-1525 Budapest 114, P.O.Box 49, Hungary

ABSTRACT

A general definition is suggested for the triangle property of matrices and the structure of such matrices is investigated. It is shown that the inverse of a tridiagonal matrix can be represented by a sum of lower and upper triangular halves of rank 1 matrices, where the triangular matrices are fully determined by the elements on the perpendicular sides. The representation also shows the location of zeros. In connection with the zeros the zero angulus property is introduced and an alternative representation is derived. The symmetrization properties lead to the introduction of composite Green's matrices.

1. INTRODUCTION

The ordinary inverse of a tridiagonal matrix has the property that the rank of all submatrices taken from the lower or upper triangular half of the matrix, may not exceed 1 . This is a consequence of a theorem by E. Asplund [1] concerning the inverse of band matrices. The property has been rediscovered recently by Barrett [2] and by Barrett and Feinsilver [3] in the general form.

Barrett [2] called this property the triangle property for a special case, because it could be associated with triangular patterns. The author believes the term appropriate so much the more that the results of this paper yield further evidences in favour of the naming, however, he thinks that the definition should be given a general form.

The present paper suggest a general definition for the triangle property and investigates the strucure of matrices which have this property.

Section 2 is an introductory section. It recalls some concepts that will be necessary in subsequent sections. Moreover, a suggestion is made for the definition of pseudo-symmetric matrices.

Section 3 defines the triangle property in its general form and gives the first representation theorem for such matrices.

Section 4 defines the angulus of a matrix and introduces the zero angulus property. For matrices that have the triangle property and the zero angulus property, a second representation theorem is stated.

Section 5 investigates symmetrization. It is shown that similar symmetrization properties hold to that of tridiagonal matrices if the zero angulus property is also assumed. At last the Green's matrices are generalized for the pseudo-symmetric case.

N o t a t i o n s

$A=[a_{ij}]$ or $[A_{pq}]$. Matrix A is given by the elements a_{ij} or by the submatrices A_{pq}:

$a(i_k, j_m)$ A matrix element for double subscripts that belongs to the i_kth row and j_mth column of matrix A.

e_i, $e(i_k)$ The ith and i_kth Cartesian unit vectors.

I_q Unit matrix of order q.

$D=\langle d_i \rangle$ Diagonal matrix defined by the elements d_i.

δ_{ij} Kronecker symbol.

Superscripts T and H denote the transpose and conjugate transpose of a matrix or vector.

sgn ( ) Sign function

$S=[(1+\delta_{ij}+\mathrm{sgn}(i-j))/2]$ A lower triangular matrix consisting of 1's in the lower half.

$\tilde{S} = S^T - I$ The complement to S.

$A \circ B = [a_{ij} b_{ij}]$ The Hadamard or logical product.

$A[i_1, i_2, \ldots, i_k; j_1, j_2, \ldots, j_k]$ The minor defined by selecting the rows $i_1, i_2, \ldots, i_k$ and columns $j_1, j_2, \ldots, j_k$ from matrix A.

2. SYMMETRIZATION OF TRIDIAGONAL MATRICES

We recall the definitions of tridiagonal and reducible matrices and suggest a definition for pseudo-symmetric matrices. Theorem 2.6 states that any tridiagonal matrix can be factored into a pseudo--symmetric and a diagonal matrix.

DEFINITION 2.1. A real or complex matrix A is tridiagonal if $a_{ij}=0$ for $|i-j|>1$.

DEFINITION 2.2. A real or complex matrix A is reducible if there is a permutation matrix P so that

$$PAP^T = \begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix},$$

where B_{11} and B_{22} are square matrices; A is irreducible if it is not reducible.

As a consequence, any reducible matrix can be brought to an upper block triangular form by some permutation matrix Q:

$$QAQ^T = \begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1s} \\ 0 & B_{22} & \cdots & B_{2s} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & B_{ss} \end{bmatrix}, \qquad (2.1)$$

where the diagonal blocks $B_{11},\ldots,B_{ss}$ are irreducible or zero square matrices.

DEFINITION 2.3. For any matrix A and a permutation matrix Q, the matrix QAQ^T is called a basic form of A if QAQ^T is of a block upper triangular form with irreducible or zero square matrices in the diagonal blocks. If matrix A is irreducible then it is thought to be of a special block upper triangular form consisting of one block only.

DEFINITION 2.4. Matrix A is said pseudo-symmetric if all diagonal blocks in its basic form are symmetric.

This definition is a common generalization of triangular and symmetric matrices. Clearly, pseudo-symmetric matrices have real eigenvalues and their inversion needs only the inversion of symmetric matrices. Pseudo-hermitian matrices can be defined in a similar manner.

THEOREM 2.5. An nxn tridiagonal matrix A is irreducible if and only if $a_{i\,i+1}a_{i+1\,i} \neq 0$ for $i=1,2,\ldots,n-1$.

For the proof of this theorem, see [4]. ∎

THEOREM 2.6. Let A be an nxn tridiagonal matrix. Then it can be factored as

$$A = DC, \tag{2.2}$$

where $D = \langle d_1,\ldots,d_n \rangle$ is a nonsingular diagonal matrix and matrix C is pseudo-symmetric.

Proof. If A is irreducible then $C=D^{-1}A$ is symmetric if $d_{i+1}^{-1}a_{i+1\ i} = d_i^{-1}a_{i\ i+1}$. Choose $d_1=1$, $d_{i+1}= d_i a_{i+1\ i}/a_{i\ i+1}$, $i=1,2,\ldots,n-1$ then the representation is complete.

If A is reducible, decompose it into irreducible diagonal blocks and apply the procedure as before to each irreducible block. If there are zero diagonal blocks, choose 1's for the corresponding d_i elements.

In order to show that matrix $D^{-1}A=C$ is pseudo-symmetric if its irreducible diagonal blocks are symmetric, partition matrix C symmetrically so that

$$C = \begin{bmatrix} C_{11} & 0 \\ C_{21} & C_{22} \end{bmatrix},$$

where the qxq submatrix C_{11} is of a basic form. Then apply the transformation

$$\begin{bmatrix} 0 & I_{n-q} \\ I_q & 0 \end{bmatrix} \begin{bmatrix} C_{11} & 0 \\ C_{21} & C_{22} \end{bmatrix} \begin{bmatrix} 0 & I_q \\ I_{n-q} & 0 \end{bmatrix} = \begin{bmatrix} C_{22} & C_{21} \\ 0 & C_{11} \end{bmatrix}.$$

If C_{22} is still not of a basic form, repeat the procedure for C_{22} and so on, until the basic form is complete. The submatrix C_{21} may still be effected in the subsequent steps, however, it is seen that only the locations of the diagonal blocks are changed , hence matrix C is pseudo--symmetric. ■

3. THE TRIANGLE PROPERTY

DEFINITION 3.1. An nxn matrix R has the triangle property if all its 2x2 minors that do not spread across but may touch the main diagonal are zero:

If $i_1 \leq i_2 \leq j_1 \leq j_2$ or $j_1 \leq j_2 \leq i_1 \leq i_2$ then

$$R[i_1,i_2;j_1,j_2] = r(i_1,j_1)r(i_2,j_2) - r(i_1,j_2)r(i_2,j_1) = 0 \tag{3.1}$$

This definition is a generalization of Barrett's definition [2]. Here $i_2=j_1$ or $j_2=i_1$ are not demanded and the additional constraints $r_{ii} \neq 0$, $i=2,3,\ldots,n-1$ are not imposed. The cases $i_1=i_2$ and $j_1=j_2$ are trivial.

