ПРИОРИТЕТНЫЙ НАЦИОНАЛЬНЫЙ ПРОЕКТ «ОБРАЗОВАНИЕ»
РОССИЙСКИЙ УНИВЕРСИТЕТ ДРУЖБЫ НАРОДОВ

К.П. ЛОВЕЦКИЙ, Л.А. СЕВАСТЬЯНОВ

МЕТОДЫ ДИФФЕРЕНЦИАЛЬНЫХ РАЗНОСТЕЙ РАСЧЕТА ОПТИЧЕСКИХ ПОКРЫТИЙ

Учебное пособие

Москва

2008

Инновационная образовательная программа
Российского университета дружбы народов

«Создание комплекса инновационных образовательных программ и формирование инновационной образовательной среды, позволяющих эффективно реализовывать государственные интересы РФ через систему экспорта образовательных услуг»

Экспертное заключение –

доктор физико-математических наук, профессор *С.И. Виницкий*

Ловецкий К.П., Севастьянов Л.А.

Методы дифференциальных разностей расчета оптических покрытий: Учеб. пособие. – М.: РУДН, 2008. – 161 с.

Пособие посвящено изложению основ метода дифференциальных разностей и его применению к численному решению систем обыкновенных дифференциальных уравнений для моделирования и проектирования современных оптических устройств на основе тонкопленочных покрытий и дифракционных оптических элементов.

Методы дифференциальных разностей включают в себя различные модификации методов Рунге–Кутта, учитывающие гамильтонову структуру уравнений Максвелла. Данные методы являются предпочтительными в том случае, когда периодическая структура существенно не бинарна. По существу эти методы являются дискретной реализацией метода геометрического интегрирования Уитни.

Для магистров и аспирантов, обучающихся по направлению «Прикладная математика и информатика».

Учебное пособие выполнено в рамках инновационной образовательной программы Российского университета дружбы народов, направление «Комплекс экспортоориентированных инновационных образовательных программ по приоритетным направлениям науки и технологий», и входит в состав учебно-методического комплекса, включающего описание курса, программу и электронный учебник.

© Ловецкий К.П., Севастьянов Л.А., 2008

Содержание

Общее описание курса

Курс «Методы дифференциальных разностей расчета оптических покрытий» является составной частью магистерской программы «Оптика наноструктур». Магистерская программа «Оптика наноструктур» реализуется в рамках направления «Прикладная математика и информатика» и направления «Прикладная математика и физика», а возможно, и других направлений. В составе магистерской программы «Оптика наноструктур» курс «Методы дифференциальных разностей расчета оптических покрытий» является обязательным, не привязанным к семестру. Для других магистерских программ этот курс может быть курсом по выбору без привязки к семестру или факультативным на усмотрение методической комиссии программы. Курс носит теоретический и практический характер.

Целью курса является подробное ознакомление студентов с устойчивыми современными методами численного решения систем обыкновенных дифференциальных уравнений, получающихся при применении модального Фурье-метода к математической модели взаимодействия электромагнитного излучения в области светового диапазона с диэлектрическими структурами нанометровых размеров. Эта область знаний особенно быстро развивается в последние годы в связи с широким применением наноэлементов и тонких (менее одного микрометра толщиной) пленок, используемых в производстве жидкокристаллических дисплеев, солнечных батарей на основе диэлектриков, фотоэмиссионных диодов, просветляющих покрытий, поляризаторов, миниатюрных лазеров, управляемых оптических элементов. Задачи оптики наноструктур практически не поддаются аналитическому решению, поэтому важным является не только освоение теоретического материала, но и изучение специальных численных методов, используемых при решении данного класса задач, приобретение навыков создания программного обеспечения

для численного моделирования различных оптических наноструктур.

Задачей курса «Методы дифференциальных разностей расчета оптических покрытий» является обучение студентов использованию специальных методов численного решения систем обыкновенных дифференциальных уравнений для моделирования и проектирования современных оптических устройств на основе тонкопленочных покрытий и дифракционных оптических элементов. Это позволит им при необходимости разрабатывать новое программное обеспечение. Безусловной задачей курса является также освоение существующего программного обеспечения, ориентированного на расчет и проектирование оптических покрытий. В результате обучения студенты получат умение и навыки правильно оценить сложность научно-исследовательских и конструкторских заданий на разработку дифракционных оптических элементов и устройств, аргументированно выбирать метод решения конструкторской задачи, а затем экономично и эффективно выполнять компьютерный дизайн требуемого дифракционного оптического покрытия или устройства.

Инновационность курса

Курс является инновационным по содержанию и по литературе, он включает в себя последние научные достижения в области решения задач дифракционной оптики, когда характерные размеры исследуемых объектов не превышают либо сравнимы с длиной волны оптического излучения. Эта область знаний интенсивно развивалась в последнее время, но лишь недавно были созданы устойчивые алгоритмы и разработаны численные методы решения задач для многослойных решеток. Следует отметить, что для оптических однослойных и многослойных решеток с характерными размерами больше длины волны оптического излучения устойчивые методы решения известны с середины прошлого века. Сейчас

алгоритмы решения оптических задач в субволновой области распространяются на объекты со сложной геометрией, такие как двумерные решетки с произвольным профилем, трехмерные решетки (фотонные кристаллы), и на анизотропные материалы. Они востребованы, поскольку позволяют создавать математические модели взаимодействия излучения с веществом в наномасштабах, а затем с их помощью проектировать новые эффективные устройства в высокотехнологичных областях медицины, энергетики, инфокоммуникаций и приборостроения.

В ходе проведения занятий по этому курсу разработчики предполагают использование традиционных методик преподавания, принятых в странах болонской системы образования, т. е. с использованием кредитной системы оценки знаний.

Наряду с традиционными элементами преподавания математических методов решения прикладных задач, разработчики курса предполагают воспользоваться хорошо зарекомендовавшим себя опытом МФТИ и подобных вузов. Для этого в рамках подпрограммы «Оптика наноструктур» осуществляется закупка уникального измерительного и аналитического оборудования для выполнения измерений разнообразных характеристик оптических наноустройств с целью использования этого оборудования в учебном процессе и для проведения научно-исследовательских работ преподавателями, аспирантами и студентами.

По окончании магистратуры по направлению «Оптика наноструктур» выпускники Российского университета дружбы народов станут конкурентоспособными специалистами в области проектирования современных оптических устройств, которые не будут испытывать затруднений при последующем трудоустройстве.

Данное направление научно-практических разработок сформировалось лишь в последние 10 – 15 лет. Поэтому наблюдается

сильный дефицит учебно-методической литературы не только в России, но и во всем мире. Разрабатываемые в рамках инновационной программы «Оптика наноструктур» учебные пособия восполнят в некоторой степени этот пробел и составят основной список литературы для слушателей курсов. Вместе с ними следует использовать несколько учебников и монографий, вышедших в свет к настоящему времени и перечисленных в списке литературы. Курс базируется на публикациях научных статей мировых лидеров исследований в данной области в научной периодике, диссертационных работах их учеников, включающих работы по непосредственному моделированию, дизайну и последующему изготовлению лабораторных образцов оптических элементов и устройств.

В качестве практических заданий, курсовых работ и тем рефератов слушателям магистерской программы будут предложены актуальные проблемы и задачи, решение которых востребовано современным уровнем развития высокотехнологичных отраслей промышленности и научно-исследовательских лабораторий.

Тема 1. МАТРИЦЫ И ДИФФЕРЕНЦИАЛЬНЫЕ УРАВНЕНИЯ

1.1. Системы линейных обыкновенных дифференциальных уравнений первого порядка

Обсудим применение теории матриц к решению линейных систем дифференциальных уравнений вида:

$$\frac{dx_i}{dt} = \sum_{l=1}^{N} a_{ij} x_j, \quad x_i(0) = c_i, \quad i = 1,2,...,N, \tag{1.1}$$

где a_{ij} — матрица коэффициентов.

В изложении материала данного раздела в основном использованы результаты работы Беллмана [1].

Векторно-матричные обозначения

Для исследования системы (1.1) введем векторы **y** и **c** с компонентами соответственно y_i и c_i и матрицу $A = \left\| a_{ij} \right\|$. Из определения разности двух векторов очевидно, что производная от вектора должна быть определена как

$$\frac{d\mathbf{y}}{dt} = \left\| \begin{array}{c} \frac{dy_1}{dt} \\ \frac{dy_2}{dt} \\ \cdot \\ \cdot \\ \cdot \\ \frac{dy_N}{dt} \end{array} \right\|. \tag{1.2}$$

Подобным образом интеграл от $\mathbf{y}(t)$ определяется как

$$\int_0^t \mathbf{y}(s)ds = \begin{Vmatrix} \int_o^t y_1(s)ds \\ \int_o^t y_2(s)ds \\ . \\ . \\ . \\ \int_o^t y_N(s)ds \end{Vmatrix}. \qquad (1.3)$$

Аналогично вводятся производные и интегралы от матриц. Отсюда следует, что уравнения (1.1) могут быть записаны как

$$\frac{d\mathbf{x}}{dt} = A\mathbf{x}, \quad \mathbf{x}(0) = \mathbf{c}. \qquad (1.4)$$

Матрица A в общем случае не симметрична.

Вектор, составляющие которого являются функциями t, будет называться векторной функцией, или, функцией от t. Векторная функция непрерывна, если все ее составляющие в рассматриваемом интервале являются непрерывными функциями аргумента t. Эту же терминологию будем использовать при описании матричных функций.

Равномерные нормы векторов и матриц

При желании можно использовать скалярную функцию (x,x) как меру вектора $\mathbf{x}$, однако удобнее использовать не эту евклидову норму, а более простое выражение:

$$\|\mathbf{x}\| = \sum_{i=1}^{N} |x_i| \qquad (1.5)$$

для вектора и

$$\|A\| = \sum_{i,j=1}^{N} |a_{ij}| \tag{1.6}$$

для матрицы. Легко проверяется, что

$$\begin{gathered} \|\mathbf{x}+\mathbf{y}\| \le \|\mathbf{x}\| + \|\mathbf{y}\|, \quad \|A+B\| \le \|A\| + \|B\|, \\ \|A\mathbf{x}\| \le \|A\|\cdot\|\mathbf{x}\|, \quad \|AB\| \le \|A\|\cdot\|B\|, \\ \|c_1\mathbf{x}\| \le |c_1|\cdot\|\mathbf{x}\|, \quad \|c_1 A\| \le |c_1|\cdot\|A\|. \end{gathered} \tag{1.7}$$

Причиной, по которой мы выбрали предыдущие выражения в качестве норм для вектора и матрицы, является то, что проверка всех результатов (1.7) чрезвычайно проста. Все нормы в равной степени применимы при рассмотрении конечномерных векторов и матриц. Выбор нормы становится делом более сложным лишь в случае, когда мы обращаемся к бесконечномерным векторам и матрицам.

Бесконечные ряды векторов и матриц

В ходе доказательства существования решений линейного векторного уравнения, введенного выше, нам понадобятся бесконечные ряды векторов и матриц. Под вектором $\sum_{n=0}^{\infty} \mathbf{x}^n$ будем понимать вектор, i-я составляющая которого есть сумма ряда $\sum_{n=0}^{\infty} x_i^n$. Поэтому сходимость векторного ряда эквивалентна одномерной сходимости N рядов $\sum_{n=0}^{\infty} x_i^n$, $i = 1,\ldots,N$. Отсюда следует, что достаточным условием сходимости векторного ряда $\sum_{n=0}^{\infty} \mathbf{x}^n$ является сходимость скалярного ряда $\sum_{n=0}^{\infty} \|\mathbf{x}^n\|$.

Подобно этому матричный ряд вида $\sum_{n=0}^{\infty} A_n$ представляет собой N^2

рядов, а для сходимости матричного ряда достаточно, чтобы сходился ряд $\sum_{n=0}^{\infty}\|A_n\|$.

Существование и единственность решений линейной системы уравнений.

Благодаря этим предварительным замечаниям можно доказать следующий основной результат.

Теорема 1. Если матрица $A(t)$ непрерывна при $t \geq 0$, то решение векторного дифференциального уравнения

$$\frac{d\mathbf{x}}{dt} = A(t)\mathbf{x}, \quad \mathbf{x}(0) = \mathbf{c}, \tag{1.8}$$

существует для всех $t \geq 0$, является единственным и может быть записано в виде

$$\mathbf{x} = X(t)\mathbf{c}, \tag{1.9}$$

где $X(t)$ — матрица, определенная единственным образом и удовлетворяющая матричному дифференциальному уравнению

$$\frac{dX}{dt} = A(t)X, \quad X(0) = I. \tag{1.10}$$

Доказательство. Для установления существования решения (1.10) используем метод последовательных приближений. Рассмотрим вместо (1.10) эквивалентное интегральное уравнение:

$$X = I + \int_0^t A(s)Xds. \tag{1.11}$$

Определим последовательность матриц $\{X_n\}$ следующим образом:

$$X_0 = I,$$
$$X_{n+1} = I + \int_0^t A(s)X_n ds, \quad i = 1,2... \quad . \tag{1.12}$$

Далее имеем

$$X_{n+1} - X_n = \int_0^t A(s)(X_n - X_{n-1})ds, \quad i = 1,2... \ . \tag{1.13}$$

Положим

$$m = \max_{0 \le t \le t_1} \|A(t)\|. \tag{1.14}$$

Здесь и далее используем определение нормы (1.5) и (1.6).

Используя (1.13), получим при $0 \le t \le t_1$

$$\begin{aligned} \|X_{n+1} - X_n\| &= \left\| \int_0^t A(s)(X_n - X_{n-1})ds \right\| \le \\ &\le \int_0^t \|A(s)\| \|X_n - X_{n-1}\| ds \le m \int_0^t \|X_n - X_{n-1}\| ds. \end{aligned} \tag{1.15}$$

Поскольку в этом же интервале

$$\|X_1 - X_0\| \le \int_0^t \|A(s)\| ds \le mt, \tag{1.16}$$

то, используя (1.15), по индукции получим

$$\|X_{n+1} - X_n\| \le \frac{m^{n+1} t^{n+1}}{(n+1)!} \tag{1.17}$$

при $0 \le t \le t_1$.

Следовательно, ряд $\sum_{n=0}^{\infty} (X_{n+1} - X_n)$ сходится равномерно в интервале $[0, t_1]$. Поэтому матрица X_n равномерно сходится к матрице $X(t)$, которая удовлетворяет уравнению (1.11), а, следовательно, и (1.10).

В предположении, что $A(t)$ непрерывна при $t \geq 0$, мы можем выбрать t_1 сколь угодно большим, и, таким образом, получим решение, существующее при всех $t \geq 0$.

Легко проверяется, что $\mathbf{x} = X(t)\mathbf{c}$ есть решение уравнения (1.8), удовлетворяющее требуемому начальному условию. Установим далее единственность решения уравнения (1.10). Пусть имеется другое решение, которое обозначим через Y. Тогда Y удовлетворяет (1.11), и, следовательно, мы имеем соотношение:

$$X - Y = \int_0^t A(s)[X(s) - Y(s)]ds. \tag{1.18}$$

Поэтому

$$\|X - Y\| \leq \int_0^t \|A(s)\| \|X(s) - Y(s)\| ds. \tag{1.19}$$

Так как матрицы Y и X дифференцируемы, а следовательно, непрерывны, то существует максимум:

$$m = \max_{0 \leq t \leq t_1} \|X - Y\|. \tag{1.20}$$

Из (1.19) получим

$$\|X - Y\| \leq m_1 \int_0^t \|A(s)\| ds \quad 0 \leq t \leq t_1. \tag{1.21}$$

Используя неравенство (1.19), получим

$$\|X - Y\| \leq m_1 \int_0^t \|A(s)\| \left(\int_0^s \|A(s_1)\| ds_1 \right) ds \leq \frac{m_1 \int_0^t \|A(s)\| ds}{2} \quad . \tag{1.22}$$

Продолжая итеративную процедуру, получим

$$\|X - Y\| \le \frac{m_1 \left(\int_0^t \|A(s)\| ds \right)^{n+1}}{(n+1)!} \quad . \tag{1.23}$$

Устремляя n к бесконечности, видим, что $\|X - Y\| \le 0$. Поэтому $X \equiv Y$.

Имея матрицу X, легко построить решение уравнения (1.8). Оно равно $X(t)\mathbf{c}$. Единственность решения уравнения (1.8) легко установить, используя те же рассуждения, что и выше.

Матричная экспонента

Рассмотрим теперь частный случай, когда $A(t)$ — постоянная матрица. В скалярном случае уравнение

$$\frac{du}{dt} = au, \quad u(0) = c, \tag{1.24}$$

имеет решение $u = e^{at}c$. Было бы очень удобно найти аналогичное решение для матричного уравнения:

$$\frac{dX}{dt} = AX, \quad X(0) = C, \tag{1.25}$$

имеющее форму $X = e^{At}C$.

По аналогии со скалярным случаем и имея в виду метод последовательных приближений, попытаемся определить матричную экспоненциальную функцию посредством ряда:

$$e^{At} = I + At + \ldots + \frac{A^n t^n}{n!} + \ldots \quad . \tag{1.26}$$

Докажем следующий результат.

Теорема 2. Матричный ряд, определенный выше, существует для всех матриц A при любом фиксированном t, и для фиксированной

матрицы A он равномерно сходится в любой конечной области комплексной плоскости t.

Доказательство. Имеем

$$\frac{\left\|A^n t^n\right\|}{n!} \le \frac{\left\|A^n\right\|\left|t^n\right|}{n!}. \tag{1.27}$$

Учитывая, что $\left\|A^n\right\|\left|t^n\right| / n!$ является общим членом разложения в ряд экспоненты $e^{\|A\|\,|t|}$, видим, что ряд (1.26) мажорируется равномерно сходящимся рядом и, следовательно, сам равномерно сходится в любой конечной области плоскости t.

Функциональные уравнения

Скалярная экспоненциальная функция удовлетворяет основному функциональному тождеству:

$$e^{a(s+t)} = e^{as} e^{at}. \tag{1.28}$$

До тех пор, пока аналогичное равенство не доказано для матричной экспоненты, мы не имеем права использовать обозначение (1.26). Покажем далее, что

$$e^{A(s+t)} = e^{As} e^{At}. \tag{1.29}$$

Используя разложение в ряд для трех экспоненциальных функций и тот факт, что члены абсолютно сходящегося ряда можно группировать произвольным образом, можно записать:

$$\begin{aligned} e^{As} e^{At} = \sum_{k=0}^{\infty} \frac{A^k s^k}{k!} \left(\sum_{l=0}^{\infty} \frac{A^l t^l}{l!} \right) = \sum_{n=0}^{\infty} A^n \left(\sum_{k+l=n} \frac{s^k t^l}{k!\,l!} \right) = \\ = \sum_{n=0}^{\infty} A^n \frac{(s+t)^n}{n!} = e^{A(s+t)}. \end{aligned} \tag{1.30}$$

Полагая в (1.29) $s = -t$, получим важный результат:

$$e^{A(-t+t)} = e^{-At}e^{At}. \tag{1.31}$$

Следовательно, матрица e^{At} всегда невырождена и ее обратная равна e^{-At}. Это матричный аналог того факта, что скалярная экспонента никогда не обращается в нуль.

Доказательство функционального равенства (1.30) является скорее проверкой результата, чем выводом. Для понимания результата обратимся к дифференциальному уравнению:

$$\frac{dX}{dt} = AX. \tag{1.32}$$

Заметим, что e^{At} есть решение этого уравнения при начальном условии $X(0) = I$, а $e^{A(s+t)}$ — решение при $X(0) = e^{As}$. Поэтому из теоремы единственности можно заключить, что $e^{A(s+t)} = e^{As}e^{At}$.

После вывода функционального равенства, обсуждавшегося выше, возникает естественный вопрос о связи между $e^{(A+B)t}$ и $e^{At}e^{Bt}$. Поскольку

$$e^{(A+B)t} = I + (A+B)t + \frac{(A+B)^2}{2}t^2 + \dots, \tag{1.33}$$

а

$$e^{At}e^{Bt} = \left(I + At + \frac{A^2}{2}t^2 + \dots\right)\left(I + Bt + \frac{B^2}{2}t^2 + \dots\right),$$

то

$$e^{(A+B)t} - e^{At}e^{Bt} = (BA - AB)\frac{t^2}{2} + \dots. \tag{1.34}$$

Следовательно, равенство $e^{(A+B)t} = e^{At}e^{Bt}$ справедливо для всех t только в случае, когда $AB = BA$, т. е. когда матрицы A и B перестановочны. Легко видеть, что это условие является и достаточным.

1.2. Однородные линейные системы дифференциальных уравнений с постоянными коэффициентами

Рассмотрим частный случай, в котором задача Коши формулируется для однородной системы дифференциальных уравнений с постоянной $n \times n$ - матрицей коэффициентов:

$$\frac{dy}{dt} = \mathbf{A}y\,, \quad y = y^0 \quad \text{при} \quad t = t_0\,.$$

Ранее было показано, что задача Коши имеет единственное непрерывно дифференцируемое решение $y = y(t)$. Поскольку **A** – постоянная матрица, то решение определено на любом конечном отрезке по t. Нам предстоит убедиться, что при этом $y(t)$ принадлежит классу сколь угодно число раз дифференцируемых функций и что $y(t)$ можно представить в виде бесконечного равномерно сходящегося ряда.

Принимая во внимание, что правые части системы не зависят от t, можно, не теряя общности, считать, что вектор начальных данных задается при $t = 0$. Действительно, после замены аргумента $t = t_0 + \tau$ задача Коши принимает вид:

$$\frac{dy}{dt} = \mathbf{A}y\,, \quad y = y^0 \quad \text{при} \quad \tau = \tau_0\,.$$

В дальнейшем будем считать, что эта замена уже проведена, полагая $t_0 = 0$.

Предположим, что вектор-функция $\mathbf{y}(t)$ имеет производные любого порядка. Поскольку функция $\mathbf{y}(t)$ – решение задачи Коши, то

$$\frac{dy}{dt}(t) = Ay(t),\ \frac{d^2y}{dt^2}(t) = A\frac{dy}{dt}(t) = A^2y(t),\ldots..,\ \frac{d^ky}{dt^k}(t) = A^ky(t),\ldots\,.$$

Используя эти выражения, запишем формальное представление $\mathbf{y}(t)$

в виде ряда Тейлора в окрестности $t = t_0 = 0$. Имеем:

$$y(t) = y(0) + \frac{t}{1!}\frac{dy}{dt}(0) + \frac{t^2}{2!}\frac{d^2y}{dt^2}(0) + \ldots + \frac{t^k}{k!}\frac{d^ky}{dt^k}(0) + \ldots =$$

$$= y(0) + \frac{t}{1!}Ay(0) + \frac{t^2}{2!}A^2y(0) + \ldots\frac{t^k}{k!}A^ky(0) + \ldots \ldots$$

Теорема. Вектор-функцию $\mathbf{y}(t)$, являющуюся решением задачи Коши

$$\frac{dy}{dt} = Ay\,, \quad y = y^0 \quad \text{при } t = 0, \tag{1.35}$$

где **A** – постоянная матрица, можно представить в виде функционального ряда:

$$\mathbf{y}(t) = y^0 + \frac{t}{1!}Ay^0 + \frac{t^2}{2!}A^2y^0 + \ldots + \frac{t^k}{k!}A^ky^0 + \ldots, \tag{1.36}$$

равномерно сходящегося на любом конечном отрезке $|t| \le T < \infty$. При этом $\mathbf{y}(t)$ имеет непрерывные производные любого порядка.

Доказательство. Равномерная сходимость ряда (1.36), члены которого – непрерывные вектор-функции, следует из оценки:

$$\left\| \frac{t^k}{k!}A^ky^0 \right\| \le \frac{|t|^k \|A\|^k}{k!}\|y^0\| \le \frac{(T\|A\|)^k}{k!}\|y^0\|\,.$$

Очевидно, числовой ряд

$$\|y^0\| + \frac{T\|A\|}{1!}\|y^0\| + \frac{(T\|A\|)^2}{2!}\|y^0\| + \ldots + \frac{(T\|A\|)^k}{k!}\|y^0\| + \ldots,$$

сходящийся к $e^{T\|A\|}\|y^0\|$, является мажорирующим, что и доказывает, согласно признаку Вейерштрасса, равномерную сходимость (1.36) к непрерывной вектор-функции $\mathbf{y}(t)$.

Рассмотрим теперь ряд, формально составленный из производных членов ряда (1.36):

$$\frac{dy}{dt}(t) = Ay^0 + \frac{t}{1!}A^2y^0 + \ldots\frac{t^{k-1}}{(k-1)!}A^k y^0 + \ldots \quad . \tag{1.37}$$

Легко заметить, что в этом случае мажорирующий числовой ряд имеет вид:

$$\|A\|\,\|y^0\| + \frac{T\|A\|^2}{1!}\|y^0\| + \ldots + \frac{T^{k-1}\|A\|^k}{(k-1)!}\|y^0\| + \ldots$$

(Его сумма равна $\|A\|e^{T\|A\|}\|y^0\|$.) Поэтому ряд (1.37) равномерно сходится к непрерывной вектор-функции. Следовательно, ряд (1.36) можно почленно дифференцировать, а его суммой является производная вектор-функции $\mathbf{y}(t)$.

Точно также можно убедиться, что в свою очередь и ряд (1.37) можно почленно дифференцировать, и т.д. В результате приходим к выводу, что ряд (2) почленно дифференцируем сколь угодно число раз.

Покажем, что вектор-функция $\mathbf{y}(t)$, представленная рядом (1.36), дает решение задачи Коши (1). Действительно,

$$Ay(t) = A(\ y^0 + \frac{t}{1!}Ay^0 + \frac{t^2}{2!}A^2y^0 + \ldots\ldots\frac{t^k}{k!}A^k y^0 + \ldots\ldots\ ) =$$

$$= Ay^0 + \frac{t}{1!}A^2y^0 + \ldots\ldots\frac{t^{k-1}}{(k-1)!}A^k y^0 + \ldots\ldots \equiv \frac{dy}{dt}(t)$$

Кроме того, $\mathbf{y}(0) = y^0$. Теорема доказана.

Матричная экспонента

Представим решение задачи Коши (1.35) в виде:

$$\mathbf{y}(t)=\left[E+\frac{t}{1!}A+\frac{t^2}{2!}A^2+\ldots+\frac{t^k}{k!}A^k+\ldots\right]y^0,$$

где E – единичная матрица. С использованием, как и ранее, признака Вейерштрасса показывается, что бесконечный ряд для матричных функций

$$\mathbf{Y}(t)=E+\frac{t}{1!}A+\frac{t^2}{2!}A^2+\ldots+\frac{t^k}{k!}A^k+\ldots \qquad (1.38)$$

равномерно сходится к непрерывной матричной функции $\mathbf{Y}(t)$ на любом конечном отрезке по t. При этом формально составленный ряд для производных членов ряда (1.38)

$$\frac{dY}{dt}(t)=\ A+\frac{t}{1!}A^2+\ \ldots+\frac{t^{k-1}}{(k-1)!}A^k+\ldots. \qquad (1.39)$$

равномерно сходится, и, следовательно, ряд (1.38) допускает почленное дифференцирование. В свою очередь, равномерно сходится ряд, составленный для производных членов ряда (1.39) и т.д. Таким образом, матричная функция $\mathbf{Y}(t)$ непрерывно дифференцируема сколь угодно число раз. Далее легко проверить, что матричная функция $\mathbf{Y}(t)$ удовлетворяет условиям:

$$\frac{dY}{dt}(t)\equiv AY(t),\quad Y(0)=E.$$

Таким образом, $\mathbf{Y}(t)$ является решением задачи Коши для матричного однородного дифференциального уравнения:

$$\frac{dY}{dt}=AY,\ Y=E\ \text{ при }\ \mathrm{t}=0. \qquad (1.40)$$

Определение. Матричная функция $\mathbf{Y}(t)$, определяемая либо равномерно сходящимся рядом (1.38), либо как решение задачи Коши (1.40), называется матричной экспонентой матрицы A и обозначается e^{tA}.

Если матричная экспонента известна, то решение задачи Коши (1.35) записывается в виде:

$$y(t) = e^{tA} y^0.$$

Свойства матричной экспоненты

1. Матрица A и матричная экспонента матрицы A перестановочны: $A\,e^{tA} = e^{tA} A.$ Это свойство становится очевидным, если воспользоваться представлением матричной экспоненты в виде бесконечного ряда:

$$e^{tA} = \mathrm{E} + \frac{t}{1!}A + \frac{t^2}{2!}A^2 + \ldots + \frac{t^k}{k!}A^k + \ldots \qquad (1.41)$$

При этом имеем

$$Ae^{tA} = \mathrm{A} + \frac{t}{1!}A^2 + \frac{t^2}{2!}A^3 + \ldots + \frac{t^k}{k!}A^{k+1} + \ldots = e^{tA} A\ .$$

2. Точно так же показывается, что если A и B – перестановочные матрицы, то перестановочными будут матрицы A и матричная экспонента матрицы B, или матрица B и матричная экспонента матрицы A:

$Ae^{tA} = e^{tA} A$, $Be^{tA} = e^{tA} B$, если $AB = BA$.