The relational sequence can easily be memorized by placing the indices i_1,i_2 and j_1,j_2 next to each other and supplying with $\leq$ signs.

The term "triangle property" can be associated with the property that an element of the matrix may be expressed in terms of three others that lie on the vertices of a triangle.

THEOREM 3.2. Let the nxn matrix A be tridiagonal and invertible: $A^{-1} = R$. Then matrix R has the triangle property if and only if A is tridiagonal.

Proof. This is a special case of Asplund's theorem [1]. See also Barrett and Feinsilver [3].

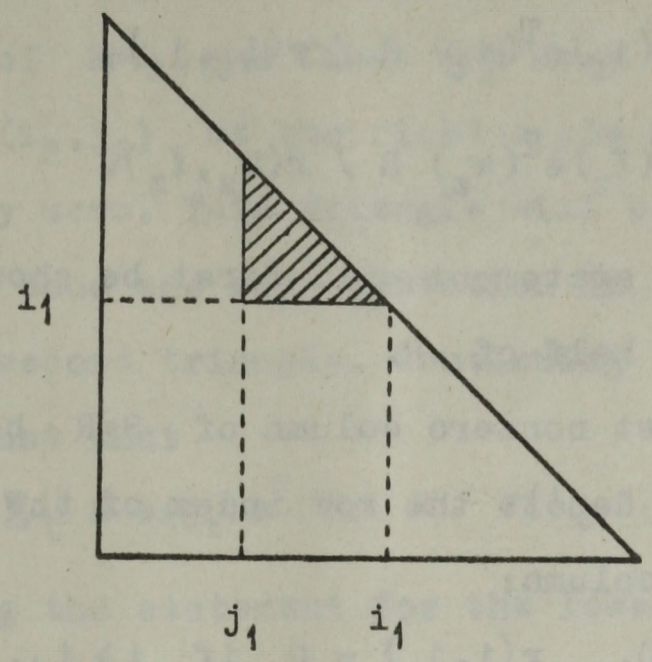


Fig. 1

For matrices that have the triangle property, a simple expansion can be given in terms of triangular halves of rank 1 matrices. For this we shall make some conventions.

Let S be a lower triangular matrix such that $s_{ij}=1$, $i \geq j$; and let $\tilde{S}=S^T-I$. For a double subscript we denote the i_tth Cartesian unit vector by $e(i_t)$ and also, we use the notation $r(i_t,j_t)= e^T(i_t)Re(j_t)$ for a matrix element that has a double subscript. The Hadamard product $S \circ R$ is defined by $S \circ R = [s_{ij}r_{ij}]$.

THEOREM 3.3. Let matrix R have the triangle property. Then there exist pairs of indices i_t,j_t and k_z,ℓ_z for some t and z such that

$$R = S \circ L + \tilde{S} \circ U \,, \tag{3.3}$$

$$I \circ L = I \circ U \,, \tag{3.4}$$

where

$$L = \sum_t R\, e(j_t) e^T(i_t)\, R \,/\, r(i_t, j_t),$$

$$U = \sum_z R\, e(\ell_z) e^T(k_z)\, R \,/\, r(k_z, \ell_z).$$

Proof. The statement will first be shown for the lower triangular half of R.

Let the first nonzero column of $S \circ R$ be indexed by j_1 and let i_1 denote the row index of the last nonzero element in this column:

$$r(i_1, j_1) \neq 0, \quad r(i, j_1) = 0 \quad \text{if} \quad i > i_1. \tag{3.5}$$

Then one has the following implications by (3.1):

$$j_1 \leq j \leq i_1 < i \;\Rightarrow\; R[i_1, i;\, j_1, j] = 0$$

$$\Rightarrow \quad r_{ij} = r(i, j_1)\, r(i_1, j) / r(i_1, j_1) = 0 \tag{3.6}$$

because of (3.5). Moreover,

$$j_1 \leq j \leq i \leq i_1 \Rightarrow R[i, i_1;\, j_1, j] = 0$$

$$\Rightarrow \quad r_{ij} = r(i, j_1)\, r(i_1, j) / r(i_1, j_1). \tag{3.7}$$

Hence matrix $S \circ R$ shows the picture as in Fig. 1 up to the i_1th column, where the shaded triangle has nonzero elements only. Relation (3.7) shows that the elements of this triangle are fully determined by the elements on the perpendicular sides. Now it is simple to check that one has

$$r_{ij} = \left\{ S \circ \left(R\, e(j_1)\, e^T(i_1)\, R \,/\, r(i_1, j_1) \right) \right\}_{ij}, \quad j \leq \min(i, i_1) \tag{3.8}$$

and this relation represents the elements of $S \circ R$ up to the i_1th column.

In the next step one repeats the same procedure for the remainder of $S \circ R$ and finds the next triangle with the element $r(i_2, j_2)$ at the right angle if the remainder is not entirely zero. This triangle will be disjoint to the first one, thus the representation in (3.8) can be enlarged by a second triangle. Continuing the process, one finds at last that

$$S \circ R = S \circ (\Sigma_t \; R \; e(j_t) e^T(i_t) \; R \; / \; r(i_t, j_t)) \qquad (3.9)$$

thereby proving the statement for the lower triangular half of R.

For the upper triangular half one can proceed similarly such that one finds

$$S^T \circ R = S^T \circ (\Sigma_z \; R \; e(\ell_z) e^T(k_z) \; R \; / \; r(k_z, \ell_z)) \qquad (3.10)$$

for some k_z and ℓ_z.

Relation (3.4) states the equality of the diagonal elements of R from both representations (3.9) and (3.10). ■

One can draw some conclusions from this theorem.

The inverse of a reducible tridiagonal matrix has the form as shown in Fig. 2, where the nonzero area is shaded. Zero off-diagonal blocks in the inverse correspond to zero off-diagonal elements in the tridiagonal matrix. Each lower and upper triangle is fully determined by the elements on the perpendicular sides of the triangles, hence the inverse

is determined by the boundary elements of the nonzero area. Note that some of the triangles may degenerate to a $1x1$ matrix. The elements lying at the right angles play a special role. They are in the denominators of (3.9) and (3.1o), they are nonzero, and their indices specify the perpendicular sides of the triangles.

DEFINITION 3.4. For a matrix of the triangle property the nonzero elements that lie at the right angles of the triangular blocks in the nonzero area (as marked by small squares in Fig.2) will be called the vertex elements of the matrix.

DEFINITION 3.5. The nonzero area of a matrix that has the triangle property, is defined as follows: From each vertex element draw a vertical and a horizontal line to the diagonal. The unification of all these triangles form the nonzero area.

If the nxn matrix R has the triangle property and it is irreducible then $r_{1n} \neq 0$ and $r_{n1} \neq 0$. Consequently, if matrix A is tridiagonal and irreducible then its inverse is determined by two triangles:

$$A^{-1} = S \bullet (A^{-1} e_1 e_n^T A^{-1} / e_n^T A^{-1} e_1) + \tilde{S} \bullet (A^{-1} e_n e_1^T A^{-1} / e_1^T A^{-1} e_n) \tag{3.11}$$

If A is also symmetric then its inverse is a Green's matrix:

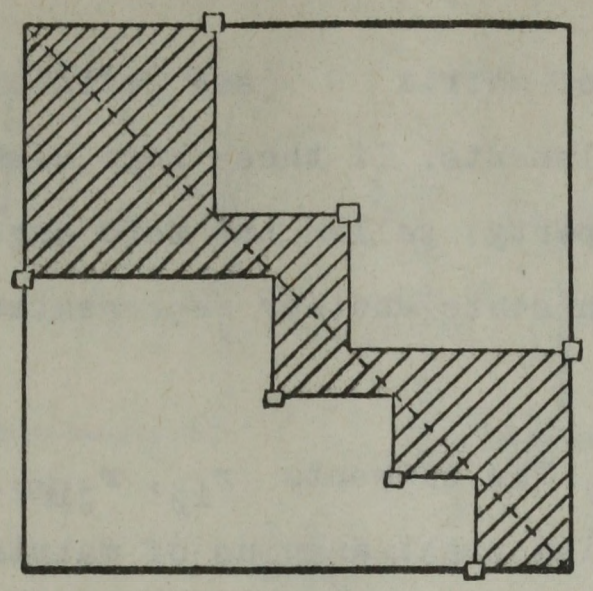


Fig. 2

DEFINITION 3.6. For two n-dimensional real or complex vectors a and b the Green's matrix is formed by

$$G(a,b) = [a_{\max(i,j)} b_{\min(i,j)}] = S\circ(ab^T) + \tilde{S}\circ(ba^T). \quad (3.12)$$

G(a,b) is symmetric, its first and last columns are $b_1 a$ and $a_n b$.

Hermitian Green's matrices can also be defined.

DEFINITION 3.7. For two complex vectors a and b that satisfy $\bar{a}_i b_i = a_i \bar{b}_i$ for all i, a Hermitian Green's matrix is formed by

$$G_H(a,b) = S\circ(ab^H) + \tilde{S}\circ(ba^H), \quad (3.13)$$

where the superscript H stands for the conjugate transpose.

For Hermitian and irreducible tridiagonal matrices one has

$$A^{-1} = G_H(A^{-1}e_1/(A^{-1})_{1n}, A^{-1}e_n), \quad (3.14)$$

Here the subscript H can be omitted if A is symmetric.

4. THE ZERO ANGULUS PROPERTY

The nonzero area of matrix R (see Definition 3.5) may still have zero elements. If these zero elements satisfy a certain property, called the zero angulus property, one can then state another representation theorem.

DEFINITION 4.1. The elements r_{ij}, r_{ji}, $j=1,2,\ldots,i$ are said to form the ith right angulus of matrix R. Similarly, the elements r_{ij}, r_{ji}, $j=i,i+1,\ldots,n$ form the ith left angulus of the nxn matrix R as shown schematically in Fig. 3a) and 3b).

DEFINITION 4.2. Let matrix R have the triangle property. We say that R has the zero angulus property if any zero in the nonzero area of R is located at an entirely zero left or right angulus.

LEMMA 4.3. Let matrix R have the triangle property so that there are no zero rows or columns in R. Then matrix R has the zero angulus property.

Proof. Choose a diagonal element r_{ii}. Assume it belongs to the upper triangle with vertex element r_{jk} and to the lower triangle with vertex element $r_{\ell m}$ (see Fig. 4, the nonzero area is shaded). Then one has, by the triangle property,

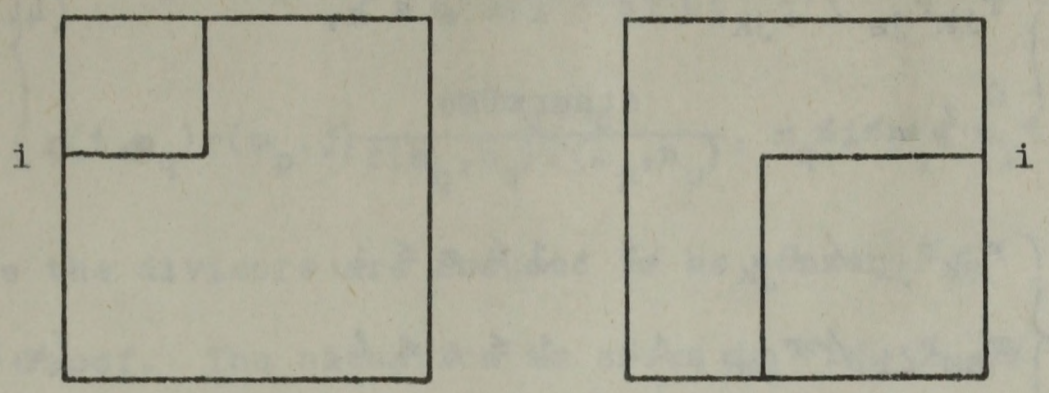


Fig. 3

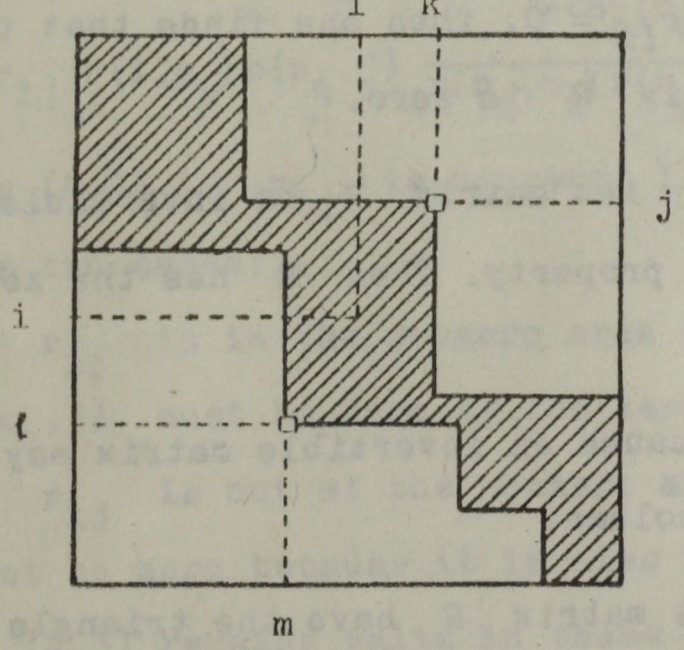


Fig. 4

$$r_{is} = \begin{cases} r_{im}r_{\ell s} / r_{\ell m}, & \text{if} \quad m \leq s \leq i, \\ r_{ik}r_{js} / r_{jk}, & \text{if} \quad i \leq s \leq k, \\ 0, & \text{otherwise} \end{cases} \tag{4.1}$$

and

$$r_{si} = \begin{cases} r_{sk}r_{ji} / r_{jk}, & \text{if} \quad j \leq s \leq i, \\ r_{sm}r_{\ell i} / r_{\ell m}, & \text{if} \quad i \leq s \leq \ell, \\ 0, & \text{otherwise.} \end{cases} \tag{4.2}$$

In particular,

$$r_{ii} = r_{im}r_{\ell i} / r_{\ell m} = r_{ik}r_{ji} / r_{jk}.$$

Now if $r_{im} = 0$ then r_{ik} or r_{ji} is zero. Assume that r_{ik} is zero. Then the ith row of matrix R is entirely zero by (4.1) and this contradicts a hypothesis of the lemma. Consequently, $r_{ji} = 0$ and the ith right angulus of R is zero by (4.1) and (4.2).