3. Пусть $AB = BA$. Составим матричную функцию $V(t) = e^{tA}\, e^{tB}$. Имеем:

$$\frac{dV}{dt}(t) = Ae^{tA}e^{tB} + e^{tA}Be^{tB}.$$

Поскольку матрицы B и e^{tA} перестановочны, то

$$\frac{dV}{dt}(t) = \quad Ae^{tA}e^{tB} + Be^{tA}e^{tB} = (A+B)e^{tA}e^{tB} = (A+B)V(t)\ .$$

Кроме того, $V(0) = E$. Следовательно, по определению (1.40)

$V(t) = e^{t(A+B)}$. Таким образом,

$$e^{tA}\ e^{tB} = e^{t(A+B)},\ \text{если}\ \ AB = BA.$$

Важно обратить внимание на то, что

$$e^{tA}\ e^{tB} \neq e^{t(A+B)},\ \ \text{если}\ \ \ AB \neq BA.$$

4. Очевидно, матрицы A и $(-A)$ перестановочны. Поэтому $e^{tA}\ e^{-tA} = e^{t(A-A)} = E.$ В результате мы приходим к выводу, что матрица e^{-tA} является обратной матрицей матричной экспоненты матрицы A:

$$(e^{tA})^{-1} = e^{-tA}.$$

(Заметим, что в данном случае для вычисления обратной матрицы достаточно в показателе экспоненты заменить t на $(-t)$.)

5. По этой причине $e^{tA}\ e^{tA} = (e^{tA})^2 = e^{2tA}$ и т. д.

$$(e^{tA})^m = e^{mtA}.$$

6. Из представления матричной экспоненты в виде (1.41) следует, что матричная экспонента сопряженной матрицы A^* равна сопряженной матричной экспоненте матрицы A:

$$e^{tA^*} = (e^{tA})^*.$$

Невырожденность решения

Ранее установлено, что матрица $e^{\mathbf{A}t}$ всегда невырождена. Докажем теперь, что этот факт вытекает из общего результата, заключающегося в том, что решение уравнения

$$\frac{dX}{dt} = A(t)X, \quad X(0) = I \tag{1.42}$$

является невырожденным в любом интервале $0 \le t \le t_1$, в котором существует интеграл $\int_o^{t_1} \|A(t)\| dt$.

Имеется несколько различных путей доказательства. Рассматриваемый метод базируется на тождестве Якоби:

$$|X(t)| = e^{\int_0^t SpA(s)ds}. \tag{1.43}$$

Для вывода этого результата рассмотрим производную от скалярной функции $|X(t)|$. Для упрощения обозначений рассмотрим двумерный случай.

Имеем

$$|X(t)| = \begin{vmatrix} x_1 & y_1 \\ x_2 & y_2 \end{vmatrix}, \tag{1.44}$$

причем

$$\begin{aligned} &\frac{dx_1}{dt} = a_{11}x_1 + a_{12}x_2; \quad \frac{dy_1}{dt} = a_{11}y_1 + a_{12}y_2; \\ &\frac{dx_2}{dt} = a_{21}x_1 + a_{22}x_2; \quad \frac{dy_2}{dt} = a_{21}y_1 + a_{22}y_2; \end{aligned} \tag{1.45}$$

и

$$\begin{aligned} &x_1(0) = 1, \quad y_1(0) = 0, \\ &x_2(0) = 0, \quad y_2(0) = 1. \end{aligned} \tag{1.46}$$

Далее имеем

$$\frac{d}{dt}|X(t)| = \begin{vmatrix} \frac{dx_1}{dt} & \frac{dy_1}{dt} \\ x_2 & y_2 \end{vmatrix} + \begin{vmatrix} x_1 & y_1 \\ \frac{dx_2}{dt} & \frac{dy_2}{dt} \end{vmatrix} =$$

$$= \begin{vmatrix} a_{11}x_1 + a_{12}x_2 & a_{11}y_1 + a_{12}y_2 \\ x_2 & y_2 \end{vmatrix} + \begin{vmatrix} x_1 & y_1 \\ a_{21}x_1 + a_{22}x_2 & a_{21}y_1 + a_{22}y_2 \end{vmatrix} = \tag{1.47}$$

$$= a_{11}\begin{vmatrix} x_1 & y_1 \\ x_2 & y_2 \end{vmatrix} + a_{22}\begin{vmatrix} x_1 & y_1 \\ x_2 & y_2 \end{vmatrix} = Sp(A(t))\left|X(t)\right|.$$

Поскольку, кроме того, $\left|X(t)\right| \equiv 1$, то

$$\left|X(t)\right| = e^{\int_0^t Sp\mathbf{A}(s)ds}. \qquad (1.48)$$

7. Пусть A и B – подобные матрицы: $A = TBT^{-1}$, где T – произвольная невырожденная матрица, $\det(T) \neq 0$. Так как $A^k = TB^kT^{-1}$, то

$$e^{tA} = E + \frac{t}{1!}TBT^{-1} + \frac{t^2}{2!}TB^2T^{-1} + ...\frac{t^k}{k!}TB^kT^{-1} + ... =$$
$$T\left(E + \frac{t}{1!}B + \frac{t^2}{2!}B^2 + ...\frac{t^k}{k!}B^k + ...\right)T^{-1} = Te^{tB}T^{-1}.$$

Таким образом, матричные экспоненты подобных матриц подобны.

8. Приведем оценки для нормы матричной экспоненты. В силу (1.41)

$$\left\|e^{tA}\right\| = \left\|E + \frac{t}{1!}A + \frac{t^2}{2!}A^2 + ... + \frac{t^k}{k!}A^k + ...\right\| \leq$$
$$\leq 1 + \left\|\frac{t}{1!}A\right\| + \left\|\frac{t^2}{2!}A^2\right\| + ... + \left\|\frac{t^k}{k!}A^k\right\| + ... \leq \left\|e^{tA}\right\|.$$

Получим теперь оценку $\left\|e^{tA}\right\|$ снизу. Имеем:

$$1 = \left\|e^{tA}\cdot e^{-tA}\right\| \leq \left\|e^{tA}\right\|\cdot\left\|e^{-tA}\right\| \leq \left\|e^{tA}\right\|\cdot e^{|t|\|A\|}, \text{ т.е. } e^{|t|\|A\|} \leq \left\|e^{tA}\right\|.$$

Итак, двухсторонние оценки нормы матричной экспоненты имеют вид:

$$e^{-|t|\|A\|} \leq \left\|e^{tA}\right\| \leq e^{|t|\|A\|}$$

Явная форма решения линейного дифференциального уравнения. Диагональные матрицы

Рассмотрим подход к решению дифференциального уравнения, отличный от предыдущего.

В уравнении

$$\frac{d\mathbf{x}}{dt} = A\mathbf{x}, \quad \mathbf{x}(0) = \mathbf{c}, \tag{1.49}$$

произведем замену переменных $\mathbf{x} = T\mathbf{y}$, где T — постоянная невырожденная матрица, которую мы определим в дальнейшем. Уравнение для $\mathbf{y}$ имеет вид

$$\frac{d\mathbf{y}}{dt} = T^{-1}AT\mathbf{y}, \quad \mathbf{y}(0) = t^{-1}\mathbf{c}. \tag{1.50}$$

Можно ли выбором матрицы T настолько упростить систему, чтобы она допускала непосредственное интегрирование?

Предположим, что мы нашли такую матрицу T, что матрица $T^{-1}AT$ является диагональной:

$$T^{-1}AT = \begin{vmatrix} \mu_1 & & & 0 \\ & \mu_2 & & \\ & & \cdot & \\ & & & \cdot \\ 0 & & & \mu_N \end{vmatrix}. \tag{1.51}$$

Если это выполнено, то уравнения (1.50) распадаются на N независимых уравнений вида:

$$\frac{dy_i}{dt} = \mu_i y_i, \quad y_i(0) = c_i', \quad i = 1,2,...,N. \tag{1.52}$$

Эти последние имеют простейшие решения: $y_i = e^{\mu_i t}c_i'$. После этого исходный вектор x легко определяется по известному вектору y.

Диагонализация матрицы

Рассмотрим весьма интересную, но трудную проблему

диагонализации матрицы A. Как известно, если матрица A является симметрической, то матрица T, обладающая требуемыми свойствами, всегда может быть найдена. При этом $T^{-1} = T'$. Но существуют и другие важные классы матриц, которые могут быть приведены к диагональной форме. Не менее важен тот факт, что помимо диагонального имеются и другие полезные канонические представления матриц.

Рассмотрим, однако, общий случай. Сразу же отметим, что набор величин $\{\mu_i\}$ должен быть точно таким же, как и $\{\lambda_i\}$, поскольку характеристические числа матрицы $T^{-1}AT$ те же, что и у матрицы A.

Отсюда следует, что столбцами T являются собственные векторы матрицы A. Обратно, если все λ_i различны, а T — матрица, столбцами которой являются соответствующие собственные векторы A, то

$$AT = T \begin{Vmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_N \end{Vmatrix}. \qquad (1.53)$$

Нами получен, таким образом, следующий важный результат.

Теорема 1. Если характеристические числа матрицы A различны, то существует матрица T такая, что

$$T^{-1}AT = \begin{Vmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_N \end{Vmatrix}. \qquad (1.54)$$

Как видим, предположение об отсутствии кратных характеристических чисел является весьма важным. Далее покажем, что такое простое представление, как (1.54), не может быть получено в общем случае.

Связь между двумя подходами

Как мы установили, один из методов решения линейного уравнения

$$\frac{d\mathbf{x}}{dt} = A\mathbf{x}, \quad \mathbf{x}(0) = \mathbf{c}, \tag{1.55}$$

приводит к выражению $\mathbf{x} = e^{At}\mathbf{c}$, в то время как другой метод, основанный на введении характеристических чисел и векторов, приводит к скалярным экспоненциальным функциям.

Установим связь между этими подходами (пока еще в предположении $\lambda_i \neq \lambda_j$), которая должна существовать вследствие единственности решения. В уравнении

$$\frac{dX}{dt} = AX, \quad X(0) = I \tag{1.56}$$

произведем замену переменных $X = TY$, где T выбрана в соответствии с (1.53). Тогда для Y получим уравнение

$$\frac{dY}{dt} = T^{-1}ATY, \quad Y(0) = T^{-1}, \tag{1.57}$$

или

$$\frac{dY}{dt} = \begin{Vmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_N \end{Vmatrix} Y, \quad Y(0) = T^{-1}. \tag{1.58}$$

Отсюда следует, что

$$Y = \begin{Vmatrix} e^{\lambda_1 t} & & & 0 \\ & e^{\lambda_2 t} & & \\ & & \ddots & \\ 0 & & & e^{\lambda_N t} \end{Vmatrix} T^{-1}. \qquad (1.59)$$

Следовательно,

$$X = e^{At} = T \begin{Vmatrix} e^{\lambda_1 t} & & & 0 \\ & e^{\lambda_2 t} & & \\ & & \ddots & \\ 0 & & & e^{\lambda_N t} \end{Vmatrix} T^{-1}. \qquad (1.60)$$

Подчеркнем еще раз, что это представление получено в предположении, что характеристические числа матрицы A различны.

1.3. Вычисление матричной экспоненты с учетом недиагонализуемости

Пусть для приближенного вычисления матричной экспоненты использовалась частичная сумма ряда (1.41):

$$e^{tA} \approx S_k = E + \frac{t}{1!}A + \frac{t^2}{2!}A^2 + \ldots + \frac{t^k}{k!}A^k.$$

Получим оценку нормы остаточного члена ряда (7):

$$\left\| e^{tA} - S_k \right\| = \left\| \frac{t^{k+1}}{(k+1)!}A^{k+1} + \frac{t^{k+2}}{(k+2)!}A^{k+2} + \frac{t^{k+3}}{(k+3)!}A^{k+3} + \ldots \right\| \le$$

$$\frac{(|t|\|A\|)^k}{(k+1)!}\left(\frac{|t|\|A\|}{1} + \frac{(|t|\|A\|)^2}{k+2} + \frac{(|t|\|A\|)^3}{(k+2)(k+3)} + \ldots \right) \le$$

$$\le \frac{(|t|\|A\|)^k}{(k+1)!}\left(\frac{|t|\|A\|}{1!} + \frac{(|t|\|A\|)^2}{2!} + \frac{(|t|\|A\|)^3}{3!} + \ldots \right) = \frac{(|t|\|A\|)^k}{(k+1)!}\,(e^{|t|\|A\|} - 1).$$

Таким образом, оценка для нормы погрешности приближения матричной экспоненты имеет вид:

$$\left\| e^{tA} - S_k \right\| \le \frac{(|t|\|A\|)^k}{(k+1)!}\,(e^{|t|\|A\|} - 1).$$

1.4. Вычисление матричной экспоненты с привлечением жордановой формы.

Обозначим через J жорданову форму $n \times n$-матрицы A, $A = TJT^{-1}$, состоящую из p стандартных жордановых $n_k \times n_k$-клеток J_k, $k = 1,2,...,p$, $n_1 + n_2 + ... + n_p = n$:

$$J = \begin{bmatrix} J_1 & & & \\ & J_2 & & \\ \ldots & \ldots & \ldots & \ldots \\ & & & J_p \end{bmatrix}.$$

Согласно свойству (1.54) $e^{tA} = Te^{tJ}T^{-1}$.

Найдем матричную экспоненту жордановой формы. По определению (1.41)

$$e^{tJ} = E + \frac{t}{1!}J + \frac{t^2}{2!}J^2 + \ldots\frac{t^k}{k!}J^k + \ldots \ . \qquad (1.61)$$

Поскольку

$$J^m = \begin{bmatrix} J_1^m & & & \\ & J_2^m & & \\ \ldots & \ldots & \ldots & \ldots \\ & & & J_p^m \end{bmatrix},$$

то после подстановки в (1.61) выражений $J, J^2, ..., J^k, ...$ получаем

$$e^{tJ} = \begin{bmatrix} e^{tJ_1} & & & \\ & e^{tJ_2} & & \\ \dots & \dots & \dots & \dots \\ & & & e^{tJ_p} \end{bmatrix}.$$

Таким образом, для определения матричной экспоненты жордановой формы осталось найти матричную экспоненту стандартной жордановой клетки, имеющей вид:

$$J_k = \begin{bmatrix} \lambda_k & 1 & & & & \\ & \lambda_k & 1 & & & \\ \dots & \dots & \dots & \dots & \dots & \dots \\ & & & & \lambda_k & 1 \\ & & & & & \lambda_k \end{bmatrix}.$$

С этой целью представим J_k в виде суммы двух матриц:

$$J_k = \begin{bmatrix} \lambda_k & & & & & \\ & \lambda_k & & & & \\ \dots & \dots & \dots & \dots & \dots & \dots \\ & & & & \lambda_k & \\ & & & & & \lambda_k \end{bmatrix} + \begin{bmatrix} 0 & 1 & & & & \\ & 0 & 1 & & & \\ \dots & \dots & \dots & \dots & \dots & \dots \\ & & & & 0 & 1 \\ & & & & & 0 \end{bmatrix} = \lambda_k E + B.$$

Так как матрицы E и B перестановочны, то (свойство 3)

$$e^{t(\lambda_k E + B)} = e^{t\lambda_k E} e^{tB} = e^{\lambda_k t} e^{tB}.$$

Далее воспользуемся представлением матричной экспоненты e^{tB} в виде бесконечного ряда. Для этого потребуются степени матрицы B. Имеем:

$$B=\begin{bmatrix} 0 & 1 & & & & \\ & 0 & 1 & & & \\ \dots & \dots & \dots & \dots & \dots & \dots \\ & & & & 0 & 1 \\ & & & & & 0 \end{bmatrix},\ B^2=\begin{bmatrix} 0 & 0 & 1 & & & \\ & 0 & 0 & 1 & & \\ \dots & \dots & \dots & \dots & \dots & \dots \\ & & & & 0 & 0 \\ & & & & & 0 \end{bmatrix},\ \dots,$$

$$B^{n_k-1}=\begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 1 \\ & 0 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ & & & \dots & 0 & 0 \\ & & & \dots & & 0 \end{bmatrix}.$$

Более высокие степени матрицы B дают нулевую матрицу. В результате матричная экспонента определяется конечной суммой ряда:

$$e^{tB}=E+\frac{t}{1!}B+\frac{t^2}{2!}B^2+\ \dots\dots+\frac{t^{n_k-1}}{(n_k-1)!}B^{n_k-1}\ .$$

После подстановки сюда степеней матрицы B получаем:

$$e^{tB}=\begin{bmatrix} 1 & \frac{t}{1!} & \frac{t^2}{2!} & \frac{t^3}{3!} & \dots & \frac{t^{n_k-1}}{(n_k-1)!} \\ & 1 & \frac{t}{1!} & \frac{t^2}{2!} & \dots & \frac{t^{n_k-2}}{(n_k-2)!} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ & & & & \dots\ 1 & \frac{t}{1!} \\ & & & & \dots & 1 \end{bmatrix}.$$

Итак, матричная экспонента стандартной жордановой клетки имеет вид:

$$e^{tJ_k} = e^{\lambda_k t}\begin{bmatrix} 1 & \frac{t}{1!} & \frac{t^2}{2!} & \frac{t^3}{3!} & \dots & \frac{t^{n_k-1}}{(n_k-1)!} \\ & 1 & \frac{t}{1!} & \frac{t^2}{2!} & \dots & \frac{t^{n_k-2}}{(n_k-2)!} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ & & & \dots & 1 & \frac{t}{1!} \\ & & & \dots & & 1 \end{bmatrix} =$$

$$= \begin{bmatrix} e^{\lambda_k t} & \frac{te^{\lambda_k t}}{1!} & \frac{t^2 e^{\lambda_k t}}{2!} & \frac{t^3 e^{\lambda_i t}}{3!} & \dots & \frac{t^{n_k-1} e^{\lambda_i t}}{(n_k-1)!} \\ & e^{\lambda_k t} & \frac{te^{\lambda_k t}}{1!} & \frac{t^2 e^{\lambda_k t}}{2!} & \dots & \frac{t^{n_k-2} e^{\lambda_k t}}{(n_k-2)!} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ & & & \dots & e^{\lambda_k t} & \frac{te^{\lambda_k t}}{1!} \\ & & & \dots & & e^{\lambda_k t} \end{bmatrix}.$$

Напомним, что жорданова форма матрицы определяется ее спектром. Пусть λ_k - собственное число матрицы A кратности m_k; r_k – ранг матрицы $(A-\lambda_k E)$; $d_k = n - r_k$ – дефект матрицы A, равный числу собственных векторов, соответствующих λ_k, $d_k \le m_k$. Если $d_k = m_k$, то λ_k соответствуют m_k одномерных жордановых клеток. В противном случае число одномерных клеток равно $d_k - 1$. К ним добавляется стандартная жорданова клетка размера $m_k - d_k + 1$. Одномерной клетке соответствует столбец матрицы T, являющийся собственным вектором матрицы A. Если клетка неодномерная, то ей соответствуют столбцы матрицы A, представляющие собственный вектор и цепочку присоединенных векторов, число которых равно $m_k - d_k$.

Рассмотрим пример. Определим матричную экспоненту (8×8) - матрицы A, $e^{tA} = T\, e^{tJ} T^{-1}$, используя жорданову форму J матрицы A, если

спектр матрицы A состоит из следующих собственных чисел: собственное число λ_1 кратности 3, ранг $(A-\lambda_1 E)$ равен 7; собственное число λ_2 кратности 1, ранг $(A-\lambda_2 E)$ равен 7; собственное число λ_3 кратности 4, ранг $(A-\lambda_3 E)$ равен 5. В этом случае жорданова форма матрицы A состоит из трех одномерных клеток, одной клетки размера 3 и одной клетки размера 2. Жорданова форма и матричная экспонента жордановой формы имеют вид:

$$J=\begin{bmatrix}
\lambda_1 & 1 & 0 & & & & & \\
0 & \lambda_1 & 1 & & & & & \\
0 & 0 & \lambda_1 & & & & & \\
. & & & \lambda_2 & & & & \\
 & & & & \lambda_3 & & & \\
 & & & & & \lambda_3 & & \\
 & & & & & & \lambda_3 & 1 \\
 & & & & & & 0 & \lambda_3
\end{bmatrix},$$

$$e^{tJ}=\begin{bmatrix}
e^{\lambda_1 t} & te^{\lambda_1 t} & \frac{t^2}{2}e^{\lambda_1 t} & & & & & \\
0 & e^{\lambda_1 t} & te^{\lambda_1 t} & & & & & \\
0 & 0 & e^{\lambda_1 t} & & & & & \\
. & & & e^{\lambda_2 t} & & & & \\
 & & & & e^{\lambda_3 t} & & & \\
 & & & & & e^{\lambda_3 t} & & \\
 & & & & & & e^{\lambda_3 t} & te^{\lambda_3 t} \\
 & & & & & & 0 & e^{\lambda_3 t}
\end{bmatrix}$$

Столбцы матрицы перехода $T=[T_1,T_2,T_3,T_4,T_5,T_6,T_7,T_8]$ удовлетворяют условиям, которые следуют из определения подобия матриц A и J: $A=TJT^{-1}$ или, что то же самое, $AT=TJ$. Приравнивая соответствующие столбцы матриц AT и TJ, получаем

$$AT_1=\lambda_1 T_1,\ AT_2=\lambda_1 T_2+T_1,\ AT_3=\lambda_1 T_3+T_2\ \text{—}$$

цепочка из собственного вектора T_1 и присоединенных векторов T_2, T_3, соответствующая собственному числу λ_1;

$$AT_4 = \lambda_2 T_4,\ AT_5 = \lambda_3 T_5,\ AT_6 = \lambda_3 T_6\ —$$

собственные векторы матрицы A, соответствующие собственным числам λ_2 и λ_3;

$$AT_7 = \lambda_3 T_7,\ AT_8 = \lambda_3 T_8 + T_7\ —$$

цепочка из собственного вектора T_7 и присоединенного вектора T_8, соответствующая собственному числу λ_3.

Заметим, что вектор–функцию $y(t)$, являющуюся решением задачи Коши (1.35), можно представить в виде:

$$y(t) = e^{tA}y^0 = Te^{tJ}T^{-1}y^0 = Te^{tJ}C,\ C = T^{-1y^0}.$$

Пусть далее $y^{[1]}(t), y^{[2]}(t),..., y^{[n]}(t)$ – столбцы матричной функции $V(t) = Te^{tJ}$. Тогда решение задачи Коши (1.35) имеет вид линейной комбинации столбцов матричной функции $V(t)$:

$$y(t) = C_1 y^{[1]}(t) + C_2 y^{[2]}(t) + ... + C_n y^{[n]}(t), \qquad (1.62)$$

где $C_1, C_2, ...C_n$ – компоненты вектора C, определяемые из системы линейных алгебраических уравнений с матрицей T:

$$TC = y^0. \qquad (1.63)$$

Поскольку T – невырожденная матрица, то формула (1.62) описывает любые решения однородной системы дифференциальных уравнений (1.35), так как коэффициенты линейной комбинации однозначно определяются из (1.63) условием прохождения решения через произвольно задаваемую точку y^0 в n – мерном евклидовом пространстве. Очевидно, каждая из вектор-функций $y^{[j]}(t),\ j = 1,2,...,n$, является

решением однородной системы.

1.5. Представление матричной экспоненты в виде матричного полинома.

Пусть $\lambda_1, \lambda_2, \ldots, \lambda_n$ – собственные числа $(n \times n)$ – матрицы A. Введем в рассмотрение матричные полиномы :

$$\begin{aligned}
&P_1(\mathbf{A}) = \mathbf{A} - \lambda_1 \mathbf{E} \\
&P_2(\mathbf{A}) = P_1(\mathbf{A})\,(\mathbf{A} - \lambda_2 \mathbf{E}) \\
&P_3(\mathbf{A}) = P_2(\mathbf{A})\,(\mathbf{A} - \lambda_3 \mathbf{E}) \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
&P_{n-1}(\mathbf{A}) = P_{n-2}(\mathbf{A})\,(\mathbf{A} - \lambda_{n-1} \mathbf{E}) \\
&P_n(\mathbf{A}) = P_{n-1}(\mathbf{A})\,(\mathbf{A} - \lambda_n \mathbf{E}).
\end{aligned} \tag{1.64}$$

Как известно из курса линейной алгебры, матрица всегда удовлетворяет своему характеристическому уравнению (теорема Гамильтона—Кэли). Поэтому $P_n(A) = 0$. Из определения матричных полиномов следуют равенства:

$$\begin{aligned}
&P_1(\mathbf{A}) + -\lambda_1 \mathbf{E} = \mathbf{A} \\
&P_2(\mathbf{A}) + \lambda_2 P_1(\mathbf{A}) = \mathbf{A} P_1(\mathbf{A}) \\
&P_3(\mathbf{A}) + \lambda_3 P_2(\mathbf{A}) = \mathbf{A} P_2(\mathbf{A}) \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
&P_{n-1}(\mathbf{A}) + \lambda_{n-1} P_{n-2}(\mathbf{A}) = \mathbf{A} P_{n-2}(\mathbf{A}) \\
&0 + \lambda_n P_{n-1}(\mathbf{A}) = \mathbf{A} P_{n-1}(\mathbf{A}).
\end{aligned} \tag{1.65}$$

Здесь учтено, что матрица A перестановочна с матричными полиномами.

Далее потребуются выражения компонент $\psi_1(t), \psi_2(t), \ldots, \psi_n(t)$ вектор-функции $\psi(t)$, которая является решением задачи Коши :

$$\frac{d\psi}{dt} = \begin{bmatrix} \lambda_1 & & & & & & \\ 1 & \lambda_2 & & & & & \\ \dots & \dots & \dots & \dots & \dots & \dots & \\ & & & 1 & \lambda_{n-1} & & \\ & & & & 1 & \lambda_n & \end{bmatrix} \psi, \quad \psi(0) = \begin{bmatrix} 1 \\ 0 \\ \dots \\ 0 \\ 0 \end{bmatrix},$$

с покомпонентной записью:

$$\begin{aligned} &\frac{d\psi_1}{dt} = \lambda_1\psi_1, \quad \psi_1(0) = 1, \\ &\frac{d\psi_2}{dt} = \lambda_2\psi_2 + \psi_1, \psi_2(0) = 0, \\ &\frac{d\psi_n}{dt} = \lambda_n\psi_n + \psi_{n-1}, \psi_n(0) = 0. \end{aligned} \tag{1.66}$$

Очевидно, первая строка (1.66) представляет задачу Коши, определяющую $\psi_1(t)$, решение которой имеет вид: $\psi_1(t) = e^{\lambda_1 t}$.

Остальные компоненты $\psi(t)$ находятся из рекуррентных соотношений:

$$\psi_k(t) = e^{\lambda_k t}\int_0^t e^{-\lambda_k s}\psi_{k-1}(s)ds \ , k=2,3....,n. \tag{1.67}$$

Действительно, предположим , что компонента $\psi_{k-1}(t)$ уже известна. Если $\psi_k(t)$ – $k-$я компонента решения задачи Коши (1.66), то $k-$е уравнение системы (1.66) можно представить в виде тождества:

$$e^{-\lambda_k t}(\frac{d\psi_n}{dt}(t) - \lambda_k\psi(t)) \equiv e^{-\lambda_k t}\psi_{n-1}(t),$$

или

$$\frac{d}{dt}(e^{-\lambda_k t}\psi_k(t)) \equiv e^{-\lambda_k t}\psi_{n-1}(t).$$

После интегрирования левой и правой частей тождества от 0 до t получаем:

$$e^{-\lambda_k t}\psi_k(t)) \equiv \int_0^t e^{-\lambda_k s}\psi_{k-1}(s)ds,$$

откуда и следует формула (1.66).