Similarly, if $r_{\ell i} = 0$, then one finds that the ith left angulus of matrix R is zero. ∎

COROLLARY 4.4. Let matrix R be invertible and let it have the triangle property. Then R has the zero angulus property.

This follows because an invertible matrix may not have a zero row or column.

LEMMA 4.5. Let matrix R have the triangle property and the zero angulus property, then

$$r_{ij}=\begin{cases} r(i,m_p)\,r(n_q,j)\,\dfrac{r(n_p,m_q)}{r(n_q,m_q)\,r(n_p,m_p)}, & m_q\le j\le n_q < m_p\le i\le n_p, \quad (4.3)\\[2ex] r(i,n_q)\,r(m_p,j)\,\dfrac{r(m_q,n_p)}{r(m_q,n_q)\,r(m_p,n_p)}, & m_q\le i\le n_q < m_p\le j\le n_p, \quad (4.4)\end{cases}$$

where the divisors are assumed to be nonzero.

Proof. The situation is shown in Fig. 5. By virtue of the triangle property one has

$$R[n_q,i;m_q,j] = R[i,n_p;j,m_p] = R[i,n_p;m_q,j] = 0.$$

Writing out each of them yields

$$r_{ij} = r(i,m_q)\,r(n_q,j)/r(n_q,m_q), \quad (4.5)$$

$$r_{ij} = r(i,m_p)\,r(n_p,j)/r(n_p,m_p), \quad (4.6)$$

$$r(i,m_q)\,r(n_p,j) = r_{ij}\,r(n_p,m_q). \quad (4.7)$$

Multiplying (4.5) and (4.6) and substituting (4.7) gives

$$r_{ij}^2 = r_{ij}\,r(i,m_p)\,r(n_q,j)\,\frac{r(n_p,m_q)}{r(n_q,m_q)\,r(n_p,m_p)}$$

which yields (4.3) if r_{ij} is nonzero. If r_{ij} is zero, there may be two cases:

i) If r_{ij} is in the nonzero area then $r(i,m_p)$ or $r(n_q,j)$ must be zero by the zero angulus property.

ii) If r_{ij} is out of the nonzero area then $r(n_p,m_q)$ must be zero because it is also in the zero area.

Hence, (4.3) remains valid in these cases. The other case is completely analogous. ∎

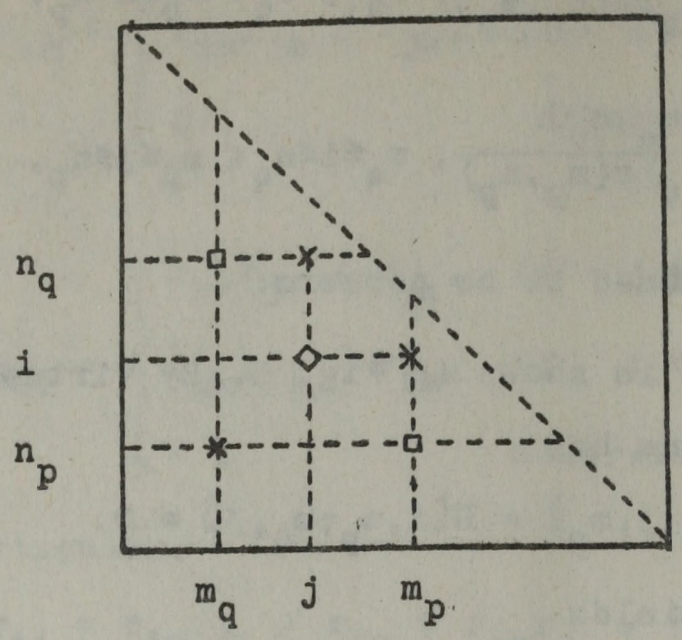


Fig. 5

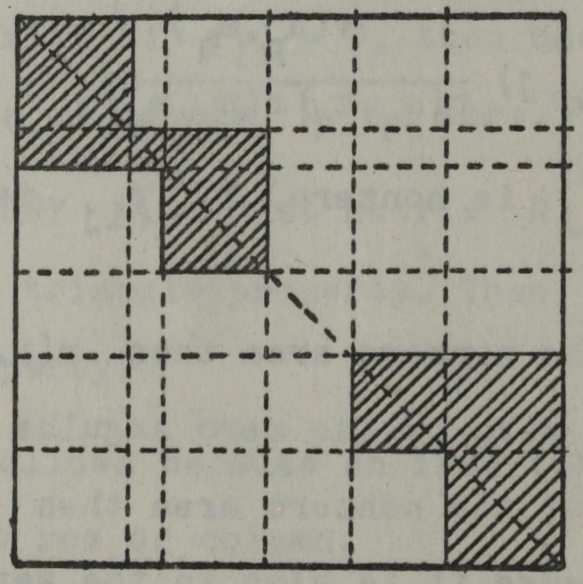


Fig. 6

THEOREM 4.6. Let the nxn matrix R have the triangle property and the zero angulus property. Let R be partitioned symmetrically with respect to the index sets $\bigcup_{k=1}^{s} N_k = \{1,2,\ldots,n\}$, $N_k = \{m_k, m_k+1, \ldots, n_k\}$, $k=1,2,\ldots,s$, $m_1=1$, $m_{k+1}=n_k+1$, $n_s=n$ so that it is done along the perpendicular sides of the determining triangles (see Theorem 3.3). Then

$$r_{ij} = \begin{cases} \dfrac{r(i,m_p)\,r(n_q,j)}{r(n_q,m_q)} \prod_{k=q+1}^{p} \dfrac{r(n_k,m_{k-1})}{r(n_k,m_k)}, & i \geq j, \ (i,j) \in N_p \times N_q \quad (4.8) \\ \dfrac{r(i,n_q)\,r(m_p,j)}{r(m_q,n_q)} \prod_{k=q+1}^{p} \dfrac{r(m_{k-1},n_k)}{r(m_k,n_k)}, & i \leq j, \ (i,j) \in N_q \times N_p \quad (4.9) \\ 0, \text{ if some of the divisors in the previous formulae are zero.} & (4.10) \end{cases}$$

The product is taken to be 1 if the lower limit exceeds the upper limit.

Proof. Assume that matrix R has a composite structure as given by Theorem 3.3. Under the assumptions of the theorem, the nonzero area may be split into separate regions as shown in Fig. 6, where two connected regions can be seen. The partitioning along the perpendicular sides of the determining triangles is shown by dashed lines.

Now take one connected part of the nonzero area. There the quadratic diagonal blocks must have nonzero left lower and right upper elements by the following argument: These elements belong to the nonzero area, hence both of them

should be either zero or nonzero by the zero angulus property. But,if they are zero then a zero angulus should pass through a vertex element of R that is a contradiction. As a consequence, Lemma 4.5 can be applied, where the divisors are the left lower and right upper elements of the nonzero diagonal blocks.

Formula (4.8) is a consequence of (4.3). The fraction $r(n_p,m_q)/r(n_p,m_p)$ in (4.3) is equal to 1 if $(i,j) \in N_q x N_q$ i. e. $p=q$ and (4.3) still remains valid by the triangle property. If $q < p$, this fraction may be rewritten in a product form: $R[n_{p-1},n_p;\ m_q,m_{p-1}] = 0$

$$\Rightarrow \quad r(n_{p-1},m_q)\,r(n_p,m_{p-1}) = r(n_{p-1},m_{p-1})\,r(n_p,m_q)$$

hence

$$\frac{r(n_p,m_q)}{r(n_p,m_p)} = \frac{r(n_{p-1},m_q)}{r(n_{p-1},m_{p-1})}\,\frac{r(n_p,m_{p-1})}{r(n_p,m_p)}$$

and if $p-1 < q$, this process can be continued to get the product in (4.4). A similar approach applies to (4.4) in order to obtain (4.9).

If some of the divisors are zero in (4.8) or (4.9) then some zero diagonal blocks can be found along the diagonal between the ith and jth position. This means that r_{ij} is in the zero area, consequently (4.1o) applies.

In this theorem the matrix elements are represented in terms of the boundary elements of the diagonal blocks and 2s-2 elements from the off-diagonal blocks.

5. SYMMETRIZATION

One expects, similarly to tridiagonal matrices that some matrices with the triangle property can be made pseudo-symmetric by multiplying by a diagonal matrix. In this section the composite Green's matrices, a generalization of Green's matrices, will be introduced and a symmetrization theorem will be stated.

DEFINITION 5.1. A pseudo-symmetric matrix that has the triangle property and the zero angulus property is called a composite Green's matrix.

THEOREM 5.2. Assume that matrix R has the triangle property and the zero angulus property. Then

$$R = D\,C\,, \tag{5.1}$$

where D is diagonal and C is a composite Green's matrix.

Proof. It can be seen in the proof of Theorem 4.6 that there is a symmetric partitioning of R, where the diagonal blocks are either entirely zero or they have nonzero left lower and right upper elements. Matrix R can be brought to a basic form by the same method as was given in the proof of Theorem 2.6. This method does not change the diagonal blocks, only their sequence along the diagonal is effected, hence it is enough to prove that a nonzero diagonal block can be symmetrized. This is stated in the next lemma.

LEMMA 5.3. Assume that the mxm matrix B has the triangle property and the zero angulus property so that b_{1m} and b_{m1} are vertex elements. Matrix B can then be symmetrized by multiplying by a diagonal matrix and the result is a Green's matrix.

Proof. As b_{1m} and b_{m1} are the vertex elements, matrix B is composed of a lower and an upper triangle by Theorem 3.3:

$$B = S\circ(Be_1e_m^TB/b_{m1}) + \tilde{S}\circ(Be_me_1^TB/b_{1m}). \qquad (5.2)$$

The idea of the proof is to multiply by a diagonal matrix $D = \langle d_i \rangle$ so that the first column will be equal to the first row:

$$d_1 = 1, \qquad d_ib_{i1} = d_1b_{1i}. \qquad (5.3)$$

The last column should also be equal to the last row:

$$d_ib_{im} = d_mb_{mi}. \qquad (5.4)$$

As the zeros are located at entirely zero anguli of B, the elements b_{i1} and b_{1i} in (5.3) should be zero or nonzero at the same time. Also, one has the same case in (5.4) for r_{in} and r_{ni}. The equations (5.3) and (5.4) are not contradictory if b_{i1}, b_{1i}, b_{im} and b_{mi} are nonzero because the triangle property yields the relation:

$$b_{ii} = b_{i1}b_{mi} / b_{m1} = b_{im}b_{1i} / b_{1m}.$$

Now all d_i's can be defined and matrix DB, where B comes from (5.2), can be reorganized so that the resulting matrix is a Green's matrix.

Similarly, one can construct a diagonal matrix D' so that BD' is symmetric.

There is also a parametric representation of composite Green's matrices.

THEOREM 5.4. A composite Green's matrix C can be given as follows: Let N_k, $k=1,2,\ldots,s$ be the same index sets as in Theorem 4.6. Then

$$c_{ij} = a_{\max(i,j)}\, b_{\min(i,j)} \prod_{k=q+1}^{p} \gamma_{\mathrm{sgn}(j-i),k}\,, \qquad (5.5)$$

where

$$(\min(i,j),\ \max(i,j)) \in N_q \times N_p\,,$$

a and b are n-dimensional vectors and the numbers $\gamma_{-1\,k}$, $\gamma_{1\,k}$, $k=2,3,\ldots,s$ satisfy $\gamma_{-1\,k}\gamma_{1\,k} = 0$.

Proof. The representation in Theorem 4.6 can be rewritten if the diagonal blocks are symmetric:

$$a_i = c(i,m_p) = c(m_p,i)\,, \qquad i \in N_p\,,$$

$$b_i = \begin{cases} c(n_p,i)/c(n_p,m_p) = c(i,n_p)/c(n_p,m_p)\,, & i \in N_p\,, \\ 0\,, \quad \text{if} \quad c(n_p,m_p) = 0\,, \end{cases}$$

$$\gamma_{-1\,k} = \begin{cases} c(n_k,m_{k-1})/c(n_k,m_k), \\ 0\,, \text{ if } \quad c(n_k,m_k) = 0\,, \end{cases}$$

$$\gamma_{1\,k} = \begin{cases} c(m_{k-1},n_k)/c(m_k,n_k), \\ 0\,, \text{ if } \quad c(m_k,n_k) = 0. \end{cases}$$

ACKOWLEDGEMENT. The author is indebted to ANNA LEE for helpful discussions and comments.

REFERENCES

1 E. Asplund, Inverses of matrices $\{a_{ij}\}$ which satisfy $a_{ij}=0$ for $j > i+p$, Math. Scand. 7:57-60 (1959).

2 W. Barrett, A theorem on inverses of tridiagonal matrices, Linear Algebra Appl., 27:211-217 (1979).

3 W. Barrett and P. Feinsilver, Inverses of banded matrices, Linear Algebra Appl., (to appear).

4 J. M. Ortega and W. C. Rheinboldt, Iterative Solution of Nonlinear Equations in Several Variables, Academic Press, New York, 1970, p. 47.