Например, если $\lambda_1 \neq \lambda_2$, то

$$\psi_2(t) = e^{\lambda_2 t}\int_0^t e^{-\lambda_2 s}e^{\lambda_1 s}ds = e^{\lambda_2 t}\frac{e^{(\lambda_2-\lambda_1)t}-1}{\lambda_2-\lambda_1} = \frac{e^{\lambda_2 t}-e^{\lambda_1 t}}{\lambda_2-\lambda_1}. \qquad (1.68)$$

Если $\lambda_1 = \lambda_2$, то

$$\psi_2(t) = e^{\lambda_1 t}\int_0^t e^{-\lambda_1 s}e^{\lambda_1 s}ds = te^{\lambda_1 t}.$$

При равенстве собственных чисел $\lambda_1, \lambda_2, \ldots, \lambda_k$ функции $\psi_1(t), \psi_2(t),....,\psi_k(t)$ имеют, согласно рекуррентной формуле, следующие выражения:

$$\psi_1(t) = e^{\lambda_1 t}, \quad \psi_2(t) = \frac{t}{1}e^{\lambda_1 t}, \quad \ldots., \quad \psi_k(t) = \frac{t^{k-1}}{(k-1)!}e^{\lambda_1 t}. \qquad (1.69)$$

Покажем теперь, что матричная функция

$Y(t) = \psi_1(t)E + \psi_2(t)P_1(A) + \psi_3(t)P_2(A) ... + \psi_n(t)P_{n-1}(A),$

где матричные полиномы $P_1(A), P_2(A),\ldots, P_{n-1}(A)$ и функции $\psi_1(t), \psi_2(t),..., \psi_n(t)$, определены в (1.64) и (1.66) соответственно, является решением задачи Коши (1.40). Действительно, $Y(0) = E$ в силу задания начальных условий задачи Коши (1.66). Составим далее выражение производной матричной функции Y(t), учитывая при этом, что компоненты вектор-функции $\psi(t)$ удовлетворяют уравнениям (1.66). Имеем:

$$\frac{dY}{dt}(t) = \frac{d\psi_1}{dt}(t)E + \frac{d\psi_2}{dt}(t)P_1(A) + \frac{d\psi_3}{dt}(t)P_3(A) + ... + \frac{d\psi_n}{dt}(t)P_{n-1}(A) =$$

$$= \lambda_1\psi_1(t)E + (\lambda_2\psi_2(t)+\psi_1(t))P_1(A) + (\lambda_3\psi_3(t)+\psi_2(t))P_3(A) + \\ + ... + (\lambda_n\psi_n(t)+\psi_{n-1}(t))P_{n-1}(A) =$$

$$= \psi_1(t)(\lambda_1 E + P_1(A)) + \psi_2(t)(\lambda_2 P_1(A) + P_2(A)) + (\lambda_3\psi_3(t) + \psi_2(t))P_3(A) + \\ + ... + \lambda_n\psi_n(t)P_{n-1}(A).$$

После замены коэффициентов при $\psi_k(t)$ на их выражения из (1.66) получаем

$$\frac{dY}{dt}(t) = \psi_1(t)A + \psi_2(t)AP_1(A) + \psi_3(t)AP_2(A) \ldots + \psi_n(t)AP_{n-1}(A) =$$

$$= A(\psi_1(t)E + \psi_2(t)P_1(A) + \psi_3(t)P_2(A) \ldots + \psi_n(t)P_{n-1}(A)) = AY(t).$$

В результате мы доказали, что Y(t) является матричной экспонентой матрицы A:

$$e^{tA} = \psi_1(t)E + \psi_2(t)P_1(A) + \psi_3(t)P_2(A) \ldots + \psi_n(t)P_{n-1}(A). \qquad (1.70)$$

1.6. Примеры вычисления матричных экспонент

Пример 1.

Найдем матричную экспоненту матрицы A и решение задачи Коши (1.35), если

$$A = \begin{bmatrix} 2 & 1 \\ 3 & 4 \end{bmatrix}.$$

Характеристическое уравнение матрицы A имеет вид:

$$\det(A - \lambda E) = \det \begin{bmatrix} 2-\lambda & 1 \\ 3 & 4-\lambda \end{bmatrix} = \lambda^2 - 6\lambda + 5 = 0.$$

Отсюда находим, что $\lambda_1 = 1$ и $\lambda_2 = 5$ - собственные числа матрицы A. Им соответствуют матричный полином

$$P_1(A) = A - \lambda_1 E = \begin{bmatrix} 1 & 1 \\ 1 & 3 \end{bmatrix}$$

и функции

$$\psi_1(t) = e^t, \quad \psi_2(t) = \frac{e^{5t} - e^t}{5-1} = \frac{e^{5t} - e^t}{4}.$$

Таким образом, согласно (1.70),

$$e^{tA} = e^t \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \frac{e^{5t} - e^t}{4} \begin{bmatrix} 1 & 1 \\ 3 & 3 \end{bmatrix} = \begin{bmatrix} \frac{e^{5t} + 3et}{4} & \frac{e^{5t} - e^t}{4} \\ 3\frac{e^{5t} - e^t}{4} & 3\frac{e^{5t} + e^t}{4} \end{bmatrix}.$$

Запишем решение задачи Коши (1.35):

$$y(t) = \begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix} = e^{tA} \begin{bmatrix} y_1^0 \\ y_2^0 \end{bmatrix} = \begin{bmatrix} \frac{e^{5t} + 3et}{4} y_1^0 + & \frac{e^{5t} - e^t}{4} y_2^0 \\ 3\frac{e^{5t} - e^t}{4} y_1^0 + & 3\frac{e^{5t} + e^t}{4} y_2^0 \end{bmatrix},$$

т.е.

$$y_1(t) = \frac{e^{5t} + e^t}{4} y_1^0 + \frac{e^{5t} - e^t}{4} y_2^0, \quad y_2(t) = 3\frac{e^{5t} - e^t}{4} y_1^0 + 3\frac{e^{5t} + e^t}{4} y_2^0.$$

Пример 2.

Рассмотрим линейное дифференциальное уравнение

$$u'' - 3u' + 2u = 0 \tag{1.71}$$

и будем искать решение, удовлетворяющее условиям

$$u(0) = 1, \quad u'(0) = 0.$$

Поскольку характеристическое уравнение

$$\lambda^2 - 3\lambda + 2 = 0 \tag{1.72}$$

имеет корни $\lambda_1 = 1$ и $\lambda_2 = 2$, то решение (1.71) должно иметь вид

$$u = c_1 e^{2t} + c_2 e^{t}. \tag{1.73}$$

Постоянные коэффициенты c_1 и c_2 определяются из условий

$$\begin{aligned} c_1 + c_2 &= 1, \\ 2c_1 + c_2 &= 0, \end{aligned} \tag{1.74}$$

что дает

$$c_1 = -1, \qquad c_2 = 2. \tag{1.75}$$

Окончательно решение записывается в форме

$$u = 2e^{t} - e^{2t}. \tag{1.76}$$

Рассмотрим вместо этой процедуры метод приведения уравнения (1.71) к системе линейных дифференциальных уравнений первого порядка (1.55). Решение выпишем в виде (1.60).

Введем обозначения: $x_1 = u$, $x_2 = u'$. Тогда уравнение (1.71) может быть записано в виде:

$$\begin{aligned} \frac{dx_1}{dt} &= x_2; \\ \frac{dx_2}{dt} &= -2x_1 + 3x_2; \end{aligned}, \tag{1.77}$$

а начальные условия примут следующий вид:

$$x_1(0) = 1, \quad x_2(0) = 0.$$

Матрица A имеет вид:

$$A = \begin{Vmatrix} 0 & 1 \\ -2 & 3 \end{Vmatrix}.$$

Собственные значения уже вычислены как корни квадратного уравнения (1.72): $\lambda_1 = 1$ и $\lambda_2 = 2$. Для получения матрицы T вычислим собственные векторы матрицы A. Первый собственный вектор,

соответствующий $\lambda_1 = 1$, должен удовлетворять условию $Ay = \lambda_1 y$, т. е. $Ay = y$. Имеем

$$\begin{Vmatrix} 0 & 1 \\ -2 & 3 \end{Vmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \Rightarrow \begin{cases} y_2 = y_1 \\ -2y_1 + 3y_2 = y_2 \end{cases} \Rightarrow y_2 = y_1.$$

В качестве собственного (ненулевого) вектора с точностью до скалярного множителя возьмем вектор $y = (1,1)$. Второй вектор должен удовлетворять уравнению $Ay = \lambda_2 y$, что при $\lambda_2 = 2$ соответствует $Ay = 2y$. Получим

$$\begin{Vmatrix} 0 & 1 \\ -2 & 3 \end{Vmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = 2\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \Rightarrow \begin{cases} y_2 = 2y_1 \\ -2y_1 + 3y_2 = 2y_2 \end{cases} \Rightarrow y_2 = 2y_1.$$

Выберем искомый вектор в виде $y = (1,2)$. Тогда матрица T имеет вид:

$$T = \begin{Vmatrix} 1 & 1 \\ 1 & 2 \end{Vmatrix}. \tag{1.78}$$

Вычислив обратную к ней матрицу

$$T^{-1} = \begin{Vmatrix} 2 & -1 \\ -1 & 1 \end{Vmatrix}, \tag{1.79}$$

мы можем записать матрицу (1.60), которая позволит определить решение для заданного вектора начальных условий:

$$X = e^{At} = T \begin{Vmatrix} e^t & 0 \\ 0 & e^{2t} \end{Vmatrix} T^{-1} = \begin{Vmatrix} 1 & 1 \\ 1 & 2 \end{Vmatrix} \begin{Vmatrix} e^t & 0 \\ 0 & e^{2t} \end{Vmatrix} \begin{Vmatrix} 2 & -1 \\ -1 & 1 \end{Vmatrix} = \begin{Vmatrix} 2e^t - e^{2t} & e^{2t} - e^t \\ 2e^t - 2e^{2t} & 2e^{2t} - e^t \end{Vmatrix}. \tag{1.80}$$

Для начальных условий $x_1(0) = 1, \quad x_2(0) = 0$ решение получим по формуле $x(t) = X(t)x(0)$

$$\begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} = \begin{pmatrix} 2e^t - e^{2t} & e^{2t} - e^t \\ 2e^t - 2e^{2t} & 2e^{2t} - e^t \end{pmatrix} \begin{pmatrix} x_1(0) \\ x_2(0) \end{pmatrix} = \\ = \begin{pmatrix} 2e^t - e^{2t} & e^{2t} - e^t \\ 2e^t - 2e^{2t} & 2e^{2t} - e^t \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 2e^t - e^{2t} \\ 2e^t - 2e^{2t} \end{pmatrix}. \tag{1.81}$$

Возвращаясь к прежним обозначениям, видим, что решение $x_1(t) = 2e^t - e^{2t}$, полученное с помощью этого метода, совпадает с решением (1.76): $u = 2e^t - e^{2t}$.

Замечание. Легко проверить, что матрица T состоит из правых собственных векторов, расположенных по столбцам. В то же время T^{-1} - матрица, в которой левые собственные векторы расположены по строкам.

Тема 2. Конечно-разностный подход к рассеянию света на оптических решетках

2.1. Изложение основ метода

В данной теме описывается подход к решению проблемы рассеяния света на оптической решетке, предложенный в работе [2]. В этом подходе система дискретизируется с точностью до второго порядка или до псевдочетвертого порядка в вертикальном направлении. Данный подход отличается от других тем, что линейная система может быть преобразована в систему с ленточной матрицей или к начальной задаче. Следует только знать соответствие между обобщенным током при z и $z + h/2$ и электрическое поле при z, или соответствие между значениями электрического поля в двух соседних точках дискретизации. В таком случае можно вывести формулу устойчивой рекурсии.

Метод обладает следующими преимуществами:

- Можно использовать матричную диагонализацию при необходимости, например, когда свойства материала не меняются в слое, а толщина велика настолько, что требует большого числа точек дискретизации.
- Требуется минимальное количество памяти компьютера. Метод решения систем с ленточной матрицей хранит всю матрицу, в нашем же методе необходимо знать матрицу лишь для отдельного шага.
- Существует численная устойчивость по сравнению с другими методами, основанными на матрице перехода, поэтому нет нужды привлекать метод стрельбы, используемый другими авторами. Показано, что метод работает в несколько раз быстрее, чем широко распространенный строгий метод связанных волн.

Конечно-разностная в вертикальном направлении формулировка

В последние несколько лет возрос интерес к использованию оптической спектрометрии для проведения измерений критических размеров (КР) линий и других структур в интегрально-оптических контурах. Наиболее распространенным методом численного моделирования спектрометрических данных является так называемый строгий метод связанных волн (rigorous coupled wave analysis - RCWA), иначе (но не вполне адекватно) называемый методом разложения Фурье [3]. Изначально RCWA-метод столкнулся с проблемами сходимости при моделировании поперечных магнитных полей. Лишь позднее было осознано, что формирование функции диэлектрической проницаемости различными способами может обеспечить значительно большую скорость сходимости [4, 5]. С физической точки зрения такие переформулировки приводят к правильным результатам, когда в статическом пределе число учитываемых гармоник становится равным 1. Это важное открытие в конце концов превратило RCWA-метод в универсальный и жизнеспособный подход к решению задач дифракции. Много усилий было направлено на уменьшение числа учитываемых в решении Фурье гармоник для небинарных (непрямоумольных) решеток [6].

В RCWA-методе электрическое (или магнитное) поле раскладывается в ряд Фурье, а система нарезается слоями в ветрикальном направлении. В каждом слое точное поле E раскладывается в ряд по волновым функциям (приближение возникает при удержании конечного числа членов функционального ряда). Собственные векторы получаются при диагонализации матрицы. Общее число операций с плавающей точкой пропорционально cNN_z, где N число Фурье слагаемых, а N_z - число слоев, используемых для вычисления, константа c зависит от метода решения системы линейных алгебраических уравнений, и основной вклад

получает при диагонализации матрицы, особенно в случае поглощающих материалах.

Другие подходы включают модальные методы, интегральные методы, специальные методы граничных интегральных уравнений (элементов) и др. Модальный метод считается лучше RCWA-метода, но он требует точного вычисления собственных значений и собственных векторов, которые считаются медленно и не очень надежно. В интегральном методе узким местом является вычисление функции Грина при работе с небольшим числом точек. Он требует также аккуратного использования многослойных функций Грина. Его численная реализация доставляет много хлопот. Для глубоких бороздок требуется большое число точек дискретизации, и прямая факторизация результирующей матрицы становится слишком медленной и не эффективной.

По мере того, как структуры критического размера становятся все более сложными, численное использование строгого метода связанных волн становится все более медленным. Даже для высокоскоростных процессоров реализация этих вычислений представляет определенные сложности. А мониторинг процесса очень желательно проводить в режиме реального времени. Для этого требуются более скоростные алгоритмы. Можно ускорить процесс исключением матричной диагонализации. Но даже после исключения матричной диагонализации остаются другие матричные операции. Допустимой альтернативой является конечно-разностная формулировка в вертикальном направлении. Развитие конечно-разностных методов, использовавшихся в прошлом, задерживалось многими обстоятельствами: локальная точность, глобальная устойчивость, скорость решения результирующего уравнения. Хорошо известно, что конечно-разностные методы могут привести к нежелательному убыванию или возрастанию энергии. В случае слабой глобальной устойчивости

требуется огромное число слоев, что приводит к бесполезности перехода от RCWA-метода, значительно менее чувствительного к шагам дискретизации. В данной работе описывается способ значительного повышения глобальной устойчивости, точности и скорости конечно-разностного подхода.

Теоретичекий базис

В проблеме рассеяния на решетке, когда электрическое поле $E(x,z)$ раскладывается по плоским волнам, или, иными словами, для многокомпонентной задачи второго порядка мы имеем следующее дифференциальное уравнение второго порядка:

$$\frac{d}{dz}p(z)\frac{d}{dz}\psi(z) = A(z)\psi(z). \tag{2.1}$$

Если воспользоваться понятием обобщенного тока J, то вышеуказанное уравнение можно переписать в виде дифференциального уравнения первого порядка:

$$\frac{d}{dz}\begin{pmatrix}\psi\\ J\end{pmatrix} = aY = \begin{pmatrix}0 & p^{-1}(z)\\ A(z) & 0\end{pmatrix}\begin{pmatrix}\psi\\ J\end{pmatrix} \equiv \begin{pmatrix}0 & B(z)\\ A(z) & 0\end{pmatrix}\begin{pmatrix}\psi\\ J\end{pmatrix}. \tag{2.2}$$

В простом случае неконической дифракции на решетке мы имеем дело раздельно с $TE-$ и $TM-$модами. Для $TE-$моды $p(z)$ является единичной матрицей, а для $TM-$моды $(\psi = H)$, $p(z) \equiv \left[\varepsilon^{-1}\right]$ — матричная функция обратной диэлектрической проницаемости, за деталями можно обратиться, например, к статье Мохарама [3]. Для конической дифракции или для дифракции на трехмерных структурах $\psi = (E_x, E_y)$ и после простого поворота матрицы получаем:

$$A = \begin{pmatrix}k_x e_x k_x + k_y e_y k_y & k_y e_y k_x - k_x e_x k_y\\ k_x e_y k_x - k_y e_x k_x & k_y e_x k_y + k_x e_y k_x - k^2\end{pmatrix},\ B = \begin{pmatrix}1 - k e_z^{-1} k & 0\\ 0 & 1\end{pmatrix},$$

где $\overline{k}_x = k_x / k$ и $\overline{k}_y = k_y / k$, а $k = (k_x^2 + k_y^2)^{1/2}$ и пространственные переменные безразмерны. Важно подчеркнуть, что A и B являются симметричными. Если матрицы являются вещественными, это частный случай более общего случая так называемых гамильтоновых матриц

$$\begin{pmatrix} D & B \\ A & -D^T \end{pmatrix}.$$

Для неконической дифракции проблема рассеяния может быть полностью сформулирована со следующими граничными условиями. Данный слой и все нижележащие слои однородны, поэтому электрические поля можно записать в диагональной форме:

$$E_j(z) = f_j\left(e^{ik_j z} + re^{-ik_j z}\right),$$

$$J_j = ip_j k_j f_j\left(e^{ik_j z} - re^{-ik_j z}\right),$$

где f_j подлежат определению. Коэффициенты k_j могут быть вычислены из рекурсивных отношений для однородных слоев многослойного материала.

Над областью периодичности мы имеем (в векторной форме):

$$E(z) = (1+r)f,$$

$$J = p_L q(1-r)f = p_L q(1-r)(1+r)^{-1}E \equiv VE, \qquad (2.3)$$

где $q \equiv ik$, r и V – диагональные матрицы. Аналогично, в среде падающего света имеем

$$J_0 = p_0 q(1-R)f_0 = p_0 q(1-R)(1+R)^{-1}E_0 = p_0 q\left(\frac{2}{1+R} - 1\right)E_0 \equiv \omega_0 E_0,$$

где R и ω_0 – заполненные матрицы. Матрица R восстанавливается по матрице отражения. Если матрица ω_0 известна, мы немедленно получаем

$$R = 2(\omega_0 + pq)^{-1} pq - 1. \tag{2.4}$$

В более общей ситуации по заданному падающему электрическому полю ψ_{in} рассеянное поле может быть вычислено по формулам:

$$R = 2(S^T \omega_0 S + q)^{-1} q - 1 = 2S^{-1}(\omega_0 + pSqS^{-1})^{-1} pSq - 1 \tag{2.5}$$

$$\psi_0 = SRf_{in} = 2(\omega_0 + pSqS^{-1})^{-1} pSqf_{in} - Sf_{in} \tag{2.6}$$

$$\begin{aligned} &= 2(\omega_0 + pSqS^{-1})^{-1} pSq\psi_{in} - \psi_{in} = \\ &= 2(\omega_0 + (S^{-1})^T qS^{-1})^{-1}(S^{-1})^T \psi_{in} - \psi_{in} = \end{aligned} \tag{2.7}$$

$$= (\omega_0 + \bar{j}_0)\psi_0 - \psi_{in}, \tag{2.8}$$

где S – матрица подобия, которая диагонализирует матрицу BA, легко вычисляемую в области падения.

Основная идея заключается в том, что не нужно вычислять точное поле при всех значениях вертикальной переменной z, а необходима лишь информация о матрице отношения между током J (H в случае TM поля) и ψ (для E поля).

Алгоритм решения начальной задачи

Введем обозначение: $Y \equiv \begin{pmatrix} \psi \\ J \end{pmatrix}$.

Решение в точке $z + h$ через решение в точке z можно выразить следующим образом

$$Y(z+h) = Te^{\int_z^{z+h} a(z')dz'} Y(z),$$

где T означает упорядоченное по времени произведение. Оператор может быть переписан в терминах разложения Магнуса [7]

$$Te^{\int_z^{z+h} a(z')dz'} = e^{\Omega(z+h,z)}$$

$$\Omega(z+h,z)=\sum_j \Omega_j(z+h,z)$$

$$\Omega_1=\int_z^{z+h} a(z')dz'$$

$$\Omega_2=\frac{1}{2}\int_z^{z+h} dz_1\int_z^{z_1} dz_2[a(z_1),a(z_2)],$$

и все старшие члены включают коммутаторы. Оператор $\Omega(z+h,z)$ продолжает сохранять свойства гамильтоновой матрицы с $D\neq 0$. Популярный RCWA-метод соответствует приближению

$$\Omega(z+h,z)\approx\Omega_1(z+h,z)\approx a(z+h/2)h$$

и вычисленной точно экспоненте от матрицы. Если нет зависимости от z, приближение точно, так как все коммутаторы равны нулю. В противном случае RCWA-метод является локально точным методом второго порядка. Следующий порядок точности включает производные первого и второго порядков.

Следовательно, адаптивный размер сетки может повысить точность схемы. Точное определение матричной экспоненты e^{Ω} требует очень затратной процедуры диагонализации матрицы Ω и других дополнительных матричных операций. Вдобавок непосредственное вычисление матричной экспоненты может быть численно неустойчивым из-за наличия растущих и убывающих экспонент. Это является также источником проблем, связанных с методом стрельбы, который очень родственен прямому методу матрицы перехода. Как было сказано ранее, решающим явилось осознание того, что единственно необходимым является отношение между J и w. Действительно, если

$$BA=S\Lambda S^{-1}\equiv Sq^2S^{-1},$$

где матрица Λ диагональна, а матрица S нормализована таким образом, что $S^T DS = 1$, и если $J(z) = w(z)\psi(z)$, можно показать, что

$$w(z-d) = (S^{-1})^T \left\{ \left[\frac{1-e^{qd}}{2q} + e^{qd}(q + S^T w(z) S)^{-1} e^{qd} \right]^{-1} - q \right\} S^{-1}. \qquad (2.9)$$

Число операций при диагонализации матрицы обычно в десять раз больше числа операций при умножении матрицы на матрицу. Мы можем воспользоваться другими возможностями. Однако остаются три проблемы, с которыми необходимо справиться:

- локальная точность;
- глобальная численная неустойчивость, вытекающая из неточности численных процедур;
- природная глобальная неустойчивость задачи (или метода).

Как упоминалось ранее, RCWA является методом второго порядка точности, если $a(z)$ зависит от z и обеспечивает абсолютную численную устойчивость благодаря численной диагонализации. Третья проблема является общей для всех схем. Даже если мы можем вычислить e^{Ω} точно, то столкнемся с проблемой неустойчивости. Именно здесь вступает в игру пересчет J через ψ.

При разработке алгоритма глобальная стабильность важнее локальной точности, потому что глобально неустойчивая схема может привести окончательный результат к полному провалу, несмотря на локальную точность. Например, явный метод Рунге—Кутта четвертого порядка точности проигрывает по сравнению с центрированной разностной схемой второго порядка при малом числе точек сетки. В последующем опишем две альтернативы RCWA-методу, основанному на диагонализации матрицы.

Центральная разностная схема

Делим всю среду на N равных сегментов длины h и обозначаем $\psi_n \equiv \psi(nh)$. Значения поля вычисляем в концевых точках, а J - в центрах отрезков (или наоборот). В результате дискретизации получаем:

$$\psi_{n-1} = \psi_n - B_{n-1/2} J_{n-1/2}$$

$$J_{n-1/2} = J_{n+1/2} - A_n \psi_n$$

$$\begin{pmatrix} \psi_{n-1} \\ J_{n-1/2} \end{pmatrix} = \begin{pmatrix} 1 & -B_{n-1/2} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\mathrm{A_n} & 1 \end{pmatrix} \begin{pmatrix} \psi_n \\ J_{n+1/2} \end{pmatrix}$$

$$B_{n+1/2} \equiv p^{-1}(z_{n+1/2}) h$$

$$A_n \equiv A(z_n) / h.$$

Существует простая схема рекурсивного использования этих соотношений, однако она не работает, так как результирующая матрица со временем расходится. Вместо этого следует положить $J_{n+1/2} \equiv w_n \psi_n$, и в силу приведенных выше уравнений имеем

$$w_{n-1} = \left(\left(w_n - A_n \right)^{-1} - B_{n-1/2} \right)^{-1} = B_{n-1/2}^{-1} + B_{n-1/2}^{-1} \left(w_n - A_n - B_{n-1/2}^{-1} \right)^{-1} B_{n-1/2}^{-1}. \quad (2.10)$$

Вторая форма оказывается численно более устойчивой, если w_n является симметричной. Для ТЕ-моды B является диагональной, и обратная к ней матрица вычисляется тривиально. Для ТМ-моды вычисляем p, а B является обратной к p. Для конической или трехмерной дифракции вычисляем B, которая является блочно-диагональной, а затем вычисляем матрицу, обратную к B. Во всех случаях вторая форма оказывается быстрее.

Для эквивалентной формулировки исключаем J, так что

$$p(z + h/2)(\psi(z+h) - \psi(z)) + p(z - h/2)(\psi(z-h) - \psi(z)) = A(z) h^2 \psi(z).$$

Задаем калибровку в виде $\psi(z+h)=W(z)\psi(z)$ и получаем простую рекурсию:

$$W(z-h)=\left[\overline{A}(z)+p^{-}+p^{+}-p^{+}W(z)\right]^{-1}p^{-}, \tag{2.11}$$

где $p^{\pm}\equiv p(z\pm h)/h,\quad \overline{A}\equiv A(z)h.$

Данное упражнение имеет целью показать, что такая схема эквивалентна так называемому LU-разложению на блочно-треугольные матрицы, что будет доказано впоследствии.

Нам также нужен способ более эффективного использования граничных условий. Воспользуемся следующей схемой второго порядка точности:

$$Y_{n-1/2}\approx e^{-\frac{h}{2}a(n-1/4)}Y_n\approx\begin{pmatrix}1+\frac{1}{2}BA & -B\\ -A & 1+\frac{1}{2}AB\end{pmatrix}Y_n$$

$$w_{n-1/2}=-A+\left(1+\frac{1}{2}AB\right)w_n \tag{2.12}$$

$$Y_{1/2}\approx e^{\frac{h}{2}a(n-1/4)}Y_0, J_{1/2}=A+\left(1+\frac{1}{2}AB\right)w_0,$$

$$\begin{aligned}w_0=\left(1+\frac{1}{2}AB\right)^{-1}\left(J_{1/2}-A\right)\approx\left(1-\frac{1}{2}AB\right)\left(J_{1/2}-A\right)\approx\\ \approx -A+\left(1-\frac{1}{2}AB\right)w_{1/2}\end{aligned} \tag{2.13}$$

Игнорируя AB, получаем первый порядок точности, и результирующая матрица w является симметричной, что делает алгоритм быстрее почти в два раза. Однако при учете AB результаты в общем случае оказываются лучше, даже если мы симметризуем w. В пределе

$h \to 0$ уравнение (2.11) можно переписать в виде

$$w(z-h)-w \approx \frac{p}{h}\left[1+hp^{-1}(w-Ah)+h^2 p^{-1}(w-Ah)p^{-1}(w-Ah)\right]-\frac{p}{h}-w$$

$$=-Ah+h(w-Ah)p^{-1}(w-Ah)$$

$$=-Ah+hwp^{-1}w,$$

который приводит к следующему дифференциальному матричному уравнению Риккати для w:

$$\frac{dw}{dz}=-A+wp^{-1}w. \qquad (2.14)$$

В работе [8] явный метод Рунге—Кутта был использован для решения этого уравнения с целью получения коэффициентов отражения. Но, вообще говоря, метод Рунге—Кутта не обладает хорошей устойчивостью в случае нелинейных уравнений. Даже в линейном случае метод центральных разностей может превосходить метод Рунге-Кутта и по скорости, и по окончательной точности. В теории управления используются полуявные методы, которые сводятся к решению так называемого уравнения Сильвестра, также весьма трудоемкому. В работе [9] при решении волнового уравнения использовалось решение матричного уравнения Риккати, похожее на метод сплетающих операторов, который мы опишем ниже, и был задействован очень трудоемкий алгоритм QR разложения.