ВЛИЯНИЕ РАЗНОСТНОГО РЕЗОНАНСА ТРЕТЬЕГО ПОРЯДКА $2\nu_z - \nu_x = 1$ НА ДВИЖЕНИЕ ЧАСТИЦ В ЦИКЛИЧЕСКИХ УСКОРИТЕЛЯХ

И.В. АМИРХАНОВ, Е.П. ЖИДКОВ, И.Е. ЖИДКОВА

Объединенный институт ядерных исследований,
Лаборатория вычислительной техники и автоматизации

Основу теории бетатронных колебаний в циклических ускорителях составляет исследование системы нелинейных дифференциальных уравнений с периодическими коэффициентами. Основной задачей теории является исследование устойчивости движения. Наличие нелинейных членов в уравнении приводит к появлению "нелинейных" резонансов, которые будут наблюдаться при выполнении условия

$$k_x \nu_x + k_z \nu_z + q = 0 ,$$

где k_x, k_z, q - целые числа, ν_x и ν_z - частоты бетатронных колебаний.

Если посмотреть на диаграмму устойчивости любого конкретного ускорителя, то трудно выбрать рабочую точку вдали от резонансов, поскольку при достаточно больших k_x, k_z и q всегда найдется резонансная линия, проходящая вблизи этой точки.

В связи с этим становится актуальной задача исследования влияния различных нелинейных резонансов на движения частиц в циклических ускорителях. Этим вопросам посвящено большое количество работ [1-8], достаточно полный список литературы можно найти в [1,2].

В работах [7,8] был рассмотрен разностный резонанс третьего порядка $2\nu_z - \nu_x = 1$, который проходит достаточно близко от рабочей точки синхрофазотрона ОИЯИ. Исследование проводилось методом Крылова -Боголюбова [9] в первом приближении. Для повышения точности расчетов, т.е. для расширения интервала времени, на котором приближенное решение незначительно отличается от точного, нужно проводить исследования во втором и более высоких приближениях. Однако, насколько нам известно, до сих пор подобные исследования резонансов не проводились, что связано с практически непреодолимыми трудностями аналитических расчетов. Ситуация существенно изменилась с появлением возможности выполнять громоздкие аналитические выкладки на ЭВМ.

В данной работе исследуется разностный резонанс третьего порядка $2\nu_z - \nu_x = 1$ во втором приближении, причем все аналитические выкладки были сделаны с использованием системы для аналитических вычислений на ЭВМ *REDUCE-2* [10].

В отсутствие электрического поля движение заряженных частиц в циклических ускорителях сводится к исследованию системы нелинейных уравнений [1,2]

$$x'' + \nu_x^2 x = \sum_{i=1}^{\infty} \varepsilon^i V_{ix} ,$$
$$z'' + \nu_z^2 z = \sum_{i=1}^{\infty} \varepsilon^i V_{iz} , \qquad (I)$$

где V_{ix}, V_{iz} полиномы от x, x', z, z' и периодическая функция от θ, штрих означает дифференцирование по θ, $\varepsilon = \frac{1}{R_0}$ – малый параметр (R_0 – радиус идеальной орбиты).

Система уравнений (I) исследуется методом усреднения во втором приближении. Для этого в правой части (I) пренебрегаем членами порядка ε^3 и выше, после чего правые части системы (I) принимают вид $\varepsilon V_{1x} + \varepsilon^2 V_{2x}$ и $\varepsilon V_{1z} + \varepsilon^2 V_{2z}$, где

$$V_{1x} = Q_1(\theta) x^2 + Q_2(\theta) z^2 + \frac{1}{2}(x')^2 + \frac{1}{2}(z')^2 ,$$
$$V_{2x} = Q_4(\theta) x^3 + Q_5(\theta) x z^2 + \frac{1}{2}(3n_0 - 4) x (x')^2 +$$
$$+ \frac{n_0}{2} x (x')^2 + \frac{1}{2} n_0 x (z')^2 - n_0 z z' x' ,$$
$$V_{1z} = Q_3(\theta) x z + x' z' , \qquad (2)$$
$$V_{2z} = Q_6(\theta) x^2 z + Q_7(\theta) z^3 - \frac{1}{2} n_0 z (x')^2 -$$
$$- \frac{3}{2} n_0 z (z')^2 + (n_0 - 2) x x' z' ,$$
$$Q_j(\theta) = A_{j0} + B_{j1} \cos\theta + A_{j1} \sin\theta , \quad j = 1, 2, \ldots 7 .$$

В окрестности резонанса $2\nu_z - \nu_x = 1 + \delta$, $\delta \ll 1$, можно положить

$$\nu_z^2 = \left(\frac{1+\nu_x}{2}\right)^2 + \varepsilon \cdot \Delta , \qquad (3)$$

где $\varepsilon \cdot \Delta$ – представляет собой расстройку. После этого система (2) примет вид

$$x'' + \nu_x^2 x = \varepsilon V_{1x} + \varepsilon^2 V_{2x} , \qquad (4)$$
$$z'' + \left(\frac{1+\nu_x}{2}\right)^2 z = \varepsilon (V_{1z} - \Delta z) + \varepsilon^2 V_{2z} .$$

Сделаем в (4) замену переменных

$$x = u_1 \sin(\nu_x \theta + u_2), \qquad z = u_3 \sin\left[\left(\frac{1+\nu_x}{2}\right)\theta + u_4\right]$$

$$x' = \nu_x u_1 \cos(\nu_x \theta + u_2), \quad z' = \frac{1+\nu_x}{2} u_3 \cos\left[\left(\frac{1+\nu_x}{2}\right)\theta + u_4\right]. \qquad (5)$$

В новых переменных имеем

$$u_n'(\theta) = \varepsilon F_{1n}(\theta, u) + \varepsilon^2 F_{2n}(\theta, u), \qquad (6)$$

где

$$u = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix}, \qquad F_k = \begin{pmatrix} F_{k1} \\ F_{k2} \\ F_{k3} \\ F_{k4} \end{pmatrix}, \qquad k = 1,2.$$

$$F_{k1} = \frac{1}{\nu_x} V_{kx} \cdot \cos(\nu_x \theta + u_2),$$

$$F_{k2} = \frac{1}{\nu_x u_1} V_{kx} \cdot \sin(\nu_x \theta + u_2), \qquad (7)$$

$$F_{k3} = \frac{2}{(1+\nu_x)} V_{kz} \cdot \cos\left[\left(\frac{1+\nu_x}{2}\right)\theta + u_4\right],$$

$$F_{k4} = -\frac{2}{(1+\nu_x) u_3} V_{kz} \sin\left[\left(\frac{1+\nu_x}{2}\right)\theta + u_4\right].$$

Дифференциальные уравнения, приведенные к виду (6), называют уравнениями в стандартной форме. Усредним систему (6) по методу Крылова-Боголюбова /9/. Усреднение ведется по θ. В этом случае уравнения второго приближения примут вид:

$$u_n'(\theta) = \varepsilon \underset{\theta}{M} \left\{ F_{1n}(\theta, u) + \varepsilon F_{2n}(\theta, u) + \right.$$

$$\left. + \varepsilon \sum_{m=1}^{4} \tilde{F}_{1m}(\theta, u) \frac{\partial}{\partial u_m} F_{1n}(\theta, u) \right\}, \quad n = 1,2,3,4,$$

где

$$F = \sum_{p} e^{ip\theta} F_p(u), \quad \tilde{F} = \sum_{p \neq 0} \frac{e^{ip\theta}}{ip} F_p(u),$$

$\underset{\theta}{M}\{F(\theta, u)\} = F_0(u)$ - оператор усреднения.