2.2. Сплетающие операторы

Начиная с уравнения $Y_{n-1}=e^{-\Omega h}Y_n$, воспользуемся расщеплением Стрэнга [10], имеющим второй порядок точности:

$$a \equiv \begin{pmatrix} 0 & 0 \\ A & 0 \end{pmatrix} \quad b \equiv \begin{pmatrix} 0 & B \\ 0 & 0 \end{pmatrix}$$

$$e^{-(a+b)} \approx e^{\frac{1}{2}b} e^{-a} e^{-\frac{1}{2}b}$$

$$= \begin{pmatrix} 1 & -\frac{1}{2}B \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -A & 1 \end{pmatrix} \begin{pmatrix} 1 & -\frac{1}{2}B \\ 0 & 1 \end{pmatrix}$$

$$e^{-(a+b)} \approx e^{-\frac{1}{2}a} e^{-b} e^{-\frac{1}{2}a}$$

$$= \begin{pmatrix} 1 & 0 \\ -\frac{1}{2}A & 1 \end{pmatrix} \begin{pmatrix} 1 & -B \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\frac{1}{2}A & 1 \end{pmatrix}.$$

Следовательно, использование второй формулы дает для w рекурсивные соотношения

$$w = \left(\left(w - A/2 \right)^{-1} - B \right)^{-1} - A/2 = p \left(p + A/2 - w \right)^{-1} p - p - A/2, \quad (2.15)$$

в которых значения и A, и B берутся в центральных точках. Отметим, что данная рекурсия очень похожа на рекурсию центральных разностей.

Полученные соотношения называются также методом «прыжок лягушки» (Верле [11]). Он работает достаточно хорошо по сравнению с такими схемами, как метод Рунге—Кутта, но не так хорошо, как центрально-разностная схема, описанная в предыдущем разделе.

Следует отметить, что для систем с постоянными коэффициентами метод сплетающих операторов почти идентичен центрально-разностной схеме. Они имеют одинаковую долгосрочную устойчивость. Единственное их отличие состоит в использовании граничных условий.

Корректирующий метод сплетающих операторов второго порядка

Когда матрица коэффициентов p является постоянной по z в рассматриваемой области, можно использовать метод конечных разностей четвертого порядка, который является методом Нумерова [12]. В этом методе p заменяется на $p - Ah^2/12$. Это подходит только для TE-мод. Когда же $p(z)$ зависит от z, правильная формула для конечных разностей четвертого порядка становится слишком сложной для работы.

Однако можно попробовать аппроксимировать матричную экспоненту $e^{\Omega h}$ до четвертого порядка, чтобы повысить глобальную численную устойчивость и локальную точность. Под влиянием метода Нумерова, где p было заменено на $p - Ah^2/12$, и вариационно-разностного метода, где p заменяется на $p - Ah^2/6$, мы также проэкспериментируем с этими заменами. Нами обнаружено, что:

- обе подстановки работают хорошо в методе сплетающих операторов;
- обе подстановки хорошо работают в центрально-разностном методе без коррекции граничных условий и со слабой коррекцией граничных условий;
- подстановка $p - Ah^2/12$ работает очень хорошо для двумерных и трехмерных задач.

В итоге мы остановились на подстановке $p - Ah^2/12$, которую назвали методом псевдо-Нумерова.

Положим

$$Q \equiv \begin{pmatrix} 0 & B \\ A & 0 \end{pmatrix} h, \quad Q_a \equiv \begin{pmatrix} 0 & 0 \\ a & 0 \end{pmatrix}, \quad Q_b \equiv \begin{pmatrix} 0 & b \\ 0 & 0 \end{pmatrix},$$

где a и b - неопределенные константы,

$$e^{Q} \approx e^{\frac{1}{2}Q_a} e^{Q_b} e^{\frac{1}{2}Q_a} \approx e^{Q_a+Q_b+\frac{1}{12}[Q_b,[Q_b,Q_a]]+\frac{1}{24}[Q_a,[Q_b,Q_a]]+\ldots},$$

$$[Q_b,Q_a]=\begin{pmatrix} ba & 0 \\ 0 & -ab \end{pmatrix},$$

$$\big[Q_b,[Q_b,Q_a]\big]=\begin{pmatrix} 0 & -2bab \\ 0 & 0 \end{pmatrix},\ \big[Q_a,[Q_b,Q_a]\big]=\begin{pmatrix} 0 & 0 \\ 2aba & 0 \end{pmatrix}.$$

Игнорируя члены более высокого порядка, потребуем, чтобы выполнялось равенство

$$Q=Q_a+Q_b+\frac{1}{6}\big[Q_b,[Q_b,Q_a]\big]+\frac{1}{12}\big[Q_a,[Q_b,Q_a]\big]=$$

$$=\begin{pmatrix} 0 & b-\frac{1}{6}bab \\ a+\frac{1}{12}aba & 0 \end{pmatrix},$$

откуда следует

$$Bh=b-\frac{1}{6}bab,\quad Ah=a+\frac{1}{12}aba.$$

Снова игнорируем члены более высокого порядка при разрешении соотношений относительно a и b, так что

$$b=Bh+\frac{1}{6}BABh^3,\ \ b^{-1}=B^{-1}/h-\frac{1}{6}Ah, \tag{2.16}$$

$$a=Ah-\frac{1}{12}ABAh^3\approx\left(1+\frac{1}{12}ABh^2\right)^{-1}Ah= \tag{2.17}$$

$$=12B^{-1}h^{-1}(12B^{-1}h^{-1}+Ah)^{-1}Ah=12p/h-12p/h(12p/h+Ah)^{-1}12p/h.$$

После проведенных замен получаем алгоритм псевдочетвертого порядка. Он приводит к значительно лучшим результатам, чем просто

метод сплетающих операторов, в случае малого числа слоев. Ниже приведены матричные умножения для первого и последнего членов.

В методе псевдо-Нумерова:

$$b^{-1} = B^{-1} / h - \frac{1}{12} Ah, \tag{2.18}$$

$$a = Ah. \tag{2.19}$$

Чтобы показать, почему это хорошо работает, рассмотрим

$$e^{Q} = \begin{pmatrix} 1 + \frac{1}{2}BA + \frac{1}{24}(BA)^2 & B + \frac{1}{6}BAB \\ A + \frac{1}{6}ABA & 1 + \frac{1}{2}AB + \frac{1}{24}(AB)^2 \end{pmatrix} + \ldots$$

$$e^{Q_a/2} e^{Q_b} e^{Q_a/2} = \begin{pmatrix} 1 + \frac{1}{2}ba & b \\ a + \frac{1}{4}aba & 1 + \frac{1}{2}ab \end{pmatrix}.$$

Если $a = A$ и $b = B + \frac{1}{12}BAB$, то

$$e^{Q_a/2} e^{Q_b} e^{Q_a/2} = \begin{pmatrix} 1 + \frac{1}{2}BA + \frac{1}{24}(BA)^2 & B + \frac{1}{12}BAB \\ A + \frac{1}{4}ABA + \ldots & 1 + \frac{1}{2}AB + \frac{1}{24}(AB)^2 \end{pmatrix}.$$

Диагональные члены совпадают с матричной экспонентой до четвертого порядка. С другой стороны, для систем с постоянными коэффициентами Q_a и Q_b различаются. Коэффициенты во внедиагональных членах также заменяются средними значениями. В результате получаем метод сплетающих операторов псевдочетвертого порядка. Вот почему он проявляет себя так хорошо по сравнению с обычными методами псевдочетвертого порядка. Что еще важно в нашей формулировке, так это простота уравнения (2.18). Оказывается, что это

очень хорошая схема дискретизации даже по сравнению с методом центральных разностей, и мы выбираем ее в качестве конечно-разностной схемы.

Наша схема имеет еще два преимущества по сравнению с методом центральных разностей:

- требуется намного меньше слоев, чтобы достичь того же результата;
- обладает меньшей чувствительностью к замене матриц A и B.

Что еще важнее, так это принадлежность метода сплетающих операторов к классу так называемых симплектических интеграторов, сохраняющих структуру [13]. Например, метод сплетающих операторов сохраняет симметричность матрицы w. С другой стороны, известно, что классический метод Рунге—Кутта симплектическим интегратором не является. Существуют симплектические интеграторы более высокого порядка, однако опыт показывает, что они слишком трудоемкие.

Алгоритм Ньюмарка

Так называемый алгоритм Ньюмарка [14] широко использовался при численном интегрировании в структурной механике. Алгоритм интегрирует динамическую систему следующим образом:

$$q_{k+1} = q_k + h\dot{q}_k + \frac{1}{2}h^2\left[(1-2\beta)a_k + 2\beta a_{k+1}\right],$$

$$\dot{q}_{k+1} = \dot{q}_k + h\left[(1-\gamma)a_k + \gamma a_{k+1}\right],$$

$$a_k = Bf(q_k),$$

где $M = p = B^{-1}$. Для линейных систем $f(q) = Aq$ получаем

$$\begin{pmatrix} 1-\beta h^2 BA & 0 \\ -\gamma Ah & 1 \end{pmatrix}\begin{pmatrix} q_{k+1} \\ J_{k+1} \end{pmatrix}=\begin{pmatrix} 1+\frac{1}{2}h^2(1-2\beta) & Bh \\ (1-\gamma)Ah & 1 \end{pmatrix}\begin{pmatrix} q_k \\ J_k \end{pmatrix}.$$

Для матриц с постоянными коэффициентами получаем

$$\begin{pmatrix} q_{k+1} \\ J_{k+1} \end{pmatrix}=\begin{pmatrix} 1 & 0 \\ \gamma Ah & 1 \end{pmatrix}\begin{pmatrix} 1+\frac{h^2}{2}(1-\beta h^2 BA)^{-1}BA & (1-\beta h^2 BA)^{-1}Bh \\ (1-\gamma)Ah & 1 \end{pmatrix}\begin{pmatrix} q_k \\ J_k \end{pmatrix}.$$

Далее, если положить $\gamma = 1/2$, получаем

$$\begin{pmatrix} q_{k+1} \\ J_{k+1} \end{pmatrix}=\begin{pmatrix} 1 & 0 \\ \frac{1}{2}Ah & 1 \end{pmatrix}\begin{pmatrix} 1 & (1-\beta h^2 BA)^{-1}Bh \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ \frac{1}{2}Ah & 1 \end{pmatrix}\begin{pmatrix} q_k \\ J_k \end{pmatrix}$$

что является симплектическим преобразованием .

Наконец, если $\beta = 1/12$, получаем

$$(1-\beta h^2 BA)^{-1}Bh=\left(B^{-1}/h-\frac{1}{12}hA\right)^{-1},$$

т.е. метод псевдо-Нумерова. Таким образом, мы показали, что наш метод является частным случаем алгоритма Ньюмарка.

Так как мы заметили, что выбор параметра $\beta = 1/6$ приводит к уменьшению по сравнению с методом центральных разностей, полагаем, что выбор для параметра β значения 1/12 является более оптимальным, чем обычный выбор 1/6 и ¼, и предлагаем следующий модифицированный алгоритм Ньюмарка:

$$\dot{q}_{1/2} = q_k + (1-\gamma)ha_k,$$

$$q_{k+1} = q_k + h\left[\dot{q}_{1/2} + \beta h(a_{k+1} - a_k)\right] = q_k + h\dot{q}_k + \beta h^2 a_{k+1} + h^2(1-\gamma-\beta)a_k,$$
$$\dot{q}_{k+1} = \dot{q}_{1/2} + \gamma h a_{k+1} = \dot{q}_k + h\left[(1-\gamma)a_k + \gamma a_{k+1}\right].$$

Алгоритм точного четвертого порядка

Для дифференциального уравнения следующего вида:

$$y'' = Ay,$$

алгоритм четвертого порядка имеет вид:

$$y_{j+1} = e^{\frac{h^2}{12}A'} e^{hA+\frac{h^3}{24}A''} e^{-\frac{h^2}{12}A'} y_j$$

$$\approx e^{\frac{h}{12}(A_{j+1}-A_j)} e^{\frac{h}{6}(A_j+4A_{j+1/2}+A_{j+1})} e^{-\frac{h^2}{12}(A_{j+1}-A_j)} y_j .$$

Для ТЕ-мод он может быть использован без особых затруднений. Но в общем случае требует дополнительных усилий.

2.3. Блочно-треугольный UL(LU)-алгоритм

Предположим, что имеем матрицу следующего вида

$$A = \begin{pmatrix} a_1 & b_1 & & & & & \\ c_2 & a_2 & b_2 & & & & \\ & c_3 & a_3 & b_3 & & & \\ & & . & . & . & & \\ & & & . & . & . & \\ & & & & c_{n-1} & a_{n-1} & b_{n-1} \\ & & & & & c_n & a_n \end{pmatrix}$$ с правой частью $$Y = \begin{pmatrix} y_1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

Разложим матрицу A следующим образом $A = UL,$ где

$$U = \begin{pmatrix} 1 & u_1 & & & & & \\ 0 & 1 & u_2 & & & & \\ & 0 & 1 & u_3 & & & \\ & & 0 & 1 & & & \\ & & & 0 & 1 & & \\ & & & & 0 & 1 & u_{n-1} \\ & & & & & 0 & 1 \end{pmatrix}$$

и

$$L = \begin{pmatrix} d_1 & & & & & \\ c_2 & d_2 & & & & \\ & c_3 & d_3 & & & \\ & & . & . & & \\ & & & . & . & \\ & & & & c_{n-1} & d_{n-1} \\ & & & & & c_n & d_n \end{pmatrix}.$$

Тогда мы получаем соотношения:

$d_n = a_n$

$u_{i+1} = b_i d_{i+1}^{-1}$

$d_i = a_i - b_i d_{i+1}^{-1} c_{i+1}$

Положим $w_i = d_i^{-1} c_i$, тогда выполняются соотношения:

$$w_i = (a_i - b_i w_{i+1})^{-1} c_i \tag{2.20}$$

и в итоге

$$L = diag(d) \begin{pmatrix} 1 & & & & & \\ w_2 & 1 & & & & \\ & w_3 & 1 & & & \\ & & . & . & & \\ & & & . & . & \\ & & & & w_{n-1} & 1 \\ & & & & & w_n & 1 \end{pmatrix}.$$

Преимуществом такого подхода является то, что нам нужно решить только одну систему $Lx = y$ вследствие специальной структуры y, т.к. $U^{-1} y = y$. Следовательно, нам нужно лишь отношение между x_1 и x_2, и нет

нужды хранить w_i и d_i.

Сравнивая уравнения (2.11) и (2.20), видим, что масштабирующий алгоритм, описанный в разделе 2.1, эквивалентен UL-разложению.

2.4. Численные процедуры

Начиная с подложки, дигонализуем BA так, чтобы $BA = S\Lambda S^{-1}$, граничные условия допускают только прошедшие или затухающие волны, так что

$$w = pSpS^{-1} = (S^{-1})^T qS^{-1}. \tag{2.21}$$

Поскольку подложка изготавливается из однородного материала, процесс диагонализации тривиален и S диагональная матрица.

Итерируем по слоям, а если слой однороден или решетка бинарная, т.е. нет зависимости от z, и число слоев должно превысить некий порог (скажем 5), то диагонализуем BA и используем уравнение (2.9), чтобы получить w на следующей границе. В противном случае нарезаем материал на N слоев (Nh - толщина покрытия) и используем любой алгоритм решения начальной задачи, чтобы получить ток при $(N-1/2)h$ из уравнения (2.12). А именно пользуемся соотношением (2.10) рекурсивно до верхнего слоя, чтобы получить связь между $J_{1/2}$ и ψ_0, например уравнение (2.13).

При альтернативном подходе, используя метод сплетающих операторов, все покрытие можно разрезать на одинаковые и неодинаковые слои и воспользоваться уравнением (2.15) для итераций по всем слоям. Это можно рассматривать в качестве простого замещения RCWA-метода, процедуры диагонализации.

В общем, может быть использована любая схема дискретизации. Хорошими кандидатами являются так называемые многошаговые

формулы обратных разностей, используемые при решении жестких уравнений. Но они обычно требуют большого числа матричных процедур и не симметричны при движении вперед и назад. Для ТЕ-мод можно использовать метод Нумерова четвертого порядка [12].

Вычислить рассеянное поле с помощью соотношения (2.7). Для метода сплетающих операторов используем $p - A/12$ вместо p, и шаги практически совпадают с методом центральных разностей. При этом больше не требуется корректировать граничные условия.

Если система симметрична, можно добавить несколько замечаний. При вычислении $pz^{-1}p$, единственно трудоемкой матричной процедуры, можно разложить $z = LDL^T$, где L – нижняя треугольная матрица; D – диагональная, возможно, с поддиагональными прямоугольными блоками:

$$pz^{-1}p = \left(L^{-1}P\right)^T D^{-1}\left(L^{-1}P\right).$$

Здесь имеется только одно матричное разложение, одно полуматричное решение $Q = L^{-1}P,$ одна практически ничтожная процедура $D^{-1}Q$ и умножение матрицы на матрицу, поскольку мы знаем, что $Q^T D^{-1}Q$ симметрична.

2.5. Сравнение с другими подходами

Описанный метод обладает несколькими преимуществами по сравнению с другими.

- Преимущество в скорости. Нет необходимости диагонализовать матрицы, как в сравниваемых подходах. Диагонализация матриц – самая трудоемкая процедура популярных решеточных алгоритмов. Она требует в десять раз больше времени, чем перемножение матриц или обращение матриц. Поэтому токовый подход в несколько раз быстрее

сравниваемых алгоритмов при том же числе точек сетки (слоев). Для покрытий с комплексной диэлектрической проницаемостью он приблизительно в 7-8 раз быстрее RCWA метода, даже с использованием быстрой процедуры диагонализации матриц.

- Легкость (простота) реализации. Это существенно прямой метод, похожий на метод прогонки для ленточных матриц. Фактически, если мы используем неравномерные сетки, не заботясь о границах, все сводится к проблеме с ленточной матрицей. В отличие от других методов, нет необходимости в итерационном решении. Более того, предложенная нами реализация не требует даже алгоритма решения систем с ленточными матрицами.
- Гладкий профиль. Дискретизация требует гладкости профиля по вертикали. Так как наш метод допускает большое число слоев, требования на гладкость профиля могут быть менее обременительными.

Тема 3. Вариационная формулировка рассеяния плоских электромагнитных волн на одномерных дифракционных решетках

3.1. Введение

Дифракционная оптика - энергично растущая технология, где средства полупроводниковой промышленности используются для изготовления оптических приборов со сложными структурными свойствами. Исследуются и создаются миниатюрные оптические компоненты, в которых реализуются весьма специфические особенности. Основанные на использовании дифракционных эффектов, такие устройства обладают функциями, которые невозможны при использовании обычных оптических устройств. Преимущества таких устройств заключаются в их малых габаритах, весе, дешевизне при массовом производстве. В настоящее время дифракционные оптические устройства активно используются при производстве миниатюрных и проекционных дисплеев, для создания антибликовых и просветляющих покрытий, высококачественных зеркал в оптических коммуникационных устройствах.

Рассмотрим достижения в области оптимального проектирования дифракционных оптических элементов. Оптимальный дизайн миниатюрных дифракционных оптических элементов гораздо более труден, чем проектирование традиционных отражающих или преломляющих оптических (макро-) элементов. Дополнительные сложности возникают в связи с тем, что поведение электромагнитных волн при взаимодействии со столь малыми объектами (сравнимыми с длиной волны излучения) весьма необычно, и зачастую может быть осознано лишь в результате численного моделирования - решения уравнений Максвелла, в то время как традиционные оптические устройства вполне адекватно описываются относительно простыми алгоритмами трассировки лучей.

Методы оптимизации, разработанные для обычных оптических устройств, таким образом, неприменимы для дифракционной оптики.

Численное решение уравнений Максвелла представляет собой весьма сложную задачу даже для простых геометрических структур дифракционных элементов, поэтому в настоящее время хорошо удается моделировать лишь не очень сложные элементы – многослойные периодические структуры.

Создание новых оптических элементов в случае более сложной структуры решетки приводит к необходимости построения соответствующих математических моделей и численных исследований для решения уравнений электромагнитных векторных полей.

Одна из наиболее простых и часто используемых геометрических конфигураций для создания дифракционных оптических структур - периодическая дифракционная решетка, протравленная на поверхности оптической подлжки (рис.1). Профиль такого типа чаще всего создается с помощью фотолитографии, иногда дополняемой нанесением на поверхность дополнительных слоев нужных материалов. Такие процессы широко используются в полупроводниковой промышленности.

В большинстве приложений дифракционной оптики используются в основном узкополосные диапазоны, поэтому математическое моделирование основано на представлении электромагнитного излучения в виде ряда плоских монохроматических волн.

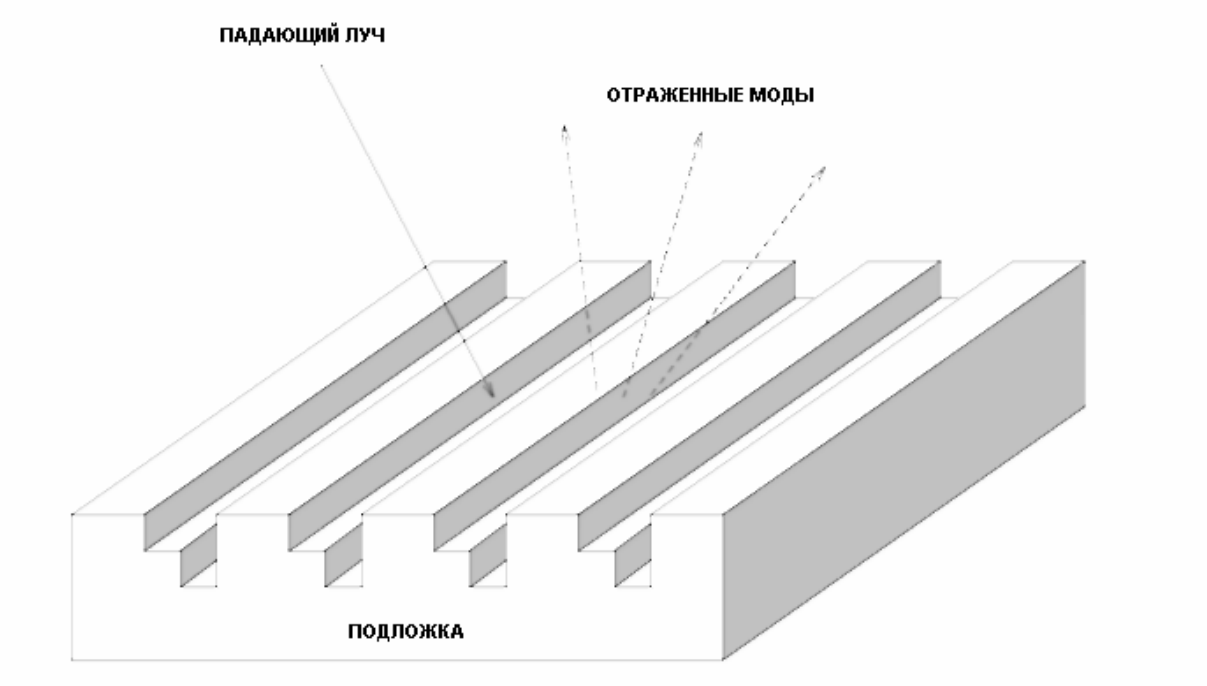

Рис. 1. Падение луча света на дифракционную решетку

Рассеяние плоских монохроматических волн на бесконечных периодических структурах - классическая проблема, изучение которой активно велось во времена Рэлея, Флоке и Блоха. Фундаментальная особенность проблемы состоит в том, что в однородных областях электромагнитное поле может быть представлено в виде бесконечной суммы плоских волн. В самой простой форме, например, для отраженного поля, зависящего от двух переменных, разложение может быть записано в виде

$$u(x_1, x_2) = \sum_{n=-\infty}^{\infty} a_n e^{i(nx_1 + \beta_n x_2)}, \qquad (3.1)$$

где a_n - неизвестные коэффициенты. Такую форму записи называют разложением Рэлея. В случае среды с действительным индексом преломления $k > 0$ коэффициенты β_n в сумме (3.1) определяются выражениями:

$$\beta_n = \begin{cases} \sqrt{k^2 - n^2} & if \quad k \geq |n|, \\ i\sqrt{n^2 - k^2} & if \quad k < |n|. \end{cases}$$

Так как β_n действительное число, по крайней мере, для некоторого ограниченного набора индексов n, из (3.1) следует, что только конечное число плоских волн из всей суммы распространяется в дальнюю область, остальные же, исчезающие волны, ниспадают по экспоненте при $x_2 \to +\infty$. Число распространяющихся мод и направления их распространения определяются частотой падающей волны, индексом диэлектрической проницаемости материала и периодом (размером ячейки) структуры. В литературе отношение энергии данной распространяющейся моды к энергии падающей волны называют коэффициентом эффективности моды. С технической точки зрения, ключевая особенность любого дифракционного элемента - эффективность каждой распространяющейся моды.

Хорошее введение в проблему рассеяния на периодических дифракционных структурах можно найти, например, в [15, 16].

Специфическая проблема, к обсуждению которой мы переходим, это проблема проектирования такой периодической структуры, что излучаемые ею (отраженные либо прошедшие) моды обладают заданной фазой и/или интенсивностью излучения. Рассмотрение базируется на анализе модели распространения плоской монохроматической волны. Ниже кратко рассмотрим вариационную формулировку этой модели: уравнение Гельмгольца для периодических структур. Затем перейдем к обсуждению аппроксимации вариационной проблемы с помощью метода конечных элементов и обсудим возникающие проблемы численной реализации. Далее переформулируем задачу, сведя проблему создания оптимальной конструкции к задаче минимизации, и опишем два

существенно различных случая, которые возникают в зависимости от предположений о классе допустимых решений. В первом случае в качестве допустимых решений берется множество профилей – это «мягкая» формулировка задачи. Во втором случае класс допустимых профилей резко ограничивается, приводя к проблеме «дизайна профиля». Конечно, в обоих случаях при формулировке задачи мы должны принимать во внимание производственные возможности и ограничения.

3.2. Прямая задача дифракции.

Прямая задача дифракции заключается в том, чтобы при заданной падающей волне произвольной поляризации описать такие характеристики, как энергия и поляризация для всех типов волн, прошедших достаточно большое расстояние от дифракционной решетки. Теоретические исследования этого вопроса имеют длительную историю, особенно в случае двумерных периодических структур.

Но только в течение последних 20 лет в двумерном случае были получены результаты о существовании, единственности и регулярности решений для негладких структур, а также для всех веществ, используемых на практике. Эти результаты основаны на вариационных формулировках задач в ограниченной периодической ячейке. Вариационный подход был разработан Неделеком и Старлингом (см. [17]).

Вариационные методы применялись к биперiодическим структурам в работах Аббоуда, Добсона, Бао и др. (см. [17]). При определенных ограничениях на параметры вещества была доказана корректность соответствующей постановки задачи. Указанные результаты в работах Шмидта и др. (см. [18]) были перенесены на трехмерные задачи дифракции. Указанный подход позволяет рассматривать более общие ситуации, соответствующие изотропному случаю, а также исследовать анизотропные структуры — так называемые скрещенные анизотропные

решетки. В работе [18] предлагается модифицированный вариант такого подхода, позволяющий получить результаты о разрешимости при предположениях, которые выполняются в соответствующих приложениях на практике.

Сформулируем задачу дифракции в простой вариационной постановке. Более подробно постановка задачи дана, например, в работах Ашду и Добсона (см. [17]). Все рассуждения, приводимые ниже, основаны на том предположении, что изучается бесконечная периодическая структура. Важно отметить, что решение задачи для бесконечной периодической системы может быть сведено к решению задачи для конечномерной периодической структуры с использованием методов Криксмана (см. [17]).

Ограничимся наиболее простым с точки зрения геометрии случаем, когда дифракционная структура постоянна в одном направлении, как представлено на рис. 1.

Обозначим через $x=(x_1,x_2,x_3)$ точку в R^3, и для удобства предположим, что стенки «пеньков» параллельны оси $\vec{x}_3$.Тогда коэффициент диэлектрической проницаемости $e(x)$ является функцией только двух переменных (x_1,x_2). В силу периодичности

$e(x_1+nL,x_2)=e(x_1,x_2)$ для всех (x_1,x_2),

где L - период; n - любое целое число. Не ограничивая общности, будем считать, что период L равен 2π.

Тогда прямая задача дифракции состоит в том, чтобы решить систему гармонических по времени (зависимость e^{-iwt} от времени) уравнений Максвелла:

$$\nabla\times E-iw\mu H=0, \qquad (3.2)$$

$$\nabla \times H - iweE = 0, \tag{3.3}$$

где E и H - векторы электрической и магнитной проницаемости соответственно при падении электромагнитной волны на дифракционную решетку сверху. В рассматриваемом случае, в зависимости от направления и поляризации падающей плоскопараллельной волны, возможны три существенно различные ситуации:

- TE (transverse electric) - поляризация: волновой вектор падающей волны ортогонален к оси x_3 и вектор напряженности электрического поля E параллелен оси x_3;
- TM (transverse magnetic) - поляризация: волновой вектор падающей волны ортогонален к оси x_3 и вектор напряженности магнитного поля H параллелен оси x_3;
- Коническая дифракция: волновой вектор падающей волны не ортогонален к оси x_3.