Основные громоздкие аналитические выкладки при получении усредненных уравнений были выполнены с использованием системы *REDUCE-2* /10/

на ЭВМ ЕС-1060. Задача считалась по частям: каждое уравнение (8) усреднялось отдельно. При этом для вычисления каждого уравнения требовалось 1200Кб памяти и около 20 мин. машинного времени. После усреднения получается система из четырех уравнений, которая преобразуется к виду:

$$\dot{u}_1 = T_{10}^{(1)} u_3^2 u_1 + T_{10}^{(2)} \frac{u_3^4}{u_1} + T_{10}^{(3)} u_1^3 +$$
$$+ T_{11}^{(1)} u_3^2 \sin\psi + T_{12}^{(1)} \frac{u_3^4}{u_1} \sin 2\psi +$$
$$+ T_{13}^{(1)} u_3^2 \cos\psi + T_{14}^{(1)} \frac{u_3^4}{u_1} \cos 2\psi ,$$

$$\dot{u}_3 = T_{30}^{(1)} u_3^3 + T_{30}^{(2)} u_1^2 u_3 + T_{31}^{(1)} u_3 u_1 \sin\psi +$$
$$+ T_{32}^{(1)} u_3 u_1^2 \sin 2\psi + T_{33}^{(1)} u_3 u_1 \cos\psi + T_{34}^{(1)} u_3 u_1^2 \cos 2\psi ,$$

$$\dot{\psi} = \frac{2\varepsilon}{1+\nu_x} \Delta + \left(-T_{20}^{(1)} + 2T_{40}^{(3)}\right) u_1^2 + \left(-T_{20}^{(2)} + 2T_{40}^{(2)}\right) u_3^2 +$$
$$+ \left(-T_{21}^{(1)} \frac{u_3^2}{u_1} + 2T_{41}^{(1)} u_1\right) \sin\psi + \qquad (9)$$
$$+ \left(-T_{22}^{(1)} \frac{u_3^4}{u_1^2} + 2T_{42}^{(1)} u_1^2\right) \sin 2\psi +$$
$$+ \left(-T_{23}^{(1)} \frac{u_3^2}{u_1} + 2T_{34}^{(1)} u_1\right) \cos\psi +$$
$$+ \left(-T_{24}^{(1)} \frac{u_3^4}{u_1} + 2T_{44}^{(1)} u_1^2\right) \cos 2\psi ,$$

где $\psi = 2u_4 - u_2$,

$$T_{10}^{(1)} = \frac{\varepsilon^2}{128 \nu_x^2 (1+\nu_x)} \Big[-3\nu_x^2 (1+\nu_x)^3 - 2\nu_x (1+\nu_x)^2 A_{30} - 8\nu_x^2 (1+\nu_x) A_{20} -$$
$$- (1+\nu_x)^3 A_{10} - 4\nu_x (B_{31} B_{21} + A_{31} A_{21}) + 7(1+\nu_x)(B_{21} B_{11} + A_{21} A_{11}) +$$
$$+ 12 A_{20} A_{10} - 16 \nu_x A_{30} A_{20}) ,$$

$$T_{10}^{(2)} = \frac{\varepsilon^2}{128\,\nu_x^2}\,(3B_{21}^2 + 3A_{21}^2 + 8A_{20}^2)\,,$$

$$T_{10}^{(3)} = \frac{\varepsilon^2}{128\,\nu_x^2}\,(5B_{11}^2 + 5A_{11}^2 + 8A_{10}^2)\,,$$

$$T_{11}^{(1)} = \frac{\varepsilon^2\Delta}{4\nu_x(\nu_x+1)}\,A_{21} + \frac{\varepsilon}{8\,\nu_x}\,B_{21}\,,$$

$$T_{12}^{(1)} = \frac{\varepsilon^2}{64\,\nu_x^2}\,B_{21}A_{21}\,,\qquad T_{14}^{(1)} = \frac{\varepsilon^2}{128\,\nu_x^2}\,(B_{21}^2 - A_{21}^2)\,,$$

$$T_{13}^{(1)} = \frac{\varepsilon^2\Delta}{4\nu_x(\nu_x+1)}\,B_{21} - \frac{\varepsilon}{8\,\nu_x}\,A_{21}\,,$$

$$T_{20}^{(1)} = \frac{\varepsilon^2}{8\,\nu_x}\,(2\nu_x^2(1-n_0) - 3A_{40})\,,$$

$$T_{20}^{(2)} = \frac{\varepsilon^2}{128\,\nu_x^2(1+\nu_x)}\Big[8\nu_x(B_{31}A_{21} - B_{21}A_{31}) + 3\nu_x(B_{21}A_{11} - B_{11}A_{21}) + 3(B_{21}A_{11} - B_{11}A_{21})\Big]\,,$$

$$T_{21}^{(1)} = \frac{\varepsilon^2\Delta}{4\nu_x(1+\nu_x)}\,B_{21} - \frac{\varepsilon}{8\,\nu_x}\,A_{21}\,,$$

$$T_{23}^{(1)} = -\frac{\varepsilon^2\Delta}{4\nu_x(1+\nu_x)}\,A_{21} - \frac{\varepsilon}{8\,\nu_x}\,B_{21}\,,$$

$$T_{22}^{(1)} = \frac{\varepsilon^2}{64\,\nu_x^2}\,(B_{21}^2 - A_{21}^2)\,,\qquad T_{24}^{(1)} = \frac{\varepsilon^2}{32\,\nu_x^2}\,B_{21}A_{21}\,,$$

$$T_{30}^{(1)} = \frac{\varepsilon^2}{128\nu_x(1+\nu_x)}\Big[-(1+\nu_x)^3\nu_x - (1+\nu_x)^2A_{30} - 2B_{31}B_{21} - 2A_{31}A_{21}\Big]\,,$$

$$T_{30}^{(2)} = \frac{\varepsilon^2}{128(1+\nu_x)^3}\,(4\nu_x^2(1+\nu_x)^3 + 4\nu_x(1+\nu_x)^2A_{30} + (1+\nu_x)^2A_{10} + 4B_{31}^2 + 4A_{31}^2 + 4A_{30}^2)$$

$$T_{31}^{(1)} = -\frac{\varepsilon^2\Delta}{2(\nu_x+1)^2}\,A_{31} - \frac{\varepsilon}{4(\nu_x+1)}\,B_{31}\,,$$