Эти проблемы, по существу, перечислены в порядке увеличивающейся трудности. В первом случае уравнения Максвелла сводятся к простому скалярному уравнению Гельмгольца для вектора электрического поля E. Во втором случае уравнения Максвелла также сводятся к простой скалярной модели (на сей раз для вектора H), однако задача перестает быть непрерывной. В третьем случае необходимо сохранение полной векторной модели, хотя геометрия задачи все еще «двумерна». Все три представленных случая, так же, как и случай полной системы уравнений Максвелла в трехмерном пространстве, интенсивно изучались в технической и математической литературе. Для простоты рассмотрим здесь лишь первый, самый простой случай TE-поляризации.

Пусть $E = u\vec{x}_3$, где $u = u(x_1, x_2)$ - скалярная функция. Тогда электрическое поле E удовлетворяют уравнению Гельмгольца:

$$\left(\Delta + k^2\right)u = 0 \text{ в } R^2, \tag{3.4}$$

где k - показатель преломления: $k^2 = \omega^2 e\mu$. Пусть $k_1^2 = \omega^2 e_1\mu$ и $k_2^2 = \omega^2 e_2\mu$, где ω - круговая частота, μ является магнитной проницаемостью (которая предполагается постоянной); и e_1 и e_2 - постоянные коэффициенты диэлектрической проницаемости в окружающей среде (воздух) и подложке соответственно. Константы k_1, k_2 могут быть, вообще говоря, комплексными (мнимая часть, отличная от нуля, соответствует поглощающему материалу). Полагаем, что $\mathrm{Im}\, k_1 = 0$, т. е. окружающая среда не является поглощающей.

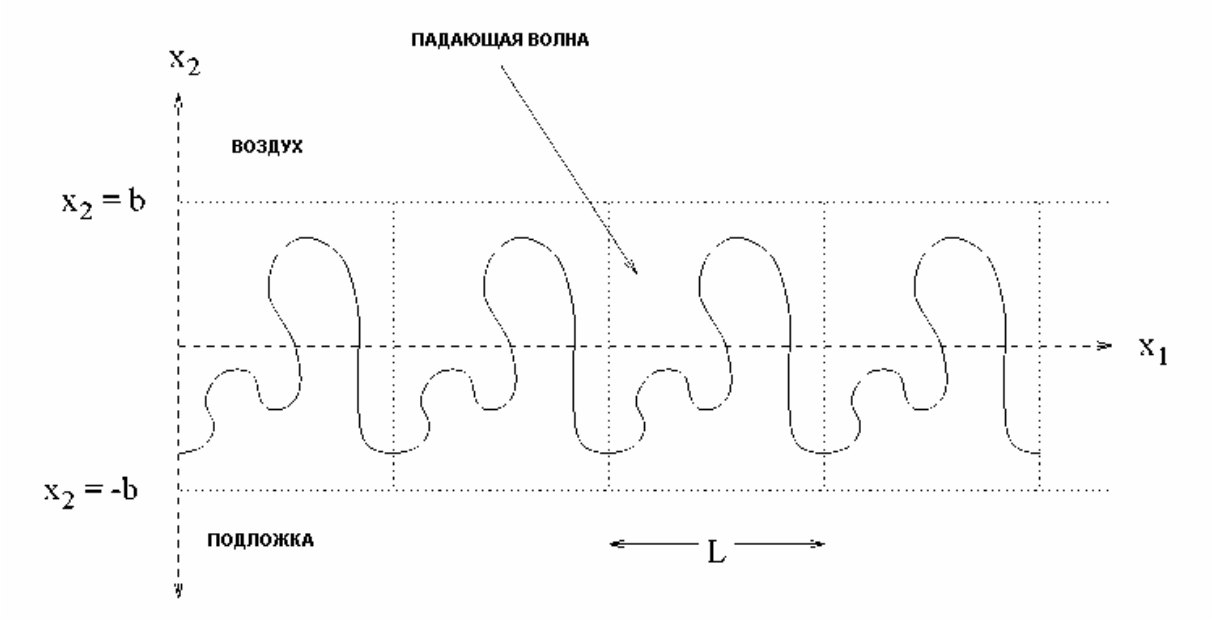

Рис. 2. Геометрия задачи. Падение ТЕ-поляризованной волны

Путь S - простая кривая в R^2 периодична с периодом 2π вдоль оси x_1 и ограничена по оси x_2. Кривая S описывает границу между окружающей средой и подложкой (рис. 2). Профиль дифракционной решетки не обязательно должен быть гладким. Для удобства выкладок задача перемасштабирована так, что период решетки равен $L = 2\pi$. Определим индекс преломления $k(x)$ в R^2 как

$$k(x)=\begin{cases}k_1, & \text{в случае, когда } x \text{ лежит над } S,\\ k_2, & \text{в случае, когда } x \text{ лежит под } S.\end{cases}$$

Требуется решить уравнение Гельмгольца (3.4) в случае, когда плоскопараллельная волна $u_* = e^{i\alpha x_1 - i\beta_1 x_2}$ падает на границу раздела из точки $(x_2 = +\infty)$.

Здесь $\alpha = k_1 \sin\theta, \quad \beta_1 = k_1 \cos\theta,$ и $-\frac{\pi}{2} < \theta < \frac{\pi}{2}$ угол падения волны. Непосредственное решение уравнения (3.4) в неограниченной области R^2 затруднительно без дополнительных предположений о непрерывности кривой S, которые облегчают переход к рассмотрению интегрального уравнения. Наш подход состоит в том, чтобы вместо этого сформулировать задачу в вариационной постановке. Для этого необходимо перейти от решения уравнения Гельмгольца во всем пространстве R^2 к эквивалентной задаче на ограниченном множестве.

Нас интересуют квазипериодические решения, т. е. решения u такие, что функции $u_\alpha = ue^{-i\alpha x_1}$ являются 2π-периодичными. Легко заметить, что если u удовлетворяет условию (3.4) , то u_α удовлетворяет условию

$$\left(\Delta_\alpha + k^2\right)u_\alpha = 0 \text{ в } R^2, \tag{3.5}$$

где оператор Δ_α определен как

$$\Delta_\alpha = \Delta + 2i\alpha\partial_1 - |\alpha|^2.$$

Поскольку далее будем рассматривать только уравнение (3.5), то опустим индекс α. Так как u и k теперь 2π- периодические функции аргумента x_1, исходная задача сведена к решению уравнения (3.5) с периодическими граничными условиями на x_1. Аналогично рассмотрим уравнение (3.5) на фактор-пространстве $Q = R^2 / \{2\pi Z, 0\}$, где

$Z=\{0,\pm1,\pm2,...\}$. Пусть $b>\max\{|x_2|:(x_1,x_2)\in S\}$. Определим периодическую полосу $\Omega=\{x\in Q:-b<x_2<b\}$ и две бесконечных области $\Omega_1=\{x\in Q:x_2>b\}$ и $\Omega_2=\{x\in Q:x_2<-b\}$ выше и ниже Ω соответственно. Определим границы $\Gamma_1=\partial\Omega_1$ и $\Gamma_2=\partial\Omega_2$

Мы хотим найти решение u в области Ω. Это требует соответствующих граничных условий Γ_1 и Γ_2, которые выведем позже. Воспользуемся нелокальными «точными» граничными операторами. Прежде всего разложим u в ряд Фурье:

$$u(x_1,x_2)=\sum_{n\in Z}u_n(x_2)e^{inx_1}, \tag{3.6}$$

где $u_n(x_2)=\frac{1}{2\pi}\int_0^{2\pi}u(x_1,x_2)e^{-inx_1}dx_1$. Определим для $j=1,2$ коэффициенты

$$\beta_j^n(\alpha)=e^{i\gamma_j/2}\left|\varpi^2k_j^2-(n+\alpha)^2\right|^{1/2},n\in Z,$$

где

$$\gamma_j=\arg\left(\varpi^2k_j^2-(n+\alpha)^2\right),0\le\gamma_j<2\pi.$$

Предположим далее, что $k_j^2\neq(n+\alpha)^2$ для всех $n\in Z,j=1,2.$ Это условие исключает «резонансные» случаи, когда волны могут распространяться вдоль оси x_1, и гарантирует, что фундаментальное решение для (3.5) существует внутри Ω_1 и Ω_2 Тогда, поскольку фундаментальное решение [13] известно, u может быть разложено в сумму плоских волн:

$$u\big|_{\Omega_j}=\sum_{n\in Z}a_j^ne^{\pm i\beta_j^n(\alpha)x_2+inx_1},j=1,2, \tag{3.7}$$

где a_j^n - комплексные величины. Учет условия излучения на бесконечности

в уравнении (3.7) гарантирует, что u состоит из ограниченных, исходящих плоских волн в Ω_1 и Ω_2 и плюс падающей волны u_* в Ω_2. Сравнение уравнений (3.6) и (3.7) приводит к явному выражению для $u\big|_{\Omega_j}$ через коэффициенты Фурье на границе $u\big|_{\Gamma_j}$:

$$u_n(x_2)=\begin{cases} u_n(b)e^{i\beta_1^n(\alpha)(x_2-b)}, & n\neq 0, \quad \text{в } \Omega_1, \\ u_0(b)e^{i\beta_1(x_2-b)}+e^{-i\beta_1 x_2}-e^{i\beta_1(x_2-2b)}, & n=0, \quad \text{в } \Omega_1, \\ u_n(-b)e^{-i\beta_2^n(\alpha)(x_2+b)}, & \text{в } \Omega_2. \end{cases}$$

Это выражение может быть продифференцировано по x_2. Суммирование по n дает:

$$\left.\frac{\partial u}{\partial x_2}\right|_{\Gamma_1}=\sum_{n\in Z} i\beta_1^n(\alpha)u^n(b)e^{inx_1}-2i\beta_1 e^{-i\beta_1 b},$$

$$\left.\frac{\partial u}{\partial x_2}\right|_{\Gamma_2}=-\sum_{n\in Z} i\beta_2^n(\alpha)u^n(-b)e^{inx_1}.$$

Определим оператор T_j^α

$$\left(T_j^\alpha f\right)(x_1)=(-1)^{j+1}\sum_{n\in Z} i\beta_j^n(\alpha)f_n e^{inx_1}, j=1,2, \tag{3.8}$$

где f_n- коэффициенты Фурье функции f. Так как коэффициенты β_j^n растут примерно как $|n|$, то можно показать, что каждый оператор $T_j^\alpha: H^{\frac{1}{2}}(\Gamma_j)\to H^{-\frac{1}{2}}(\Gamma_j)$ является непрерывным. Имеем

$$T_j^\alpha\left(u\big|_{\Gamma_j}\right)=\left.\frac{\partial u}{\partial x_2}\right|_{\Gamma_j}, j=1,2.$$

Другими словами, T_j^α является отображением Дирихле—Неймана. Операторы T_j^α определяют транспарентные граничные условия на $\partial\Omega$.

Задача дифракции теперь может быть переформулирована следующим образом: найти $u \in H^1(\Omega)$ такую, что

$$\left(\Delta_\alpha + k^2\right)u = 0 \text{ в } \Omega, \tag{3.9}$$

$$\left(T_1^\alpha - \frac{\partial}{\partial x_2}\right)u = 2i\beta_1 e^{-i\beta_1 b} \text{ на } \Gamma_1, \tag{3.10}$$

$$\left(T_2^\alpha - \frac{\partial}{\partial x_2}\right)u = 0 \text{ на } \Gamma_2. \tag{3.11}$$

Заметим, что условия (3.10), (3.11) уже включают «условие исходящих волн» по построению операторов T_j^α. Задача в формулировке (3.9)-(3.11) допускает вариационную постановку:

$u \in H^1(\Omega)$ удовлетворяет условию

$$\int_{\Omega_0} \nabla u \cdot \nabla \vec{v} - \int_{\Omega_0} \left(\varpi^2 k^2 - \alpha^2\right) u\vec{v} - 2i\alpha \int_{\Omega_0} \left(\partial_1 u\right)\vec{v} -$$

$$\int_{\Gamma_1} \left(T_1^\alpha u\right)\vec{v} - \int_{\Gamma_2} \left(T_2^\alpha u\right)\vec{v} = -\int_{\Gamma_1} 2i\beta_1 e^{-i\beta_1 b}\vec{v}$$

для всех $v \in H^1(\Omega)$. Здесь $\int_{\Gamma_j}$ представляет собой спаривание пространств $H^{-\frac{1}{2}}(\Gamma_j)$ и $H^{\frac{1}{2}}(\Gamma_j)$. Можно показать, что (3.9)-(3.11) допускает единственное слабое решение $u \in H^2(\Omega)$ для всех достаточно малых $k \in L^\infty$, и, на самом деле, для всех k, кроме дискретного множества k_1, k_2.

Случай *TM*-поляризации во многом аналогичен рассмотренному для *TE*-поляризации. Предполагая, что поля и геометрия решетки инвариантны вдоль x_3, вектор магнитного поля H направлен вдоль x_3, $H = u\vec{x}_3$ где $u(x_1, x_2)$ - скалярная функция. Уравнения Максвелла (3.2), (3.3)

трансформируются в систему уравнений

$$\nabla_\alpha\left(\frac{1}{k^2}\nabla_\alpha u\right)+u=0 \text{ в } R^2,$$

где операторы ∇_α определяются как $\nabla+i(\alpha,0)$. Так как k – фиксированная константа в $\Omega_j\ (j=1,2)$, можно получить граничные условия для u на Γ_j, подобно тому, как это было сделано выше, что приводит к соответствующей слабой формулировке.

3.3. Ослабленный метод оптимального проектирования

Итак, дана периодическая кривая $S\subset\Omega$, задающая профиль дифракционной решетки. Вещество над кривой S имеет показатель преломления k_1, и материал под кривой S имеет показатель преломления k_2, Для явного описания функции показателя преломления в зависимости от S определим

$$a_s(x)=\begin{cases}k_1^2, & \text{в случае, когда } x \text{ лежит над } S,\\ k_2^2, & \text{в случае, когда } x \text{ лежит под } S.\end{cases}$$

Далее рассмотрим задачу:

$$(\Delta_\alpha+a_S)u=0 \text{ в } \Omega, \tag{3.12}$$

$$\left(T_1^\alpha-\frac{\partial}{\partial x_2}\right)u=2i\beta_1 e^{-i\beta_1 b} \text{ на } \Gamma_1, \tag{3.13}$$

$$\left(T_2^\alpha-\frac{\partial}{\partial x_2}\right)u=0 \text{ на } \Gamma_2. \tag{3.14}$$

Предположим, что материалы, периодичность структуры и частота падающей волны фиксированы. Тогда имеется лишь конечное число рассеянных мод, и каждая из них соответствует такому показателю преломления m, для которого постоянная распространения волны β_j^n

является действительной переменной. Определим два множества индексов рассеянных мод:

$$\Lambda_j = \left\{ n \text{ целое} : \beta_j^n \text{ действительное} \right\}, j = 1,2.$$

Набор Λ_1 содержит индексы отраженных мод; Λ_2 соответствует прошедшим модам. Коэффициенты каждой отраженной моды определяются коэффициентами Фурье следа u на границе Γ_1:

$$\begin{aligned} r_n &= u_n(b) e^{-i\beta_1 b}, \quad \text{для } n \neq 0, n \in \Lambda_1, \\ r_0 &= u_0(b) e^{-i\beta_1 b} - const, \quad \text{для } n = 0. \end{aligned} \tag{3.15}$$

Точно так же коэффициенты прошедших мод могут быть записаны:

$$t_m = u_m(-b) e^{-i\beta_2 b}, m \in \Lambda_2. \tag{3.16}$$

Записывая коэффициенты отражения и пропускания в виде векторов

$$r = (r_n)_{n \in \Lambda_1}, \quad t = (t_m)_{m \in \Lambda_2},$$

обозначим пару $(r,t) = F$. Коэффициенты r_n и t_m и, следовательно, F являются функциями пограничного профиля S. Обозначим эту зависимость через $F(a_S)$. Общая проблема оптимизации конструкции состоит в том, чтобы найти такой профиль S, который порождает зависимость $F(a_S)$, как можно более близкую к некоторой заданной функции q. Рассматривая такую задачу о близости $F(a_S)$ к q с точки зрения метода наименьших квадратов, приходим к задаче:

$$\min_{a_S \in A} J(a_S) = \frac{1}{2} \left\| F(a_S) - q \right\|_2^2.$$

Выбор допустимого множества коэффициентов A весьма важен. Есть два правила, которые позволяют получить хорошо обусловленную задачу оптимизации. Согласно первому необходимо выбирать

относительно небольшое множество допустимых значений, компактное по отношению к топологии, определяемой отображением $J(a_S)$ и, таким образом, гарантируя существование решения. Правда, такой подход чреват тем, что появляется возможность введения искусственных ограничений, которые могут привести не к глобальному, а локальному оптимуму.

Другой подход: для начала ввести большой класс допустимых кривых и «ослабить» задачу: увеличив допустимый набор, включить соответствующие «смеси» материалов. Это может быть выполнено следующим образом. Рассмотрим множество всех непрерывных простых кривых $S \subset \Omega$ в качестве допустимых кривых. Обозначим этот набор профилей через S. Определим множество допустимых коэффициентов:

$$\tilde{A} = \{a_S : S \in \mathrm{S}\}.$$

Найдем замыкание $\tilde{A}$ относительно функционала J. Рассмотрим множество:

$$A = \left\{a = k_2^2\gamma + k_1^2(1-\gamma) : \gamma \in L^\infty(\Omega), 0 \le \gamma \le 1\right\},$$

представляющее собой множество всех возможных смесей из двух материалов.

Предполагая, что рассматривается низкочастотное излучение, можно показать, что задача (3.12) - (3.14) имеет слабое решение для любого показателя преломления $a \in A$. Кроме того, можно ограничить $\|u\|_{H^1(\Omega)}$ независимо от рассматриваемой смеси $a \in A$ (см. [17]). Тогда для каждой смеси $a \in A$ можно определить соответствующие векторы отражения $r(a)$ и преломления $t(a)$. Используя слабую сходимость по аргументам, можно показать, что для каждого $a \in A$ существует последовательность $a_n \in \tilde{A}$, такая, что $r(a_n) \to r(a)$ и аналогично для $t(a)$. В этом смысле A является

замыканием $\tilde{A}$ относительно $F(a)$.

Теперь можно сформулировать «слабую» задачу минимизации

$$\min_{a\in A} J(a) = \frac{1}{2}\left\|F(a) - q\right\|_2^2 \tag{3.17}$$

Решение таких проблем хорошо изучено в работе Добсона (см. [17]), где также представлены и численные результаты.

Можно было, конечно, обобщать постановку задачи на случай широкого диапазона углов падения или диапазона частот. Например, задача дизайна антиотражающих структур изучена в работе (см. [17]), где доказано существование решений задачи минимизации, которая аналогична (3.17). Теория проиллюстрирована в этой же работе примерами дизайна оптимальных антиотражающих структур.

Тема 4. Вычислительный электромагнетизм с вариационными интеграторами и дискретными дифференциальными формами

В данном разделе мы следуем работе [19], в которой введено общее семейство вариационных мультисимплектических численных методов решения системы уравнений Максвелла с использованием дифференциальных форм на пространстве-времени. На этом пути получены новые результаты. Во-первых, показано, что метод Йе и родственные ему методы являются мультисимплектическими и выводятся из дискретного вариационного принципа Лагранжа. Во-вторых, проведено обобщение схемы Йе на случай нерегулярных сеток не только пространственных, но и временных. Это снимает необходимость использовать равномерную дискретизацию по времени и даже выделенные временную и пространственные переменные. Наконец, в качестве примера такого общего использования описанного метода введен асинхронный вариационный метод решения системы уравнений Максвелла. Эти результаты проиллюстрированы эскизом компьютерного моделирования, который демонстрирует замечательное поведение характеристик системы.

4.1. Вариационные методы

Метод Йе, известный под названием (finite-difference time-domain, FDTD), был предложен в работе Йе [20] и остается одним из наиболее успешных методов в области численного решения задач электромагнетизма. Хотя он не является методом высокого порядка, его предпочитают во многих приложениях, потому что он сохраняет существенные структурные характеристики уравнений Максвелла, тогда как другие методы не обладают таким свойством. В качестве существенной характеристики выступает сохранение уравнения $\rho = div\,\mathbf{D}$ в дискретной форме, а также стационарность электростатического решения вида $\mathbf{E} = -\nabla\varphi$

(см. [21]). Мы показываем, что эти желаемые свойства являются непосредственными следствиями дискретной вариационной и дискретной дифференциальной структур метода Йе, которые отражают геометрию уравнений Максвелла. Далее показано, как строить другие вариационные методы, в итоге обладающие теми же численными свойствами, однако применимые в более широком классе дискретизаций.

Вариационные численные методы и симметрии

Методы численного интегрирования использовались первоначально для компьютерного моделирования классических механических систем. Такие характеристики которых как симплектичность, сохранение импульса и сохранение энергии весьма существенны. (Хороший обзор этих методов и их применение см. в [22].) В ряду таких методов вариационные интеграторы (вариационные методы численного решения) разработаны на основе дискретизации вариационного принципа Лагранжа для системы с последующим требованием, чтобы полученные численные траектории удовлетворяли дискретной версии принципа стационарного действия Гамильтона. Такие методы автоматически являются симплектическими, и они точно сохраняют дискретные импульсы, ассоциированные с симметриями лагранжиана. А именно у систем с трансляционной инвариантностью сохраняются линейные дискретные импульсы, а у систем с вращательной инвариантностью - дискретные моменты вращения и т. д. Вдобавок вариационные интеграторы демонстрируют хорошее долговременное поведение энергии без искусственного численного затухания (см. [23].)

Этот вариационный подход был развит до дискретизации общей мультисимплектической теории поля с приложениями к нелинейным уравнениям (см. [19]). Были также построены (см. [19]) асинхронные вариационные интеграторы (АВИ), с помощью которых стало возможным использовать различные шаги по времени для каждого элемента

нерегулярной пространственной сетки, сохраняя при этом вариационную и геометрическую структуры. Применение этих методов показало не только их практичность, но и превосходство над существовавшими методами для задач эластодинамики. Дальнейшее продвижение в этом направлении было достигнуто в работе [24].

Хотя разработанные в работе [24] асинхронные вариационные интеграторы могут быть использованы при численном решении задач электродинамики непосредственно, проблемы возникнут из-за калибровочной инвариантности уравнений Максвелла. Покажем, как обойти возникающие препятствия, комбинируя вариационные методы с дискретным внешним исчислением (ДВИ). Эта дифференциально-калибровочная структура оказывается важной при численной реализации метода и является отличительной чертой метода Йе.

Сохранение дискретной дифференциальной структуры

В качестве примера рассмотрим базовое соотношение $B = curl\, A,$ где B - индукция магнитного поля, а A - векторный магнитный потенциал. В силу тождеств векторного исчисления $div\, curl = 0$ и $curl\, \mathrm{grad} = 0$, из этого уравнения немедленно следуют два важных следствия. Во-первых, **B** автоматически является недивергентным.

Во-вторых, любое калибровочное преобразование $A \mapsto A + \nabla f$ не изменяет **B**; в этом проявляется калибровочная инвариантность, относительно которой сохраняющимся импульсом служит плотность заряда $\rho = divD$. Аналогичные рассуждения объясняют также инвариантность электростатических решений, так как $\mathbf{E} = -\nabla\varphi$ является безвихревым и инвариантным относительно добавления произвольной константы к скалярному потенциалу φ. Поэтому правильный вариационный интегратор для задач электромагнетизма должен сохранять и дискретные аналоги дифференциальных тождеств.

Этого можно добиться, рассматривая электромагнитные поля не в качестве векторных полей, а в качестве дифференциальных форм на четырехмерном пространстве-времени, как это и делается в классической теории поля. Дискретизируя эти дифференциальные формы в рамках дискретного внешнего исчисления, получаем вариационные интеграторы, автоматически сохраняющие дискретные дифференциальные тождества, такие как $d^2 = 0$ (которые включают в себя рассмотренные ранее div-curl-grad соотношения), и теорему Стокса. Следовательно, они также сохраняют и калибровочные симметрии уравнений Максвелла, и соответствующие дискретные моменты.

Практические следствия учета геометрической структуры

Покажем, что метод Йе является методом именно этого типа, что объясняет многие из его численных свойств. Например, одной из его замечательных черт является то, что электрическое поле **E** и магнитное поле **H** определены не в одних и тех же пространственно-временных точках, а в различных узлах, расположенных в шахматном порядке решеток. Причины, по которым такое специальное расположение приводит к улучшению вычислений, не очевидны: если мы рассматриваем **E** и **H** лишь как векторные поля в трехмерном пространстве,— математические объекты в точности одного типа — почему бы им не быть расположенными в одних и тех же пространственно-временных узлах? Действительно, при реализации многих методов конечных элементов поступают именно так, дискретизируя их в одних и тех же узлах единой сетки. Однако с точки зрения дифференциальных форм в пространстве-времени становится ясно, что размещение их в различных узлах расположенных в шахматном порядке решеток лучше соответствует структуре уравнений Максвелла. Мы увидим, что **E** и **H** получаются из объектов, дуальных друг другу

(пространственно-временные формы F и $*F$), и следовательно, они должны располагаться на двух сдвинутых относительно друг друга сетках.

Аргументом в пользу такого подхода является не только теоретический интерес: геометрия уравнений Максвелла имеет важное значение для численного моделирования. Например, используемая в методе конечных элементов совместная дискретизация в одних узлах, основанная на векторном взгляде на электрическое и магнитное поля, приводит к фиктивным артефактам, в основе появления которых лежит игнорирование геометрической структуры. С другой стороны, метод Йе приводит к резонансным спектрам в полном соответствии с теорией без фиктивных мод (см. [21]). Позднее было показано, что метод сдвинутых сеток позволяет развить быстрые численные методы для электромагнетизма даже в задачах с составными средами с такими сильно разрывными характеристиками, как диэлектрическая и магнитная проницаемости.

Развивая сохраняющие структуру уравнений Максвелла геометрические дискретизации, мы не только поймем метод Йе и его преимущества, но и сможем разработать более общие методы, обладающие этими преимуществами. Семейство этих методов включает в себя Йе-подобный метод Боссавита—Кеттунена [25], который является первым обобщением метода Йе на неструктурированные сетки (например, скорее симплициальные сетки, чем прямоугольные решетки). Общие методы такого типа являются весьма желательными: прямоугольные сетки не всегда практичны и удобны в приложениях с кривыми областями и косыми границами. Если использовать дискретизацию общего вида, но сохраняющую геометрические свойства, можно получить комбинацию лучших черт метода конечных элементов и метода Йе.

Перечень основных результатов

Используя дискретное внешнее исчисление, как сохраняющую структуру геометрическую основу для дискретных сеток общего вида, получаем следующие результаты:

- Метод Йе является, по существу, вариационным интегратором, т. е. он может быть получен с помощью применения принципа стационарного действия Гамильтона к дискретному лагранжиану.
- Следовательно, метод Йе является мультисимплектическим и сохраняет дискретные аналоги импульсов (т.е. сохраняет величины, аналогичные своим непрерывным аналогам в электромагнетизме). В частности, сохраняющаяся плотность электрического заряда понимается как дискретный аналог момента данного интегратора, тогда как сохранение электростатического потенциала решения относится к тождеству $d^2 = 0$, где d – дискретный оператор внешнего дифференцирования.
- Мы также создаем фундамент для более общих методов, допускающих произвольные дискретизации пространства-времени, а не только равные по времени шаги на пространственной сетке. В качестве такого метода ниже приведен асинхронный вариационный интегратор для уравнений Максвелла, у которого каждый пространственный элемент обладает собственным шагом по времени и эволюционирует «асинхронно» со своими соседями. Это значит, что можно выбирать меньшие шаги для областей, где требуется большая точность, используя в других местах при этом большие шаги. Поскольку измельчение шагов в одной части сетки не ограничивает временных шагов в других ее частях, асинхронный

вариационный интегратор может быть численно эффективным и численно устойчивым при меньшем общем числе итераций. Вдобавок к схеме асинхронного вариационного интегратора описывается также полностью ковариантная пространственно-временная схема интегратора для электромагнетизма, не требующая даже расщепления на пространственные и временные компоненты.

Основные выводы

Начнем с обозрения уравнений Максвелла, вначале выражая их через дифференциальные формы из лагранжева вариационного принципа, а далее показывая эквивалентность данной формулировки привычной формулировке в терминах векторного анализа. Затем дается обзор внешних дифференциальных форм, приводятся определения дискретных внешних дифференциальных форм и связанных с ними дискретных операторов на сетках. Эти инструменты дискретного внешнего исчисления используются для формулировки дискретных уравнений Максвелла. Также показывается, что возникающие при этом численные алгоритмы приводят к алгоритмам Йе и Боссавита—Кетунена в качестве специальных случаев, равно как и к асинхронно вариационному методу. В заключение показано, что дискретные уравнения Максвелла могут быть выведены из дискретного вариационного принципа и исследованы другие его геометрические свойства, включая мультисимплектичность и сохранение отображения момента.