$$T_{32}^{(1)} = \frac{\varepsilon^2}{16(1+\nu_x)^2}\,B_{31}A_{31}\,,\qquad T_{34}^{(1)} = \frac{\varepsilon^2}{32(1+\nu_x)^2}\,(B_{31}^2 - A_{31}^2)\,,$$

$$T_{33}^{(1)} = -\frac{\varepsilon^2\Delta}{2(1+\nu_x)^2}\,B_{31} + \frac{\varepsilon}{4(1+\nu_x)}\,A_{31}\,,$$

$$T_{40}^{(2)} = \frac{\varepsilon^2}{32\nu_x(1+\nu_x)}\left[3n_0\nu_x(1+\nu_x)^2 - 24\nu_x A_{70} - 2B_{31}A_{21}\right],$$

$$T_{40}^{(3)} = \frac{\varepsilon^2}{32\nu_x(1+\nu_x)}\left[8n_0\nu_x^3 - 16\nu_x A_{60} - 2B_{31}A_{11} + B_{11}A_{31}\right],$$

$$T_{41}^{(1)} = \frac{\varepsilon^2\Delta}{2(1+\nu_x)^2}B_{31} - \frac{\varepsilon}{4(1+\nu_x)}A_{31},$$

$$T_{42}^{(1)} = \frac{\varepsilon^2}{16(1+\nu_x)^2}\left(-B_{31}^2 + A_{31}^2\right),$$

$$T_{43}^{(1)} = -\frac{\varepsilon^2\Delta}{2(1+\nu_x)^2}A_{31} - \frac{\varepsilon}{4(1+\nu_x)}B_{31},$$

$$T_{44}^{(1)} = \frac{\varepsilon^2}{8(1+\nu_x)^2}B_{31}A_{31}.$$

Если в системе (9) пренебречь членами порядка ε^2, то получаются усредненные уравнения в первом приближении. Эти уравнения полностью совпали с уравнениями первого приближения, полученными вручную в работе [7], что является косвенным подтверждением правильности уравнений (9), полученных с помощью системы REDUCE-2.

Переходим к исследованию системы (9). Из первых двух уравнений системы (9) получаем равенство

$$\frac{d}{d\theta}\left[4\nu_x U_1^2 + (1+\nu_x)U_3^2\right] = \varepsilon^2 M, \qquad (10)$$

где M – полином от $U_1, U_3, \sin\Psi, \cos\Psi, \sin 2\Psi, \cos 2\Psi$.

В равенстве (10) правая часть пропорциональная ε^2, что свидетельствует о медленном изменении выражения $[4\nu_x U_1^2 + (1+\nu_x)U_3^2]$. Для значений аргумента θ на интервале $0 \le \theta \le \frac{1}{\varepsilon^2}$ это выражение практически остается постоянным. Тогда из (10) имеем

$$4\nu_x U_1^2 + (1+\nu_x)U_3^2 = H_0, \qquad (11)$$

где H_0 – определяется из начальных условий и практически остается постоянным на интервале $0 \le \theta \le \frac{1}{\varepsilon^2}$.

Из (11) следует, что амплитуды обоих видов колебаний U_1 и U_3 остаются в процессе движения ограниченными и резонанс $2\nu_z - \nu_x = 1$ не приводит к неустойчивости на интервале $0 \le \theta \le \frac{1}{\varepsilon^2}$. Так как в (11)

входит явно расстройка Δ_{ν}, то амплитуды остаются конечными даже точно в резонансе. В работе[7] были получены аналогичные результаты, когда система (I) усредняется в первом приближении в окрестности резонанса $2\nu_z - \nu_x = 1$. Результаты настоящей работы позволяют существенно расширить интервал изменения аргумента θ, на котором оказались справедливыми те же утверждения, что и в работе[7].

Литература

I. Коломенский А.А., Лебедев А.Н. Теория циклических ускорителей. Физматгиз, М., I962.

2. Брук Г. Циклические ускорители заряженных частиц. Атомиздат, М., I970.

3. Schoch A. CERN-Report, 57-21, Geneve, 1958.

4. Hagedorn R., Schoch A. CERN-Report, 57-14, Geneve, 1957.

5. Василишин Б.В. и др. ОИЯИ, 9-7498, Дубна, I973.

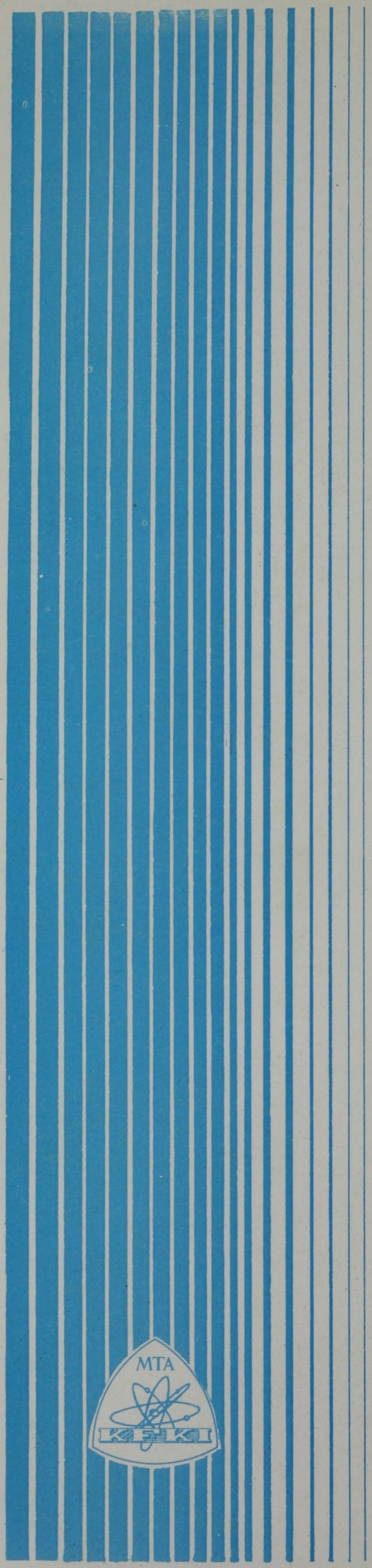
6. Безногих Ю.Д. и др. ОИЯИ, Р9-9II5, Дубна, I975, Р9-9I20, Дубна, I975.

7. Амирханов И.В. и др. ОИЯИ, 9-8663, Дубна, I975.

8. Амирханов И.В. и др. ОИЯИ, РII-9I07, Дубна, I975, РII-8780, Дубна, I975; РII-9I08, Дубна, I975.

9. Боголюбов Н.Н., Митропольский Ю.А. Асимптотические методы в теории нелинейных колебаний. "Наука", М., I974.

I0. Hearn A.C., REDUCE-2 User's Manual, UCP-19, University of Utah, Salt Lake City, 1973.



Kiadja a Központi Fizikai Kutató Intézet
Felelős kiadó: Lőcs Gyula
Szakmai lektor: Németh Géza
Nyelvi lektor: Harvey Shenker
Példányszám: 305 Törzsszám: 83-110
Készült a KFKI sokszorositó üzemében
Felelős vezető: Nagy Károly
Budapest, 1983. március hó