4.2. Уравнения Максвелла

Кратко рассмотрим метод дифференциальных форм в применении к задачам электромагнетизма, подготовливая читателя к дискретной формулировке задачи, изучаемой в следующем разделе. Для более детального ознакомления читатель может обратиться к работам Боссавита

[26], Гросса и Котиуга [27].

От векторных полей к дифференциальным формам

Уравнения Максвелла (в отсутствие свободных зарядов и токов) традиционно выражаются в терминах четырех векторных полей в трехмерном пространстве: электрическое поле **E**, магнитное поле **H**, электрическая индукция **D** и магнитная индукция **B**. Чтобы перевести их на язык дифференциальных форм, начнем с замещения электрического поля 1-формой E и магнитной индукции 2-формой B. В координатной записи они выглядят следующим образом

$$E = E_x dx + E_y dy + E_z dz$$

$$B = B_x dy \wedge dz + B_y dz \wedge dx + B_z dx \wedge dy,$$

где $E = \left(E_x, E_y, E_z\right)$ и $B = \left(B_x, B_y, B_z\right)$.

Мотивацией для выбора E в виде 1-формы и B в виде 2-формы служит интегральная формулировка закона Фарадея:

$$\oint_C E \cdot dl = -\frac{d}{dt}\int_S B \cdot dA,$$

где E интегрируется по кривой, а B интегрируется по поверхности. Аналогично в законе Ампера:

$$\oint_C H \cdot dl = -\frac{d}{dt}\int_S D \cdot dA$$

H интегрируется по кривым, а D интегрируется по поверхностям, таким образом мы можем ввести 1-форму H и 2-форму D.

Теперь **E** и **B** связаны с **D** и **H** посредством обычных уравнений связи:

$$D = \varepsilon E,$$

$$B = \mu H.$$

Боссавит и Кеттунен показали, что можно рассматривать ε и μ как соответствующие операторы Ходжа $*_\varepsilon$ и $*_\mu$, которые отображают 1-формы «поля» в 2-формы «потоков индукции» в пространстве.

Отметим, что в вакууме, когда $\varepsilon = \varepsilon_0$ и $\mu = \mu_0$ являются константами, можно записать уравнения лишь в терминах E и B, выбирая подходящие единицы измерения (Гаусса), такие что $\varepsilon_0 = \mu_0 = c = 1$, игнорируя таким образом различие между E и D и между B и H. В дальнейшем ограничимся рассмотрением уравнений Максвелла в вакууме в системе единиц Гаусса в отсутствие свободных зарядов и токов. Для определенности мы будем считать E и H 1-формами, D и B -2-формами.

2-формы Фарадея и Максвелла

В лоренцовском пространстве-времени мы можем объединить E и B в единый объект: фарадеевскую 2-форму:

$$F = E \wedge dt + B.$$

Имеется теоретическое преимущество в объединении электрического поля и магнитной индукции в единый пространственно-временной объект: на этом пути электромагнитные явления могут быть описаны в релятивистски ковариантном виде, без разделения преимущественно на пространственную и временную компоненты. На деле можем вывернуть предыдущую конструкцию наизнанку: взять F в качестве фундаментального объекта, так что E и B появляются только при специфическом выборе системы координат. Если применить звезное отображение Ходжа к F, мы снова получим двойственную 2-форму:

$$*F = H \wedge dt - D,$$

называемую 2-формой Максвелла. Это описывает отношения

двойственности между E и B, с одной стороны, и D и H - с другой, в конструктивной форме.

Электромагнитный вариационный принцип

Пусть A - электромагнитной потенциал, являющийся 1-формой. Он удовлетворяет уравнению $F = dA$ на пространственно-временном многообразии M. Определим 4-форму плотности лагранжиана:

$$L = \frac{1}{2} dA \wedge *dA$$

и соответствующий функционал действия

$$S(A) = \int_M L.$$

Если определить L^2 внутреннее произведение дифференциальных k-форм в виде:

$$\langle\langle \alpha, \beta \rangle\rangle = \int_M \alpha \wedge *\beta = \int_M \beta \wedge *\alpha,$$

то во введенных обозначениях мы можем переписать действие в виде:

$$S(A) = \frac{1}{2} \langle\langle dA, dA \rangle\rangle.$$

Теперь рассмотрим вариацию α потенциала A, такую, что α исчезает на границе ∂M. Тогда вариация функционала действия, соответствующая α, равна:

$$dS(A) \cdot \alpha = \left. \frac{d}{d\varepsilon} \right|_{\varepsilon=0} S(A + \varepsilon\alpha) = \langle\langle dA, d\alpha \rangle\rangle = \langle\langle \delta dA, \alpha \rangle\rangle,$$

где в последнем равенстве мы проинтегрировали по частям с использованием кодифференциального оператора δ, L^2 - сопряженного к d, и того факта, что α исчезает на границе.

Гамильтонов принцип постоянного действия требует, чтобы эта вариация обращалась в ноль при произвольном α, приводя, таким образом, к электромагнитному уравнению Эйлера—Лагранжа:

$$\delta dA = 0. \tag{4.1}$$

Вариационное происхождение уравнений Максвелла

Поскольку $F = dA$, то совершенно ясно, что уравнение (4.1) эквивалентно $\delta F = 0$. Более того, так как $d^2 = 0$, то из этого следует, что $dF = d^2 A = 0.$ Следовательно, уравнения Максвелла относительно фарадеевской 2-формы могут быть записаны в виде:

$$dF = 0, \tag{4.2}$$

$$\delta F = 0. \tag{4.3}$$

Предположим теперь, что выбрана стандартная система координат (x, y, z, t) в R^4. Определим E и B с помощью соотношения $F = E \wedge dt + B.$ Тогда простое вычисление показывает, что уравнение (4.2) эквивалентно паре уравнений:

$$curlE + \frac{\partial B}{\partial t} = 0 \tag{4.4}$$

$$divB = 0. \tag{4.5}$$

Аналогично, если положить $*F = H \wedge dt - D,$ то уравнение (4.3) эквивалентно уравнениям

$$curlH - \frac{\partial D}{\partial t} = 0 \tag{4.6}$$

$$divD = 0. \tag{4.7}$$

Следовательно, этот лагранжиан и метод дифференциальных форм к уравнениям Максвелла в точности эквивалентен классической формулировке в терминах векторных полей на гладком пространстве-

времени. Однако в дискретном пространстве-времени мы увидим, что метод дифференциальных форм не эквивалентен произвольной дискретизации векторных полей, а требует специального выбора дискретных объектов.

Редукция уравнений

Решая начальную задачу, нет необходимости использовать все уравнения Максвелла, чтобы проследить эволюцию системы во времени. Действительно, роторные уравнения (4.4) и (4.6) автоматически сохраняют величины $\mathrm{div}\,\mathbf{B}$ и $\mathrm{div}\,\mathbf{D}$. Поэтому дивергентные уравнения (4.5) и (4.7) можно рассмотривать лишь как ограничения на начальные условия, тогда как роторные уравнения полностью описывают эволюцию системы во времени.

Имеются различные пути доказательства правомерности исключения дивергентных уравнений. Непосредственный способ состоит в применении оператора дивергенции к уравнениям (4.4) и (4.6). Но так как $div\ curl = 0,$ мы получаем соотношения:

$$\frac{\partial}{\partial t} divB = 0,$$

$$\frac{\partial}{\partial t} divD = 0\,.$$

Поэтому если дивергентные ограничения выполняются в начальный момент времени, то они выполняются всегда, так как дивергентные члены постоянны.

В другом подходе заметим, что лагранжиан L зависит лишь от внешней производной dA электромагнитного потенциала, а не от значений самого A. Таким образом, система обладает калибровочной симметрией: любое калибровочное преобразование $A \mapsto A + df$ оставляет dA, а,

следовательно, L, неизменными. Фиксируя координату времени, фиксируем калибровку так, что скалярный электрический потенциал обращается в ноль: $\phi = A(\partial/\partial t) = 0$ (так называемая калибровка Вейля), и, таким образом, A имеет только пространственные компоненты. Действительно, эти три оставшиеся компоненты соответствуют таковым из обычного векторного потенциала A. Редуцированные уравнения Эйлера—Лагранжа при такой калибровке состоят только из уравнений (4.6), в то время как остающаяся калибровочная симметрия $A \mapsto A + \nabla f$ порождает отображение импульса, которое автоматически сохраняет дивиргенцию D во времени. Уравнения (4.4) и (4.5) автоматически сохраняются под действием тождества $d^2 A = 0$, и в действительности не являются частью уравнений Эйлера—Лагранжа. Более детально эти вычисления изложены в работе [19].

4.3. Внешнее дискретное исчисление

В этом разделе предлагается краткий обзор фундаментальных объектов и операций дискретного внешнего исчисления, сохраняющего структуру исчисления внешних дифференциальных форм. По построению дискретное внешнее исчисление автоматически сохраняет ряд важных геометрических структур, включая теорему Стокса, интегрирование по частям (с правильным учетом границ), комплекс де Рама, двойственность Пуанкаре, лемму Пуанкаре и теорему Ходжа. Поэтому внешнее дифференциальное исчисление порождает полностью дискретный аналог инструментов, использовавшихся в предыдущем разделе, для формулировки уравнений Максвелла с помощью дифференциальных форм. В последующих разделах мы будем пользоваться этой структурой для формулировки дискретных уравнений Максвелла по аналогии с непрерывной версией.

Логическое обоснование использования внешнего дискретного исчисления в вычислительной электродинамике

Современная вычислительная электродинамика возникла в 1960-х гг., когда метод конечных элементов, основанный на сеточных функциях, успешно использовался для дискретизации дифференциальных уравнений, описывающих двумерную задачу электростатики, сформулированную в терминах скалярного потенциала. К сожалению, первоначальный успех метода конечных элементов оказалось невозможно перенести на трехмерные задачи без появления численных артефактов. После введения Неделиком [1980] граничных элементов пришло понимание, что лучшая дискретизация геометрической структуры электромагнитной теории Максвелла является ключем для преодоления возникших трудностей (см. Гросс и Котиуга [2004]). Математический аппарат, развитый Вейлем и Уитни в 1950-х гг., в контексте алгебраической топологии, как оказалось, обеспечил необходимую базу, на которой могут быть построены устойчивые численные методы решения задач электродинамики(см. работу Боссавита [1998]).

Основанное на этой ранней работе дискретное внешнее исчисление является попыткой построить общую вычислительную схему, учитывающую эти важные геометрические структуры. Базируясь на картановском внешнем исчислении дифференциальных форм на гладких многообразиях, внешнее дискретное исчисление является дискретным исчислением, построенным изначально на дискретных многообразиях с целью сохранения ковариантной природы используемых величин. Этот вычислительный аппарат основан на понятиях дискретных цепей и коцепей, используемых в качестве основных строительных блоков для дискретизации, совместимой с такими важными геометрическими структурами, как комплексы де Рама. Представления с помощью цепей и коцепей являются привлекательными для вычислительных методов

благодаря своей концептуальной простоте и элегантности. Кроме того, это представление берет свое начало в работах Уитни [1957], который ввел отображения Уитни и де Рама, устанавливающие изоморфизм между симплициальными коцепями и дифференциальными формами Липшица.

Сетки и двойственные сетки

Дискретное внешнее исчисление имеет дело с задачами, в которых гладкие n-мерные многообразия заменяются дискретными сетками, точнее комплексами ячеек, образующими ориентированное многообразие и допускающими метрику. Простейшим примером такой сетки является конечный симплициальный комплекс - триангуляция 2-мерной поверхности. Обычно мы будем обозначать комплекс через K, а ячейку комплекса - через σ.

По сетке K можно построить двойсвенную сетку $*K$, когда каждой k-ячейке σ соответствует двойственная $(n-k)-$ячейка $*\sigma$ ($*K$ является «двойственным» к K в смысле двойственности графов.) Подробности построения приведены в [19].

Дискретные дифференциальные формы

Фундаментальными объектами внешнего дифференциального исчисления являются дискретные дифференциальные формы. Дискретная k-форма α^k приписывает каждой ориентированной k-мерной ячейке σ^k комплекса K действительное число. (Индекс k не обязательное, но полезное обозначение, напоминающее о том, с каким порядком формы или ячейки имеем дело.) Это число обозначается как $\langle \alpha^k, \sigma^k \rangle$, и может быть итерпретировано как интеграл формы α^k по элементу σ^k:

$$\langle \alpha, \sigma \rangle \equiv \int_{\sigma} \alpha .$$

Например, 0-формы приписывают значения вершинам, 1-формы

приписывают значения ребрам и т.д. Это соответствие может быть продолжено по линейности на дискретную область интегрирования: простым сложением значений форм на каждой ячейке, входящей в область интегрирования. Если ячейка входит в область с противоположной ориентацией, значение добавляется со знаком минус. Формально эта область, составленная из k-мерных ячеек, называется цепью, а дискретные дифференциальные формы - коцепями, так что $\langle \cdot, \cdot \rangle$ - спаривание коцепей с цепями.

Дифференциальные формы могут быть определены или на комплексе K, или на двойственном ему $*K$. При этом мы будем говорить о них как о первичных формах и двойственных формах соответственно. Существует естественное соответствие между первичными k-формами и двойственными $(n-k)$ - формами, так как каждая первичная k-ячейка имеет двойственную $(n-k)$ - ячейку. Это - важное свойство, которым мы воспользуемся, чтобы определить дискретное отображение Ходжа.

Дискретное внешнее дифференцирование

Дискретное внешнее дифференцирование d строится так, чтобы выполнялась теорема Стокса, которая в непрерывном случае записывается следующим образом:

$$\int_{\sigma} d\alpha = \int_{\partial\sigma} \alpha .$$

Поэтому, если α - дискретная дифференциальная k-форма, то $(k+1)$ - форма $d\alpha$ определена на каждой $(k+1)$ - цепи σ соотношением

$$\langle d\alpha, \sigma \rangle = \langle \alpha, \partial\sigma \rangle,$$

где $\partial\sigma$ - граница k-цепи σ. Поэтому d часто называют кограничным оператором в теории когомологий.

Дискретное отображение Ходжа

Дискретное отображение Ходжа преобразовывает k-формы, заданные на основной сетке в $(n-k)$ - форму на двойственной сетке, и наоборот. В нашем изложении мы будем использовать так называемое диагональное приближение отображения Ходжа (Bossavit [1998]) по причине его простоты, но отметим, что для более высоких порядков аппроксимации потребуются некоторые модификации. Для данной дискретной формы α, ее преобразование Ходжа *α определено соотношением

$$\frac{1}{|*\sigma|}\langle *\alpha, *\sigma\rangle = k(\sigma)\frac{1}{|\sigma|}\langle \alpha, \sigma\rangle,$$

где $|\sigma|$ и $|*\sigma|$ объемы этих элементов, а k - функция причинности, которая равняется +1, когда σ пространственно подобен, и -1 в противном случае.

4.3.6. Дискретное внутреннее произведение.

Определим L^2 внутреннее произведение $\langle\langle \cdot,\cdot \rangle\rangle$ для двух основных k-форм соотношением

$$\langle\langle \alpha, \beta\rangle\rangle = \sum_{\sigma^K} k(\sigma)\binom{n}{k}\frac{|CH(\sigma, *\sigma)|}{|\sigma|^2}\langle \alpha,\sigma\rangle\langle \beta,\sigma\rangle =$$

$$= \sum_{\sigma^k} k(\sigma)\frac{|*\sigma|}{|\sigma|}\langle \alpha,\sigma\rangle\langle \beta,\sigma\rangle,$$

где сумма берется по всем k-мерным элементам σ, а $CH(\sigma, *\sigma)$ n-мерная выпуклая оболочка $\sigma \cup *\sigma$. Последнее равенство справедливо в силу использования кругоцентрической двойственности, так как σ и $*\sigma$

ортогональны друг другу, и, следовательно, $\left|CH(\sigma,*\sigma)\right| = \binom{n}{k}^{-1} |\sigma||*\sigma|$.

(Действительно, это - одно из преимуществ использования кругоцентрической двойственности, так как надо хранить лишь информацию об объемах основных и двойственных ячеек, а не о выпуклых оболочках основных и двойственных ячеек.) Это внутреннее произведение может быть выражено в терминах $\alpha \wedge *\beta$ как и в непрерывном случае, если выбрано специальное дискретное V-произведение.

Дискретное кодифференцирование

Наконец, определяем кодифференциальный оператор δ, который преобразует дискретную $(k+1)$ - форму в k-форму по формуле

$$\delta\beta = (-1)^{nk+1+Ind(g)} * d * \beta,$$

где $Ind(g)$ - индекс метрики. (Для лоренцовской метрики пространства-времени $Ind(g) = 1$.) Этот оператор играет важную роль в интегрировании по частям, так как для сеток без границ или в случае, когда граничные вклады обнуляются, мы имеем

$$\langle\langle d\alpha, \beta\rangle\rangle = \langle\langle \alpha, \delta\beta\rangle\rangle. \qquad (4.8)$$

Эти отношения двойственности и являются причиной названия кодифференцирования.

Применение дискретного внешнего исчисления

Дискретное внешнее исчисление просто и эффективно применяется с помощью линейной алгебры. K-форма α может быть представлена вектором, компонентами которого являются значения α на каждой k-ячейке сетки. Таким образом, для данного списка k-ячеек σ_i^k компоненты вектора равны $\alpha_i = \langle \alpha, \sigma_i^k \rangle$. Внешняя производная d, преобразующая k-

формы в $(k+1)$-формы, тогда представляется матрицей: по сути, это в точности матрица смежности k-ячеек и $(k+1)$-ячеек сетки, со значениями ±1. Отображение Ходжа, преобразующее основные k-формы в двойственные $(n-k)$-формы, представляется квадратной матрицей, а в случае диагонального отображения Ходжа - диагональной матрицей со значениями $k(\sigma_i^k)|*\sigma_i^k|/|\sigma_i^k|$. Дискретное внутреннее произведение является при этом квадратичной формой матрицы отображения Ходжа. Из-за этого прямого изоморфизма между дискретным внешним исчислением и линейной алгеброй проблемы, сформулированные на языке дискретного внешнего исчисления, могут воспользоваться преимуществом существования библиотек программ по решению задач линейной алгебры.

Начальные и граничные условия в дискретном внешнем исчислении

Особое внимание следует обратить на корректное задание начальных и граничных условия на дискретной пространственно-временной границе ∂K. Например, в электродинамике, пусть даже мы хотим задать начальные условия для E и B в момент времени t_0 - но, хотя B определено на ∂K в t_0, E не определено. В действительности, E «живет» на ребрах нормалей к сечениям пространства-времени фиксированным значением t_0. Следовательно, чтобы не модифицировать наши определения, мы должны задавать значения E на полушагах $t_{1/2}$. (Эти полушаги появляются в стандартной схеме Йе.) В некоторых приложениях можно задавать значения E и B в различные моменты времени (например, если полязадаются случайным образом и интегрируются в течение долгого времени при расчете резонансных спектров), но мы хотим работать и с более общими ситуациями. Хотя дискретное внешнее исчисление в основном на границах областей, такие понятия, как дуальные ячейки,

должны быть корретно определены на границе ∂K и рядом с ней.

Для основной сетки K, двойственная сетка $*K$ определена как двойственная по Вороному к K, ограниченная на K. При этом усекаются части двойственных ячеек, выходящие за пределы K (см. иллюстрации 3.1 и 3.2 в работе [19] для сравнения). Это новое определение приводит к добавлению вершин, двойственных по Вороному, к каждому граничному $(n-1)$-симплексу, к имевшимся ранее вершинам, двойственным внутренним n-симплексам. Для завершения двойственной сетки $*K$, добавляем двойственные ребра между смежными вершинами, двойственными к границе, а также между двойственными граничными вершинами и их соседними внутренними двойственными вершинами; процесс продолжается по возрастанию размерности двойственных ячеек. Интуитивно можно представить $(n-1)$-мерную границу как исчезающе тонкую n-мерную оболочку. Таким образом, каждый граничный $(k-1)$-симплекс можно представить в виде призматической k-ячейки, полученной из предыдущей утолщением по нормали к границе. Этот процесс весьма похож на использование виртуальных ячеек границы в общепринятом методе конечных элементов. При таком процессе построения дуальной сетки ее граница совпадает с границей основной сетки и при этом $\partial(*K) = *(\partial K)$.

Вернемся к примеру о начальных условиях для E и B и вспомним, что E определено на ребрах, нормальных к срезам времени t_0. Поэтому, благодаря надлежащему ограничению диаграммы Вороного на область, мы можем теперь определить E на вершинах ∂K в момент времени t_0, при этом данные вершины надо понимать как исчезающе короткие нормали к границе. Наконец отметим, что конструкция $*K$ задает отношение двойственности между условиями Дирихле на двойственной сетке и

условиями Неймана на основной сетке, т. е. между основными полями и двойственными индукциями.

Дискретное интегрирование по частям с учетом ненулевых граничных условий

Когда двойственная сетка определена корректно, двойственные формы могут теперь быть определены и на границе. Поэтому дискретная двойственность между d и δ (уравнение(4.8)) может быть обобщена на случай неисчезающих граничных условий. Если α - основная $(k-1)$-форма и β - основная k-форма, то

$$\langle\langle d\alpha,\beta\rangle\rangle = \langle\langle \alpha,\delta\beta\rangle\rangle + \langle \alpha \wedge *\beta, \partial K\rangle. \qquad (4.9)$$

На границе в этом интеграле α по-прежнему все еще является основной $(k-1)$-формой на ∂K, в то время как $*\beta$ является $(n-k)$-формой, взятой на двойственной границе $*(\partial K)$. Формула (4.9) легко доказывается с использованием знакомого метода дискретного «суммирования по частям», и таким образом согласуется с формулой интегрирования по частям для непрерывных дифференциальных форм.

4.4. Применение дискретного внешнего исчисления к уравнениям Максвелла

В этом разделе покажем, как вывести численные алгоритмы решения уравнений Максвелла с помощью дискретного внешнего исчисления. Для этого поступим следующим образом. Сначала найдем приемлемый способ определения дискретной фарадеевской 2-формы F на сетке. Затем применим дискретные операторы d и δ, чтобы вывести дискретные уравнения Максвелла. И хотя мы еще не доказали, что эти уравнения являются вариационными в дискретном смысле, в работе [19] показано, что вывод непрерывных уравнений Максвелла из лагранжиана остается справедливым и неизменным при использовании дискретных операторов.

Наконец, обсудим, как эти уравнения можно использовать для получения численного метода для вычислительной электродинамики.

В частности, для прямоугольной сетки, покажем, что наша конструкция приводит к традиционной схеме Йе. Для общей триангуляции пространства с равными шагами по времени получим в результате схему Кеттанена и Боссавита. Далее рассмотрим метод асинхронных вариационных интеграторов, где разным пространственным элементам могут соответствовать разные шаги по времени, так что интегрирование по времени уравнений Максвелла может выполняться на элементах асинхронно. Наконец, упомянем об уравнениях на полностью обобщенной пространственно-временной сетке, которая делит R^4 на произвольные 4-симплексы.

Прямоугольная сетка

Предположим, что имеем прямоугольную сетку в R^4, ориентированную вдоль осей (x, y, z, t). Для упрощения выкладок (хотя это не является необходимым) предположим, что сетка взята с равными шагами в пространстве и времени $\Delta x, \Delta y, \Delta z, \Delta t$. Отметим, что наш подход, основанный на внешнем дискретном исчислении, применяется непосредственно к несимплициальным прямоугольным сеткам, так как n-угольник вписан в $(n-1)$-сферу.

Построение.

Так как F является 2-формой, ее значения должны задаваться на 2-гранях этой сетки. В соответствии с непрерывным выражением для F:

$$F = E_x dx \wedge dt + E_y dy \wedge dt + E_z dz \wedge dt +$$

$$+ B_x dx \wedge dz + B_y dz \wedge dx + B_z dx \wedge dy,$$

и благодаря тому, что прямоугольная сетка имеет природу тензорного

произведения, точное задание каждой 2-грани становится простым: шести компонентам F соответствуют в точности шесть типов 2 граней в 4-мерной прямоугольной сетке. Просто задайте значения $E_x \Delta x \Delta t$ на грани, параллельной xt-плоскости, $E_y \Delta y \Delta t$ - на грани, параллельной yt-плоскости и $E_z \Delta z \Delta t$ - на грани, параллельной zt-плоскости. Аналогично задайте $B_x \Delta y \Delta z$ на грани, параллельной yz-плоскости, $B_y \Delta z \Delta x$ - на грани, параллельной xz-плоскости, и $B_z \Delta x \Delta y$ - на грани, параллельной xy-плоскости (рис. 3).

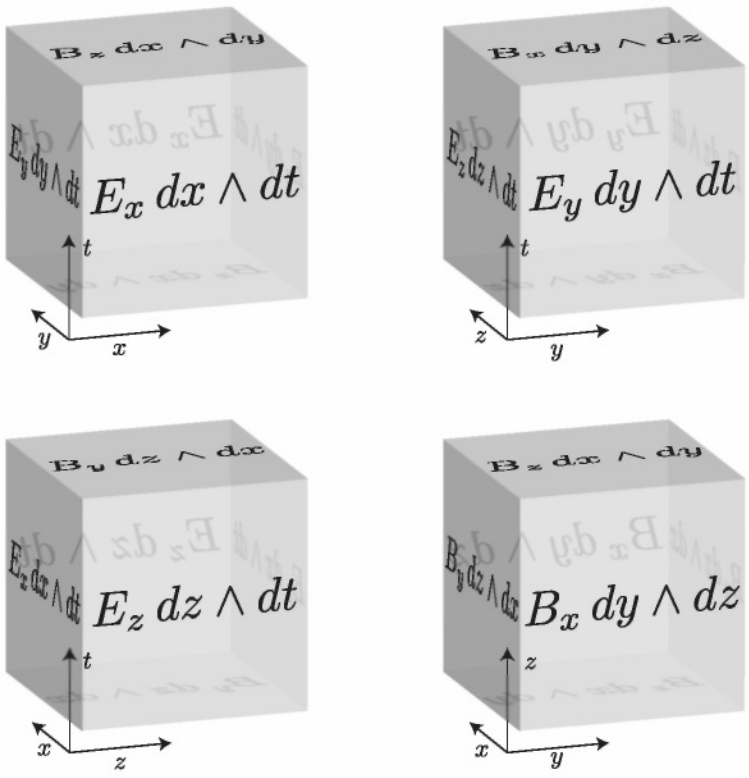

Рис. 3. Значения F на 2-гранях 4-мерной прямоугольной сетки

На рис. 3 показаны три смешанные пространственно-временные 3 ячейки, и единственная чисто пространственная 3 ячейка (внизу справа). Рассмотрим значения F на гранях 4-х мерного прямоугольника $[x_k, x_{k+1}] \times [y_l, y_{l+1}] \times [z_m, z_{m+1}] \times [t_n, t_{n+1}]$. Для упрощения обозначений, мы можем проиндексировать каждую значение F средней точкой 2-грани, на

которой она задана: например, $F\Big|_{k+\frac{1}{2},l,m}^{n+\frac{1}{2}}$ расположен на грани $[x_k, x_{k+1}]\times\{y_l\}\times\{z_m\}\times[t_n, t_{n+1}]$, параллельной плоскости xt. Следующие значения заданы на соответствующих 2-гранях:

$$xt-face: E_x\Big|_{k+\frac{1}{2},l,m}^{n+\frac{1}{2}} \Delta x\Delta t$$

$$yt-face: E_y\Big|_{k,l+\frac{1}{2},m}^{n+\frac{1}{2}} \Delta y\Delta t$$

$$zt-face: E_z\Big|_{k,l,m+\frac{1}{2}}^{n+\frac{1}{2}} \Delta z\Delta t$$

$$yz-face: B_x\Big|_{k,l+\frac{1}{2},m+\frac{1}{2}}^{n} \Delta y\Delta z$$

$$xz-face: B_y\Big|_{k+\frac{1}{2},l,m+\frac{1}{2}}^{n} \Delta z\Delta x$$

$$xy-face: B_z\Big|_{k+\frac{1}{2},l+\frac{1}{2},m}^{n} \Delta x\Delta y.$$

Видим, что «смещенная сетка» является результатом того факта, что E и B заданы на 2-гранях, а не в вершинах 4-грани.

Уравнения движения

Дискретные уравнения движения записываются, как и в непрерывном случае, в виде:

$$dF = 0,$$

$$\delta F = 0,$$

однако теперь эти уравнения интерпретируются в терминах дискретного внешнего исчисления. Рассмотрим вначале интерпретацию dF в терминах

дискретного внешнего исчисления. Так как dF является дискретной 3-формой, ее значения задаются на 3-гранях 4-мерного прямоугольника. Эти значения следующие:

$$xyt-\text{грань}: -\left(E_x \Big|_{k+\frac{1}{2},l+1,m}^{n+\frac{1}{2}} - E_x \Big|_{k+\frac{1}{2},l,m}^{n+\frac{1}{2}} \right)\Delta x\Delta t$$

$$+\left(E_y \Big|_{k+1,l+\frac{1}{2},m}^{n+\frac{1}{2}} - E_y \Big|_{k,l+\frac{1}{2},m}^{n+\frac{1}{2}} \right)\Delta y\Delta t$$

$$+\left(B_z \Big|_{k+\frac{1}{2},l+\frac{1}{2},m}^{n+1} - B_z \Big|_{k+\frac{1}{2},l+\frac{1}{2},m}^{n} \right)\Delta x\Delta y$$

$$xzt-\text{грань}: -\left(E_x \Big|_{k+\frac{1}{2},l,m+1}^{n+\frac{1}{2}} - E_x \Big|_{k+\frac{1}{2},l,m}^{n+\frac{1}{2}} \right)\Delta x\Delta t$$

$$+\left(E_z \Big|_{k+1,l,m+\frac{1}{2}}^{n+\frac{1}{2}} - E_z \Big|_{k,l,m+\frac{1}{2}}^{n+\frac{1}{2}} \right)\Delta z\Delta t$$

$$-\left(B_y \Big|_{k+\frac{1}{2},l,m+\frac{1}{2}}^{n+1} - B_y \Big|_{k+\frac{1}{2},l,m+\frac{1}{2}}^{n} \right)\Delta x\Delta z$$

$$yzt-\text{грань}: -\left(E_y \Big|_{k,l+\frac{1}{2},m+1}^{n+\frac{1}{2}} - E_y \Big|_{k,l+\frac{1}{2},m}^{n+\frac{1}{2}} \right)\Delta y\Delta t$$

$$+\left(E_z \Big|_{k,l+1,m+\frac{1}{2}}^{n+\frac{1}{2}} - E_z \Big|_{k,l,m+\frac{1}{2}}^{n+\frac{1}{2}} \right)\Delta z\Delta t$$

$$+\left(B_x \Big|_{k,l+\frac{1}{2},m+\frac{1}{2}}^{n+1} - B_x \Big|_{k,l+\frac{1}{2},m+\frac{1}{2}}^{n} \right)\Delta y\Delta z$$

$$xyz-\text{грань}:\left(B_x\Big|^{n}_{k+1,l+\frac{1}{2},m+\frac{1}{2}}-B_x\Big|^{n}_{k,l+\frac{1}{2},m+\frac{1}{2}}\right)\Delta y\Delta z$$

$$+\left(B_y\Big|^{n}_{k+\frac{1}{2},l+1,m+\frac{1}{2}}-B_y\Big|^{n}_{k+\frac{1}{2},l,m+\frac{1}{2}}\right)\Delta x\Delta z$$

$$+\left(B_z\Big|^{n}_{k+\frac{1}{2},l+\frac{1}{2},m+1}-B_z\Big|^{n}_{k+\frac{1}{2},l+\frac{1}{2},m}\right)\Delta x\Delta y\,.$$

Полагая каждое из них равным нулю, мы приходим к следующим четырем уравнениям:

$$\frac{B_x\Big|^{n+1}_{k,l+\frac{1}{2},m+\frac{1}{2}}-B_x\Big|^{n}_{k,l+\frac{1}{2},m+\frac{1}{2}}}{\Delta t}=$$

$$\frac{E_y\Big|^{n+\frac{1}{2}}_{k,l+\frac{1}{2},m+1}-E_y\Big|^{n+\frac{1}{2}}_{k,l+\frac{1}{2},m}}{\Delta z}-\frac{E_z\Big|^{n+\frac{1}{2}}_{k,l+1,m+\frac{1}{2}}-E_z\Big|^{n+\frac{1}{2}}_{k,l,m+\frac{1}{2}}}{\Delta y}$$

$$\frac{B_y\Big|^{n+1}_{k+\frac{1}{2},l,m+\frac{1}{2}}-B_y\Big|^{n}_{k+\frac{1}{2},l,m+\frac{1}{2}}}{\Delta t}=$$

$$\frac{E_z\Big|^{n+\frac{1}{2}}_{k+1,l,m+\frac{1}{2}}-E_z\Big|^{n+\frac{1}{2}}_{k,l,m+\frac{1}{2}}}{\Delta x}-\frac{E_x\Big|^{n+\frac{1}{2}}_{k+\frac{1}{2},l,m+1}-E_x\Big|^{n+\frac{1}{2}}_{k+\frac{1}{2},l,m}}{\Delta z}$$

$$\frac{B_z\Big|^{n+1}_{k+\frac{1}{2},l+\frac{1}{2},m}-B_z\Big|^{n}_{k+\frac{1}{2},l+\frac{1}{2},m}}{\Delta t}=$$

$$\frac{E_x\Big|^{n+\frac{1}{2}}_{k+\frac{1}{2},l+1,m}-E_x\Big|^{n+\frac{1}{2}}_{k+\frac{1}{2},l,m}}{\Delta y}-\frac{E_y\Big|^{n+\frac{1}{2}}_{k+1,l+\frac{1}{2},m}-E_y\Big|^{n+\frac{1}{2}}_{k,l+\frac{1}{2},m}}{\Delta x}$$

и $$\frac{B_x\Big|^{n}_{k+1,l+\frac{1}{2},m+\frac{1}{2}} - B_x\Big|^{n}_{k,l+\frac{1}{2},m+\frac{1}{2}}}{\Delta x} + \frac{B_y\Big|^{n}_{k+\frac{1}{2},l+1,m+\frac{1}{2}} - B_y\Big|^{n}_{k+\frac{1}{2},l,m+\frac{1}{2}}}{\Delta y}$$

$$+\frac{B_z\Big|^{n}_{k+\frac{1}{2},l+\frac{1}{2},m+1} - B_z\Big|^{n}_{k+\frac{1}{2},l+\frac{1}{2},m}}{\Delta z} = 0. \qquad (4.10)$$

Эти уравнения являются дискретными аналогами уравнений

$$\frac{\partial B}{\partial t} = -curlE\,,$$

$$divB = 0\,.$$

Более того, так как E и B являются дифференциальными формами, их можно также рассматривать как дискретизацию уравнений Максвелла в интегральной форме. Поскольку дискретное внешнее исчисление удовлетворяет дискретной теореме Стокса, то автоматически сохраняется эквивалентность между дифференциальной и интегральной формулировками электродинамики.

Делая то же самое с уравнением $\delta F = 0$ и задавая значения в этот раз на ребрах вместо 3-граней, получаем еще четыре уравнения:

$$\frac{D_x\Big|^{n+\frac{1}{2}}_{k+\frac{1}{2},l,m} - D_x\Big|^{n-\frac{1}{2}}_{k+\frac{1}{2},l,m}}{\Delta t} =$$

$$\frac{H_z\Big|^{n}_{k+\frac{1}{2},l+\frac{1}{2},m} - H_z\Big|^{n}_{k+\frac{1}{2},l-\frac{1}{2},m}}{\Delta y} - \frac{H_y\Big|^{n}_{k+\frac{1}{2},l,m+\frac{1}{2}} - H_y\Big|^{n}_{k+\frac{1}{2},l,m-\frac{1}{2}}}{\Delta y}$$

$$\frac{D_y\Big|_{k,l+\frac{1}{2},m}^{n+\frac{1}{2}} - D_y\Big|_{k,l+\frac{1}{2},m}^{n-\frac{1}{2}}}{\Delta t} =$$

$$\frac{H_x\Big|_{k,l+\frac{1}{2},m+\frac{1}{2}}^{n} - H_x\Big|_{k,l+\frac{1}{2},m-\frac{1}{2}}^{n}}{\Delta z} - \frac{H_z\Big|_{k+\frac{1}{2},l+\frac{1}{2},m}^{n} - H_z\Big|_{k-\frac{1}{2},l+\frac{1}{2},m}^{n}}{\Delta x}$$

$$\frac{D_z\Big|_{k,l,m+\frac{1}{2}}^{n+\frac{1}{2}} - D_z\Big|_{k,l,m+\frac{1}{2}}^{n-\frac{1}{2}}}{\Delta t} =$$

$$\frac{H_y\Big|_{k+\frac{1}{2},l,m+\frac{1}{2}}^{n} - H_y\Big|_{k-\frac{1}{2},l,m+\frac{1}{2}}^{n}}{\Delta x} - \frac{H_x\Big|_{k,l+\frac{1}{2},m+\frac{1}{2}}^{n} - H_x\Big|_{k,l-\frac{1}{2},m+\frac{1}{2}}^{n}}{\Delta y}$$

и

$$\frac{D_x\Big|_{k+\frac{1}{2},l,m}^{n+\frac{1}{2}} - D_x\Big|_{k-\frac{1}{2},l,m}^{n+\frac{1}{2}}}{\Delta x} + \frac{D_y\Big|_{k,l+\frac{1}{2},m}^{n+\frac{1}{2}} - D_y\Big|_{k,l-\frac{1}{2},m}^{n+\frac{1}{2}}}{\Delta y}$$

$$+ \frac{D_z\Big|_{k,l,m+\frac{1}{2}}^{n+\frac{1}{2}} - D_z\Big|_{k,l,m-\frac{1}{2}}^{n+\frac{1}{2}}}{\Delta z} = 0. \qquad (4.11)$$

Они следуют из задания $*F$ на двойственной сетке, как показано на рис. 4. Это множество уравнений является дискретным аналогом уравнений:

$$\frac{\partial D}{\partial t} = \operatorname{curl} H,$$

$$\operatorname{div} d = 0.$$

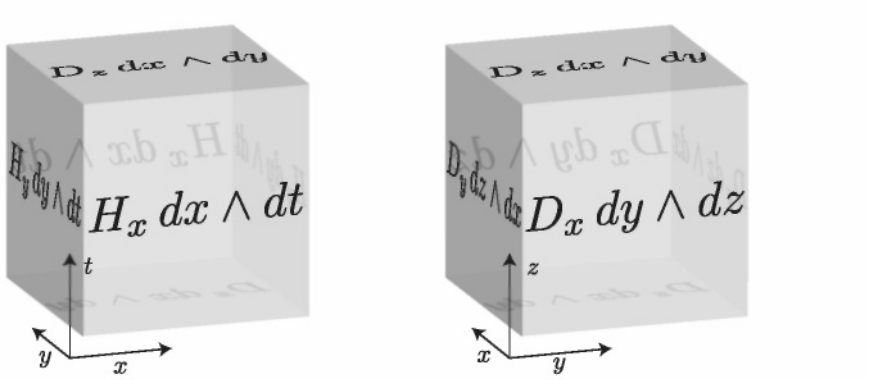

Рис. 4. Значения $*F$ на двойственных 2-гранях прямоугольной сетки.

На рис. 4 показаны смешанная пространственно-временная двойственная 3-ячейка (слева), соответствующая пространственно-подобному основному ребру, и чисто пространственная двойственная 3-ячейка (справа), соответствующая времениподобному основному ребру. Имеются также две других смешанных пространственно-временных ячейки (рис. 3), которые здесь не показаны.

После исключения избыточных дивергентных уравнений (4.10) и (4.11) (см. работу [19]) и подстановки $D = \varepsilon E$, $B = \mu H$, остающиеся уравнения совпадают со схемой Йе.

Неструктурированная пространственная сетка с равномерными шагами по времени

Рассмотрим случай неструктурированной пространственной сетки, но с равномерными шагами по времени. Предположим, что вместо прямоугольной сетки как по пространственным, так и по временным переменным, мы имеем произвольную дискретизацию пространства, однородные шаги по временной оси. (Например, можно взять тетраэдральную пространственную сетку.) Эта сетка содержит два различных типа 2-граней. Во-первых, треугольные грани, принадлежащие пространственной сетке при фиксированном значении времени. Каждое ребро грани такого типа пространственноподобно, т. е. имеет положительную длину, так что функция причинности, определенная в

работе [19], принимает значение $k=1$. Во-вторых, имеются прямоугольные грани, содержащие временной отрезок. Эти грани состоят из простого пространственного ребра, умноженного на временной отрезок. Поскольку они содержат по одному времениподобному ребру, для этих граней $k=-1$. Еще раз убеждаемся, что конструкция нашего внешнего дискретного исчисления учитывает непосредственно лишь сетки такого типа.

Построение.

Можно охарактеризовать дискретные значения F, основываясь на непрерывном выражении

$$F = E \wedge dt + B.$$

Для этого зададим значения B на чисто пространственно подобных гранях и $E\Delta t$ на смешанных пространственно-временных гранях. Из выражения для $*F$ видно, что смешанные двойственные грани должны содержать $H\Delta t$, а пространственно подобные двойственные грани должны содержать D (рис. 5).

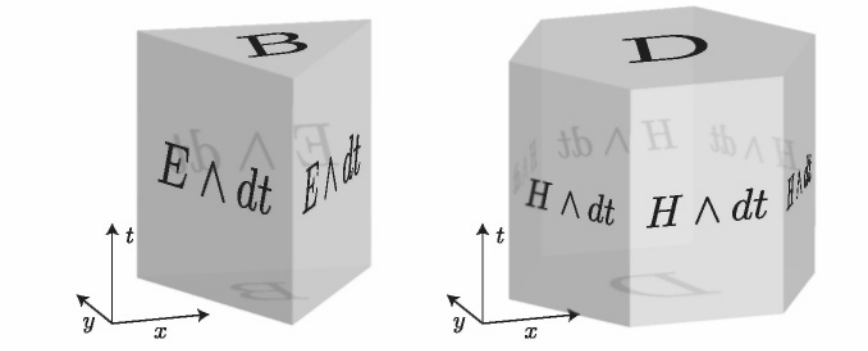

Рис. 5. F на неструктурированной пространственной сетке.

На рис.5 F задается на основных 2 гранях (слева), в то время как $*F$ задается на двойственных 2 гранях (справа), там же показаны значения на смешанных прстранственно-временных 3-ячейках. (Чисто пространственные 3-ячейки, которые соответствуют дивергентным уравнениям и не входят в уравнения движения, здесь не показаны.)

Уравнения движения.

Можно представить дифференциальную форму в виде вектора ее значений на пространственных элементах по методу, описанному в разделе 5.3. Это приводит к векторам B^n и H^n по целочисленным (n) значениям времени и $E^{n+1/2}$ и $D^{n+1/2}$ - по значениям времени в точках $(n+1/2)$. Обозначим через R матрицу инцидентности между ребрами и гранями в пространственной области. То есть R - матрица, соответствующая дискретной внешней производной из основных 1-форм в основные 2-формы, заданной только в пространстве. Транспонированная матрица R^t соответствует внешней копроизводной из пространственных двойственных 1-форм в двойственные 2-формы. Тогда уравнение $dF = 0$ на всех призматических 3-гранях принимает вид:

$$\frac{B^{n+1} - B^n}{\Delta t} = -RE^{n+1/2}.$$

Аналогично $\delta F = 0$ на всех пространственноподобных ребрах, принимает вид:

$$\frac{D^{n+1/2} - D^{n-1/2}}{\Delta t} = R^t H^n.$$

Допустимо задание $dF = 0$ на пространственноподобных 3-гранях (т. е. на тетраэдрах), равно как $\delta F = 0$ на времениподобных ребрах; но это просто дискретный аналог дивергентных условий для B и D, который может быть исключен.

Таким образом, метод дискретного внешнего исчисления для таких сеток эквивалентен Йе-подобной схеме Боссавита-Кеттонена; более того, когда пространственная сетка прямоугольная, наш метод совпадает со стандартной схемой Йе. Однако теперь у нас есть прочная основа для развития этого подхода на случай асинхронных данных.

Неструктурированная пространственная сетка с асинхронными шагами по времени

Вместо того, чтобы выбирать одинаковые шаги по времени для каждого элемента пространственной сетки, как мы делали в предыдущих двух пунктах, часто бывает эффективнее каждому элементу назначать свой оптимизированный шаг по времени. В этом случае не вся сетка целиком эволюционирует во времени синхронно на всех элементах, но каждый элемент эволюционирует со своей скоростью, асинхронно с другими - отсюда название метода: асинхронный вариационный интегратор. Позднее покажем, что такой асинхронный процесс не нарушает вариационной структуры схемы интегрирования. Здесь мы снова допускаем неструктурированность пространственной сетки.

Построение.

После выбора основной пространственной сетки поставим в соответствие каждой пространственной 2 грани (треугольнику) K его собственный набор дискретных времен:

$$\Theta_K = \left\{ t_K^0 < \ldots < t_K^{N_K} \right\}.$$

Например, можно поставить в соответствие каждой грани фиксированный шаг по времени $\Delta t_K = t_K^{n+1} - t_K^n$, сохраняя постоянный шаг по времени для каждого элемента, но позволяя шагу Δt изменяться от элемента к элементу. Для простоты потребуем, чтобы кроме начального момента времени ни у каких двух граней временные значения не совпадали, т. е. $\Theta_K \cap \Theta_{K'} = \left\{ t_0 \right\}$ для $K \neq K'$.

Чтобы првильно задать время на ребре e, на котором разные грани встречаются в разные моменты времени, положим

$$\Theta_e = \bigcup_{K \ni e} \Theta_K = \left\{ t_e^0 \leq \ldots \leq t_e^{N_e} \right\}.$$

Поэтому смешанной прстранственно-временной с 2-грани, которая содержит ребро e с множественным временем, поставим в соответствие набор промежуточных времен

$$\Theta'_e = \left\{ t_e^{1/2} \le \ldots \le t_e^{N_{e-1/2}} \right\},$$

где $t_e^{k+1/2} = \left(t_e^{k+1} + t_e^k \right) / 2$. Значения, заданные на основной сетке асинхронного вариационного интегратора, показаны на рис. 6.

Поскольку $\Theta_e \supset \Theta_K$ при $e \subset K$, каждое пространственное ребро e обладает большим набором дискретных значений времени, чем любая из инцидентных ему граней.

В результате в общем случае нельзя построить кругоцентрическую двойственную пространственно-временную сетку асинхронного вариационного интегратора целиком, так как сетка не призматична и кругоцентры могут не существовать.

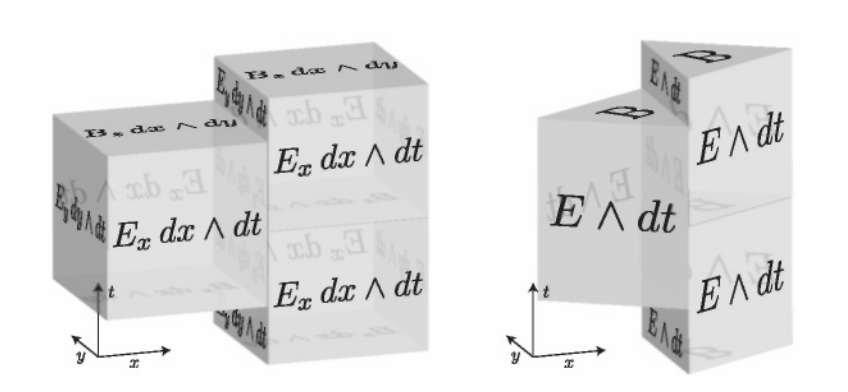

Рис. 6. Часть сетки асинхронного вариационного интегратора

Для прямоугольной пространственной сетки (рис.6 слева) и для пространственно неструктурированной сетки (рис.6 справа) различные высоты пространственно-временных призм отражают тот факт, что элементы эволюционируют с разными шагами по времени. Более того, эти шаги могут быть асинхронными даже для элементов одной горизонтальной грани.

Вместо этого строим кругоцентрически двойственную сетку к пространственной сетке и ставим в соответствие одинаковые шаги по времени основной и двойственной пространственным сеткам:

$$\Theta_{*K} = \Theta_K$$

$$\Theta_{*e} = \Theta_e.$$

В результате получаем корректно определенные основные и двойственные ячейки для каждого пространственно-временного 2-элемента и, следовательно, отображение Ходжа для элементов этого порядка. (Отображение Ходжа для элементов другого порядка не нужно определять для решения уравнений Максвелла.)

Уравнения движения.

Уравнение $dF = 0$ при задании значений на смешанных пространственно-временных 3-ячейках принимает вид:

$$\frac{B_K^{n+1} - B_K^n}{t_K^{n+1} - t_K^n} = -R \sum_{t_K^n < t_e^{m+1/2} < t_K^{n+1}} E_e^{m+1/2} \tag{4.12}$$

Аналогично, уравнение $d * F = 0$ принимает вид

$$\frac{D_e^{m+1/2} - D_e^{m-1/2}}{t_e^{m+1/2} - t_e^{m-1/2}} = R^t \left(H_K^n \mathrm{I}_{\left\{t_K^n = t_e^m\right\}} \right), \tag{4.13}$$

где $\mathrm{I}_{\left\{t_K^n = t_e^m\right\}}$ равно 1, если грань K имеет $t_K^n = t_e^m$ для некоторых n, и 0 в противном случае.

Таким образом, решение начальной задачи Коши может быть систематизировано последовательностью действий.

- Выбрать минимальное время t_K^{n+1}, при котором B_K^{n+1} еще не вычислялось.
- Вычислить B_K^{n+1} в соответствии с уравнением (4.12).

- Добавить значения $H_K^{n+1} = *\mu B_K^{n+1}$
- Вычислить $D_e^{m+3/2}$ на соседних ребрах $e \subset K$ в соответствии с уравнением (4.13).
- Добавить значения $E_e^{m+3/2} = *_e D_e^{m+3/2}$.

Последовательная по времени итерационная схема.

Явная дополняемая схема асинхронного вариационного интегратора может быть реализована выбором элементов сетки из очереди сортировкой по времени, после чего продвигаемся дальше. Однако, как описано ранее, схема не является строго итерационной, так как уравнение (4.13) содержит предыдущие значения E. Это может быть легко устранено, если переписать схему асинхронного вариационного интегратора в виде итераций по переменным A и E, где потенциал A эффективно накапливает кумулятивные вклады E для значения B на соседних гранях. Алгоритм представлен в виде псевдокода. Заметим, что если все элементы имеют одинаковые шаги по времени, наш асинхронный вариационный интегратор сводится к схеме Боссавита—Кеттунена.

Псевдокод асинхронного вариационного интегратора с использованием очередей по приоритету для хранения и обновления данных

// Initialize fields and priority queue

for each spatial edge e **do**

$A_e \leftarrow A_e^0, E_e \leftarrow E_e^{1/2}, \tau_e \leftarrow t_0$ // Store initial field values and times

for each spatial face K **do**

Compute $\tau_K \leftarrow t_0$ the next update time t_K^1

$Q\ \left(t_K^1, K\right)$ // Push element onto queue with its next update time

// Iterate forward in time until the priority queue is empty

repeat

$(t, K) \leftarrow Q.pop(\ )$ // Pop next element K and time t from queue

for each edge e of element K **do**

$A_e \leftarrow A_e - E_e(t - \tau_e)$ // Update neighboring values of A at time t

if t < final-time **then**

$B_k \leftarrow RA_e$

$H_K \leftarrow {}^*\mu B_K$

$D_e \leftarrow {}^*_e E_e$

$D_e \leftarrow D_e + R(e, K) H_K(t - \tau_K)$

$E_e \leftarrow {}^*_e D_e$

$\tau_K \leftarrow t$ // Update element's time

Compute the next update time tnKext

$Q.push(t_K^{next}, K)$ // Schedule K for next update

until $(Q.isEmpty(\ ))$

Полностью неструктурированная пространственно-временная сетка

Наконец, рассмотрим наиболее общий возможный случай: произвольная дискретизация прстранства-времени, случай симплициального 4-комплекса. Такая сетка полностью релятивистски ковариантна, так что F не может быть объективным способом разделено на компоненты E и B вне координатной системы. В большинстве инженерных приложений релятивистские эффекты несущественны, так что 3+1-сетка (как в предыдущих пунктах) почти всегда адекватна, и освобождает от дополнительных сложностей при построении прстранственно-временной сетки. Однако мы полагаем, что существуют научные приложения, в которых ковариантная дискретизация

электродинамики может оказаться очень полезной. Например, имеется много приложений в численной общей теории относительности (например, исчисление Редже), формулируемых на симплициальных 4-комплексах. Возможно, кому-то понадобится промоделировать взаимодействие гравитационного и электромагнитного полей или заряженной материей на такой сетке.

Построение прстранственно-временной сетки.

Вначале предостережение о построении сетки: так как лоренцева метрика не положительно определена, то возможно появление ребер нулевой длины, соединяющих две разные точки в $R4$ (так называемые «светоподобные ребра»). Сетки, содержащие такие ребра, вырождены, подобно евклидовым сеткам, содержащим треугольники с двумя идентичными вершинами. В частности, дискретное отображение Ходжа не определено на элементах с нулевым объемом (нельзя делить на ноль). Но даже без элементов с нулевым объемом для пространственно-временных сеток возможно нарушение причинности, поэтому следует соблюдать особую осторожность.

Когда сетка не предоставляет естественного выбора направления времени, не существует канонического способа разделения F на E и B. Поэтому необходимо задавать значения F сразу на 2-гранях (или, что эквивалентно, задавая значения A на 1 ячейках). Для такой задачи можно выбрать сетку таким образом, чтобы начальный и заключительный шаги по времени ячейки были призматическими, т.е. чтобы можно было воспользоваться начальными и конечными значениями E и B, допуская произвольную дискретизацию F во внутренних ячейках.

Уравнения движения.

Уравнения $dF = 0$ и $\delta F = 0$ могут быть записаны непосредственно в дискретном внешнем исчислении. Но поскольку сетка, вообще говоря, не

структурирована, не существует простого алгоритма вроде тех, что были изложены ранее. Вместо этого уравнения для F записываются в виде разреженной линейной системы, которая после задания корректных граничных условий может быть решена прямым или итеративным методом. Однако ясно, что предыдущие три примера, а именно методы Йе, Боссавита—Кеттунена и асинхронного вариационного интегратора, являются частными случаями решения специальных систем уравнений.

Построение сеток и сохранение энергии.

Известно, что хотя вариационные интеграторы в механике не сохраняют строго энергию, они сохраняют ее в среднем, так что она колеблется вблизи точного значения. Но даже это справедливо лишь тогда, когда берутся шаги одного размера; выбор нерегулярной по времени сетки приводит к плохим результатам (по энергии).

Поэтому нет причин ожидать, что произвольная пространственно-временная сетка будет давать такие же хорошие результаты для энергии, как методы Йе, Боссавита—Кеттунена и асинхронного вариационного интегратора. Более того, ковариантный подход с произвольными прстранственно-временными сетками не допускает даже корректного определения величины энергии. Поэтому конкретизация свойств сетки должна определяться решаемой задачей.

Предложенный в данном разделе набор ковариантных алгоритмов решения электродинамических задач дает обоснование давно известным и хорошо зарекомендовавшим себя методам решения уравнений Максвелла. Показана их естественная связь друг с другом и связь с обобщенным ковариантным алгоритмом, изложенным в последнем разделе. Понимание сути алгоритмов дает возможность их модификации и обобщения на те задачи, которые до сих пор не поддавались решению.

Изложенная теория допускает обобщение на случай любых калибровочных теорий, популярных в современной физике.

5. Численные методы решения обыкновенных дифференциальных уравнений. Практические задания.

Работа должна быть выполнена с использованием графического интерфейса Windows – GUI, решения систем дифференциальных уравнений должны быть представлены в графическом виде.

Можно использовать подпрограммы свободно распространяемого пакета **dMath[29]**. Примерные варианты заданий (системы уравнений) приводятся в конце приложения.

Краткий обзор численных методов решения обыкновенных дифференциальных уравнений

Практическая работа ставит целью сравнительный анализ вычислительных схем решения задачи Коши для обыкновенных дифференциальных уравнений (ОДУ). Такая задача формулируется в следующем виде:

$$\frac{d\vec{y}}{dt} = \vec{f}(\vec{y},t), \quad t \in [t_0,t_n], \quad \vec{y}(t_0) = \vec{y}_0, \ \vec{y}(t) \in R^m, \quad \vec{f}(\vec{y},t) \in R^m. \qquad (5.1)$$

Далее везде будем предполагать, что условия существования решения задачи (5.1) выполнены. Там, где это не принципиально, будем рассматривать методы решения скалярной задачи (5.1).

Вычислительные схемы

В принципе, наиболее простым способом построения решения в точке t_{n+1}, если оно известно в точке t_n, является способ, основанный на разложении функции решения в ряд Тейлора:

$$y(t_{n+1}) = y(t_n) + h \cdot F(t_n, y_n, h), \qquad (5.2)$$

где

$$F(t,y,h) = y'(t) + \frac{h}{2}y''(t) + \frac{h^2}{3!}y'''(t) + \ldots.$$

Если этот ряд оборвать и заменить $y(t_n)$ приближенным значением y_n, то получим приближенную формулу:

$$y_{n+1} = y_n + hf(t_n, y_n) + \frac{h}{2}f'(t_n, y_n) + \ldots. + \frac{h^p}{p!}f^{(p)}(t_n, y_n).$$

При *p=1* она представляет собой вычислительную схему явного метода Эйлера:

$$y_{n+1} = y_n + hf(t_n, y_n). \qquad (5.3)$$

Применение формулы (5.2) ограничено лишь теми задачами, где легко вычисляется производные высших порядков функции *f(y,t)* правой части уравнения (5.1). Заметим, что обычно это не так.

В начале XX века Рунге, Хойн и Кутта предложили подход, основанный на построении формулы для y_{n+1} вида

$$y_{n+1} = y_n + h \cdot \Phi(t_n, y_n, h), \qquad (5.4)$$

в которой функция Φ близка к F, но не содержит производных от функции правой части уравнения. Было получено семейство явных и неявных методов, требующих *s*-кратного вычисления функции правой части на каждом шаге интегрирования (*s*-этапные методы).

Формулы этих методов идеально приспособлены для практических расчетов: они позволяют легко менять шаг интегрирования, являются одношаговыми, достаточно экономичны, по крайней мере, до формул четвертого порядка включительно. Возможно, наиболее известной является формула четырехэтапного метода четвертого порядка:

$$
\begin{aligned}
&y_{n+1} = y_n + \frac{h}{6}(k_1 + k_2 + k_3 + k_4) \\
&k_1 = f(t_n, y_n), \\
&k_2 = f(t_n + \tfrac{h}{2}, y_n + \tfrac{h}{2}k_1) \\
&k_3 = f(t_n + \tfrac{h}{2}, y_n + \tfrac{h}{2}k_2) \\
&k_4 = f(t_n + h, y_n + hk_3)
\end{aligned}
\qquad (5.5)
$$

Одна из основных проблем, связанных с применением методов (5.3) - (5.5) (а в действительности, всех явных методов), состоит в выборе величины шага интегрирования h, обеспечивающей устойчивость вычислительной схемы.

За счет некоторого вычислительного усложнения этих формул был получен класс неявных методов, у которых отмеченная проблема в значительной степени снята. Неявные вычислительные схемы представляют собой алгебраические уравнения, в общем случае нелинейные, относительно значений y_{n+1}. Например:

неявный метод Эйлера

$$y_{n+1} = y_n + hf(t_{n+1}, y_{n+1}); \qquad (5.6)$$

метод трапеции

$$y_{n+1} = y_n + \frac{h}{2}[f(t_{n+1}, y_{n+1}) + f(t_n, y_n)]. \qquad (5.7)$$

Получив приближение к решению в точках t_1, t_2,.., t_n можно использовать их для нахождения решения в точке t_{n+1}. Эта идея приводит к группе многошаговых методов, например, к методам Адамса. В форме Лагранжа вычислительные схемы неявных методов этого класса (методы Адамса—Мултона) имеют вид:

$$y_{n+1} = y_n + h \cdot \sum_{r=1}^{p} \beta_r f_{n-r+1}. \qquad (5.8)$$

Заметим, что неявный метод Эйлера и метод трапеции являются

частными случаями последней вычислительной схемы. Очевидно, использование многошаговых формул ставит задачу вычисления p начальных значений y_1 , $y_2 \ldots$ y_p , точность задания которых должна быть не хуже точности соответствующей формулы.

Отмеченная выше трудность выбора шага интегрирования *h*, обеспечивающего численную устойчивость метода, делает неявные схемы предпочтительнее для практического использования. В этой связи отметим некоторые проблемы их реализации.

Из (5.8) получаем

$$y_{n+1} = h \cdot \beta_0 \cdot f(y_{n+1}, t_{n+1}) + g_n . \qquad (5.9)$$

Можно показать, что если выполняется неравенство

$$h < \frac{1}{L} \cdot \frac{1}{|\beta_0|},$$

где L - константа Липшица для функции *f(y)*, то существует единственное решение алгебраического уравнения (5.9), которое можно получить методом простой итерации:

$$y_{n+1}^{(k+1)} = h \cdot \beta_0 \cdot f(y_{n+1}^{(k)}, t_{n+1}) + g_n, \quad k = 0,1,\ldots . \qquad (5.10)$$

Очевидна желательность выбора «хорошего» начального приближения $y_{n+1}^{(0)}$. Его надлежащий выбор обеспечивается посредством явной формулы того же порядка точности. В этом случае явная схема выполняет роль прогнозирующей, а неявная формула (5.10) реализует коррекцию решения, и весь комбинированный процесс становится методом прогноза-коррекции - П(ВК)^k, где k – номер итерации в процессе (5.10).

Реализация неявных вычислительных методов по схеме (5.10) не является единственно возможной. Для решения получающихся

нелинейных уравнений и их систем (как в методах Рунге—Кутты) широко применяется метод Ньютона. Однако проблема вычисления начального приближения и общая идея схемы «прогноз-коррекция» остаются без изменений. В этой связи обратим внимание на то, что переход к неявным схемам в значительной степени снял проблему выбора величины шага интегрирования *h* как фактора, определяющего устойчивость метода, но привел к проблемам выбора начального приближения и величины шага *h*, обеспечивающих сходимость итерационного процесса решения нелинейных алгебраических уравнений. Ясно, что для линейных задач этих проблем нет, и применение неявных вычислительных схем для них существенно проще и надежней.

Наличие жестких задач (жесткость - свойство задачи, а не метода) сделало неявные вычислительные схемы особенно привлекательными и привело к разработке группы специальных жесткоустойчивых методов (методы Гира). Напомним, что явление жесткости возникает в системах ОДУ, являющихся моделями физических систем, в которых протекают процессы с существенно (на несколько порядков) отличающимися постоянными времени.

Для того чтобы определить, является ли данная задача Коши жесткой, необходимы сведения о ее поведении в окрестности частного решения *y(t)*. В окрестности такого решения уравнение (5.1) можно хорошо аппроксимировать линеаризованным уравнением:

$$\frac{d\vec{y}}{dt} = J(t_n)\cdot[\vec{y}(t) - \vec{y}(t_n)] + \vec{f}(t_n, \vec{y}(t_n)), \qquad (5.11)$$

где $J(t_n)$ - матрица Якоби векторной функции $\vec{f}(t,\vec{y})$, вычисленная в точке $(t_n, \vec{y}(t_n))$.

Если *J(t)* на некотором интервале *t* изменяется мало и s_i -

собственные числа матрицы Якоби, то задача (5.11) будет жесткой в некотором интервале $T=[t_0,t_n]$, если для $t \in T$ выполняется:

$$\mathrm{Re}(s_i)<0, \quad i=1,2,...m$$

$$\omega(t)=\max[\mathrm{Re}(s_i)]/\min[\mathrm{Re}(s_i)]>>1.$$

Систему уравнений можно считать жесткой, если $\max\omega(t)$ больше 10, однако во многих задачах радиоэлектроники, физики, химической кинетики, управления этот коэффициент достигает значений порядка 10^6 и более.

Введение понятия жесткой устойчивости метода позволило сконструировать группу методов, в которых величина шага интегрирования выбирается так, чтобы быстро затухающие и не оказывающие существенного влияния компоненты решения аппроксимировались устойчиво, тогда как для компонент с большими постоянными времени гарантировалась точность аппроксимации.

Наиболее распространенным классом линейных многошаговых методов для жестких задач являются «формулы дифференцирования назад» (более общее название - методы Гира). Вычислительная формула их основывается на интерполяционной формуле Лагранжа для правой части уравнения (5.1)

$$f(t_{m+1,}y_{m+1})=-\frac{1}{h_{m+1}}\sum_{i=0}^{k}a_i y_{m-i+1}\,, \qquad (5.12)$$

$$a_i=\frac{t_m-t_{m-1}}{t_m-t_{m-i}}\prod_{p=1,p\neq i}^{k}\frac{t_m-t_{m-p}}{t_{m-i}-t_{m-p}}, \quad i=1,2,...,k$$

$$a_0=-\sum_{j=1}^{k}a_j\,.$$

Выражение (5.12) представляет собой нелинейное уравнение

относительно y_{m+1}. Для вычисления начального приближения $y_{m+1}^{(0)}$ по значениям $y_m, y_{m-1}, ..., y_{m-k}$ строится интерполяционный полином *k-го* порядка для функции *y(t)* и посредством экстраполирования вычисляется значение $y_{m+1}^{(0)}$. Заметим, что по формулам (5.12) можно вести интегрирование с переменным шагом. Вместе с тем, при фиксированном шаге коэффициенты a_i не изменяются, что существенно упрощает метод.

Погрешность численного решения и методы ее оценки

Главным источником погрешности численного решения y_n в точке t_n:

$$e_n = \| \vec{y}(t_n) - \vec{y}_n \| \tag{5.13}$$

является погрешность аппроксимации метода, которая возникает на каждом шаге интегрирования. Для этой погрешности, в предположении гладкости решения, справедлива оценка

$$e_a = \| \vec{y}(t_0 + h) - y_1 \| \le Ch^{p+1}, \tag{5.14}$$

где p – порядок метода. В общем случае глобальная (полная) погрешность решения определяется как погрешностью аппроксимации, так и тем, как уравнение метода преобразует предыдущие погрешности $e_{n-1}, e_{n-2} \ldots$. Доказано, что если погрешность аппроксимации удовлетворяет соотношению (5.14), а функция метода $\vec{\Phi}(t, \vec{y}, h)$ такова, что выполняется неравенство

$$\| \vec{\Phi}(t, \vec{y}_1, h) - \vec{\Phi}(t, \vec{y}_2, h) \| \le L \| \vec{y}_1 - \vec{y}_2 \|,$$

то для полной погрешности (5.13) справедлива оценка:

$$| e_n | \le h^p \cdot (C/L) \cdot \exp[L(t_n - t_0) - 1], \tag{5.15}$$

где $h = \max\{h_i\}$, $1 \le i \le n$ и L – константа Липшица.

Очевидно, что даже упрощенные оценки погрешности по приведенным формулам не представляют практического интереса, поскольку требуют вычисления или оценивания верхних границ частных производных высших порядков от $f(t,y)$ (постоянные C и L). Для практики, однако, оценки погрешности необходимы, чтобы реализовать выбор достаточно малой величины h_n для получения требуемой точности, и одновременно - достаточно большой для минимизации вычислительных затрат.

Самый старый способ оценки погрешности для одношаговых методов, который использовал еще Рунге, состоит в двукратном вычислении значения y_n с шагом h и $h/2$. При этом погрешность решения, полученного с шагом h, оценивается выражением

$$e_{n+1} = \frac{y_{n+1}(h/2) - y_{n+1}(h)}{2^p - 1} \cdot 2^p, \tag{5.16}$$

а решения, полученного с шагом $h/2$, -

$$e_{n+1} = \frac{y_{n+1}(h/2) - y_{n+1}(h)}{2^p - 1}. \tag{5.17}$$

Еще одна идея оценки величины ошибки численного решения состоит в выполнении шага от точки t_{n-1} к t_n методами порядка p и $p+1$ для получения оценки ошибки интегрирования методом порядка p:

$$y_{n+1}^{(p)} - y(t_{n+1}) \cong y_{n+1}^{(p)} - y_{n+1}^{(p+1)}. \tag{5.18}$$

Очевидна вычислительная «дороговизна» такого способа оценки ошибки. Эта идея нашла наиболее элегантное воплощение в предложенных Инглендом (1967) и Фельбергом (1969) вложенных методах. Они представляют собой формулы Рунге—Кутты, которые одновременно дают возможность получить решения порядка p и $p+1$.

Первые методы такого типа были предложены Мерсоном (1957). Однако в его методе решение $y(p+1)$ имеет пятый порядок точности только для линейных уравнений с постоянными коэффициентами. Поэтому этот метод в общем случае переоценивает погрешность при малых h, но все же работает вполне удовлетворительно.

Формулы Фельберга не имеют подобных ограничений. Фельберг получил целое семейство таких методов различного порядка, но наиболее популярным стал шестиэтапный метод 4-го порядка. Следует иметь в виду, что Фельберг минимизировал коэффициент погрешности аппроксимации для решения низшего порядка (4-го). Поэтому точность результата высшего (5-го) порядка трудно поддается оценке, и этот результат используется только для регулирования величины шага.

Использование процедуры прогноза-коррекции для методов Гира дает простой и эффективный способ оценки погрешности аппроксимации. Брайтон (1972) показал, что ее главную часть можно выразить через разность предсказанного и скорректированного значений решения. Так, для метода k-го порядка справедлива оценка:

$$e_n = h_n \frac{y_n - y_n^{(0)}}{t_n - t_{n-k-1}}. \qquad (5.19)$$

Общие проблемы реализации численных методов

Одной из центральных проблем реализации алгоритмов численного интегрирования задачи Коши является выбор стратегии регулирования шага интегрирования и оценки допустимой его величины при интегрировании с постоянным шагом.

При интегрировании с постоянным шагом выбор величины шага определяется, главным образом, условиями устойчивости метода. Для линейных задач исчерпывающую информацию на этот счет несут

собственные числа матрицы Якоби правой части системы. Для нелинейных задач максимальный шаг можно оценить величиной $h_{\max} = C / L$, где L – константа Липшица функции $\vec{f}(\vec{y},t)$ и C – постоянная, существенно зависящая от метода, но редко превосходящая 10.

Методы, обладающие более высокими показателями устойчивости являются неявными, и их реализация приводит к необходимости решения систем нелинейных алгебраических уравнений. Ограничения, накладываемые свойствами устойчивости метода, при этом могут быть существенно ужесточены условиями сходимости итерационного процесса, применяемого для решения алгебраических систем уравнений, порождаемых численным методом интегрирования. Метод простой итерации вида (5.10) часто практически неприемлем (для жестких задач особенно), так как для его сходимости требуется выполнение условия, накладывающего на величину шага столь же сильное ограничение, как и то, которого старались избежать, переходя от явных схем к неявным.

В таких ситуациях для решения нелинейных уравнений оказывается более эффективным метод Ньютона, который позволяет нередко добиться сходимости при существенно меньших ограничениях на величину шага. Точность прогноза для $\vec{y}_{n+1}$, используемого для начала итерационного ньютоновского процесса, может оказывать заметное влияние на скорость сходимости. Поэтому следует иметь в виду, что дисперсия величины прогноза $\vec{y}_{n+1}$ быстро растет с ростом числа значений $\vec{y}_{n+1-i}$ i = 1,2,…,k, участвующих в предсказывающей формуле, что может привести к существенному снижению эффективности многошаговых формул при $k > 4$.

Если предпринять все необходимые меры для снятия ограничения на величину шага по условиям устойчивости метода и сходимости

итерационного процесса, то выбор шага может быть обусловлен лишь требуемой точностью решения. При решении жестких систем возможны две стратегии выбора шага.

Если влияние на решение быстро затухающих компонент фундаментального решения не представляет интереса, то процесс интегрирования можно провести с постоянным и достаточно большим, по сравнению с малыми постоянными времени системы, шагом. В этом случае можно получить значительный выигрыш в вычислительных затратах за счет упрощения вычислительной схемы (постоянства коэффициентов метода).

Если в решении требуется точное представление быстро изменяющихся компонент решения, то начальная величина шага должна быть меньше наименьшей постоянной времени задачи. Поскольку эти компоненты решения быстро затухают, то шаг в процессе интегрирования можно увеличить до значений порядка постоянной времени существенной (медленной) компоненты решения.

Для эффективного использования последней стратегии необходим относительно простой механизм изменения шага интегрирования в соответствии с оценками погрешности аппроксимации, приведенными выше. На основании формул (5.16)-(5.19) вычисляется оценка ошибки решения:

$$err = \max_{1 \le i \le m} \frac{e_{n,i}}{d_i} \tag{5.20}$$

где d_i - масштабирующий множитель для i-й компоненты вектора решения (при вычислении абсолютной погрешности - d_i=1, а при вычислении относительной погрешности - $d_i = | y_{n,i} |$). Для повышения надежности программы разумно использовать масштабирование вида

$d_i = \max\{|y_{n,i}|, |y_{n,0}|, 1\}$.

Величина ошибки (5.20) сравнивается с заданной допустимой погрешностью Tol и на основании оценки (5.14) вычисляется оптимальная величина следующего шага $h_{new} = \left(\frac{Tol}{err}\right)^{1/p} \cdot h_n$.

Соблюдая определенную осторожность, этот шаг следует несколько уменьшить умножением на 0.8 или 0.9. Не следует допускать и слишком быстрых изменений величины шага – разумно ограничить его вариации величиной $[0.5h_{new}, 2h_{new}]$.

Если $err \le Tol$, то решение, полученное с шагом h_n, следует считать удовлетворительным, и процесс интегрирования можно продолжить с шагом $h_{n+1} = h_{new}$. В противном случае решение с шагом h_n отбрасывается, и интегрирование из точки t_{n-1} начинается с шагом $h_n = h_{new}$.

Значительные трудности возникают при определении величины Tol - допустимой ошибки на одном шаге. Брайтон с соавт. предложили определять ее из условия:

$$Tol = \frac{e_{\lim}}{T} h_n,$$

где $e_{\lim}$ - предельная допустимая величина ошибки на полном интервале интегрирования T. Заметим, что такая стратегия определения этой величины для очень многих, особенно нежестких, задач приводит к чрезмерно сильным ограничениям величины шага.

Тестовые задачи

1. $\frac{d\vec{y}}{dt} = \begin{bmatrix} 0 & 1 \\ -2 & -2 \end{bmatrix} * \vec{y}$, $\vec{y}_0(0) = (0,1)^t$.

2. $$\frac{d\vec{y}}{dt} = \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix} * \vec{y} \quad , \quad \vec{y}_0(0) = (3, -1)^t .$$

3. Жесткая задача.

$$\frac{d\vec{y}}{dt} = \begin{bmatrix} 998 & 1998 \\ -999 & 1999 \end{bmatrix} * \vec{y} \quad , \quad \vec{y}_0(0) = (1,1)^t$$

4. Задача с регулируемой жесткостью

$$\frac{d\vec{y}}{dt} = \begin{bmatrix} d & 1/eps \\ 0.0 & -1/eps \end{bmatrix} * \vec{y} \quad , \quad \vec{y}_0(0) = (0,1)^t$$

$$d < 0 \ , \quad eps > 0 \ , \ |eps| < 1.0$$

5. $$\frac{d\vec{y}}{dt} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} * \vec{y} \quad , \quad \vec{y}_0(0) = (0,1)^t$$

6. $$\frac{d\vec{y}}{dt} = \frac{1}{t^2+1} \begin{bmatrix} -t & +1 \\ -1 & -t \end{bmatrix} * \vec{y} \quad , \quad \vec{y}_0(0) = (0,1)^t$$

7. $$\frac{d\vec{y}}{dt} = \begin{bmatrix} 0 & 1 \\ -(d^2 + \frac{2}{t^2}) ; & \frac{2}{t} \end{bmatrix} * \vec{y} \quad , \quad \vec{y}_0(0) = (0,1)^t , \quad t_0 > 0$$

8. Нелинейная задача с регулируемой жесткостью.

$$\frac{d\vec{y}}{dt} = \begin{bmatrix} -(eps^{-1} + 2.0) * y_1 + y_2^2 / eps \\ y_1 - y_2 - y_2^2 \end{bmatrix}, \quad \vec{y}_0(0) = (1,1)^t \ , \ 0 \le t \le 1$$

9. $$\frac{d\vec{y}}{dt} = \begin{bmatrix} -y_2 + y_1 * (y_1^2 + y_2^2 - 1) \\ y_1 + y_2 * (y_1^2 + y_2^2 - 1) \end{bmatrix}, \ \vec{y}_0(0) = (1+C)^{-1/2}, \ 0 \ )^t$$

10. $$\frac{d\vec{y}}{dt} = \begin{bmatrix} y_2 - y_3 \\ y_1^2 + y_2 \\ y_1^2 + y_3 \end{bmatrix} \quad , \quad \vec{y}_0(0) = (1, -1, \ 0)^t$$

11. $$\frac{d\vec{y}}{dt} = \begin{bmatrix} t + \frac{2}{t} * y_1 - \sqrt{y_2} \\ 2 * \sqrt{y_2} \end{bmatrix}, \quad \vec{y}_0(1.0) = (-1,\ 4)^t$$

12. $$\frac{d\vec{y}}{dt} = \begin{bmatrix} -y_1 + 3y_3 + y_2 * y_3 - 3 * Exp(-3t) - Exp(-5t) \\ y_1 - 2y_2 + y_1 * y_3 - Exp(-t) - Exp(-4t) \\ 2y_2 - 3y_3 + y_2 * y_1 - 2 * Exp(-2t) - Exp(-3t) \end{bmatrix},$$

$$\vec{y}_0(0) = (1,1,1)^t$$

13. $$\frac{d\vec{y}}{dt} = \begin{bmatrix} y_2 \\ \frac{d}{t^3} * (t * y_2 - y_1)^2 \end{bmatrix}, \quad t_0 > 0$$

$$y_1(t_0) = \frac{t_0}{d} \ln \frac{t_0}{t_0 + 0.1}, \quad y_2(t_0) = \frac{1}{d} \ln \left[\frac{t_0}{t_0 + 0.1} + \frac{0.1}{t_0 + 0.1} \right]$$

14. $$\frac{d\vec{y}}{dt} = \begin{bmatrix} -1 & 1 & 0 & 0 \\ -100 & -1 & 0 & 0 \\ 0 & 1 & -100 & 0 \\ 0 & 0 & -10000 & -100 \end{bmatrix} \vec{y}, \quad \vec{y}_0(0) = (1,0,0,0)^t, T_{\max} \le 4.0$$

15. $$\frac{d\vec{y}}{dt} = \begin{bmatrix} -y_1 + y_2^2 + y_3^2 + y_4^2 \\ -10y_2 + 10 * (y_3^2 + y_4^2) \\ -40y_3 + 40y_4^2 \\ -100y_4 + 2.0 \end{bmatrix}, \quad \vec{y}_0(0) = (1,\ 0,\ 0,\ 0)^t.$$

Задания

1. При интегрировании нескольких систем линейных уравнений на основе информации о значениях собственных чисел матрицы системы оценить максимальную величину шага устойчивого интегрирования и проверить эту оценку экспериментально. Зафиксировать величину шага,

при которой метод теряет устойчивость.

2. При интегрировании нескольких систем нелинейных уравнений обратить внимание на эффективность различных методов реализации схемы прогноз-коррекция.

3. Исследовать поведение полной ошибки численного решения при интегрировании с постоянным шагом методами различного порядка. (методы Эйлера, РК2 – РК4), т.е. получить зависимость максимальной погрешности решения задачи от порядка метода при условии постоянства шага интегрирования, и методами одного порядка при вариациях величины шага. Дать качественное описание поведения функции полной погрешности решения.

При интегрировании жестких задач:

- получить экспериментальные характеристики эффективности явных методов (методы РК4, явный Эйлера);
- установить возможность и условия интегрирования задачи неявными методами с большим и постоянным шагом (метод трапеции);
- сравнить эффективность применения метода Гира второго порядка и метода трапеции.

Содержание отчета

Отчет должен включать:

- формулировку цели эксперимента по каждому пункту задания;
- протокол необходимых экспериментальных результатов;
- оценку степени соответствия результатов теоретическим представлениям и объяснение возможных причин несоответствия.

Литература

1. *Беллман Р.* Введение в теорию матриц. — М.: Наука, 1976.
2. *Chu H.* Finite Difference Approach to Optical Scattering of Gratings //Proceedings of the SPIE, Volume 5188, pp. 358-370 (2003).
3. *Moharam M.G., Grann E.B., Pommet D.A.* Formulation for stable and efficient implementation of the rigorous coupled-wave analysis of binary gratings //J. Opt. Soc. Am. A12, 1068(1995).
4. *Li L.* Use of Fourier series in the analysis of discontinuous structures //J. Opt. Soc. Am. A13, 1870(1996).
5. *Lalanne P.* Improved formulation of the coupled-wave method for two-dimensional gratings //J. Opt. Soc. Am. A14, 1592(1997).
6. *Popov E., Nevitere M.* Grating theory: new equations in Fourier space leading to fast converging results for TM polarization //J. Opt. Soc. Am. A17, 1773(2000).
7. *Magnus W.* On the exponential solution of differential equations for linear operator //Comm. Pure Appl. Math. V7 649(1949).
8. *Barabanenkov Y.N., Kouznetsov V.L., Barabanenkov M.Y.* Transfer relation for electromagnetic wave scattering from periodic dielectric one-dimensional interface: TE polarization. Progress in Electromagnetics Research, Pier 24, 39(1999).
9. *Li R.* Unconventional Reflexive Numerical Methods for Matrix Differential Riccati Equations, Technical Report 2000-36, Department of Mathematics University of Kentucky, 2000.
10. *Strang G.* On the construction and comparison of difference schems, SIAM J. Num. Anal. 5, 506(1968).
11. *Verlet L.* Computer “experiments” on classical fluids. I.

Thermodynamical properties of Lennard- Jones molecules //Phys . Rev. 159(1967).

12. *Numerov B.* Publ. Observatoire Central Astrophys. Russ. 2, 188(1933).

13. *Hairer E., Lubich C., Wanner G.* Geometric Numerical Integration. Structure-Presevering Algorithms for Ordinary Differential Equations. Springer Series in Computational Mathematics 31, Springer, Berlin (2002).

14. *Newmark N.M.* A method of computation for structured dynamics //Proc. of Am. Soc. of Civil Engineers 85 (EM 3), 67(1959).

15. *Cadilhac M.* Rigorous vector theories of diffraction gratings, in Progress in Optics 21, E. Wolf, ed., North-Holland, New York, 1984.

16. Electromagnetic Theory of Gratings, Topics in Current Physics, Vol. 22, edited by R. Petit, Springer-Verlag, Heidelberg, 1980.

17. *Dobson D.C.* Controlled Scattering Of Light Waves: Optimal Design Of Diffractive Optics, in "Control Problems in Industry", Birhauser, Boston, 1995, 97-118.

18. *Шмидт Г.* Электромагнитное рассеяние на периодических структурах //Современная математика. Фундаментальные направления.- 2003, Т. 3, -С. 113-128.

19. *Stern A., Tong Y., Desbrun M., Marsden J. E.* Computational electromagnetism with variational integrators and discrete differential forms, 2007. (arXiv preprint math.NA 0707.4470v2).

20. *Yee K.S.* Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media. //IEEE Trans. Ant. Prop. 14(3), 302–307 (1966).

21. *Bondeson A., Rylander T., Ingelstr¨om P.* Computational

electromagnetics, Texts in Applied Mathematics, vol. 51. Springer, New York (2005).

22. *Hairer E., Lubich C., Wanner G.* Geometric numerical integration //Springer Series in Computational Mathematics, vol. 31. Springer-Verlag, Berlin (2002).

23. *Marsden J.E., West M.* Discrete mechanics and variational integrators //Acta Numer. 10, 357–514 (2001).

24. *Lew A., Marsden J.E., Ortiz M., West M.* Variational time integrators. Internat. //J. Numer. Methods Engrg. 60(1), 153–212 (2004).

25. *Bossavit A., Kettunen L.* Yee-like schemes on staggered cellular grids: A synthesis between FIT and FEM approaches //IEEE Trans. Magn. 36(4), 861–867 (2000).

26. *Bossavit A.* Computational electromagnetism. Electromagnetism. Academic Press Inc., San Diego, CA (1998).

27. *Gross P.W., Kotiuga P.R.* Electromagnetic theory and computation: a topological approach //Mathematical Sciences Research Institute Publications, vol. 48. Cambridge University Press, Cambridge (2004).

28. http://www.cde.spbstu.ru/CD_ED/consulting/svs/coshi_lab.html

29. http://www.unilim.fr/pages_perso/jean.debord/tpmath/tpmath.htm

ОПИСАНИЕ КУРСА И ПРОГРАММА

Цели и задачи магистерской программы «Оптика наноструктур»

Целью учебно-методического комплекса магистерской программы «Оптика наноструктур» является формирование у студентов четкого представления об основных принципах функционирования современных дифракционных оптических элементов и устройств, тонкопленочных многослойных покрытий; о законах взаимодействия электромагнитного излучения видимого диапазона с материалом. Целью является также изучение способов и возможностей математического синтеза и компьютерного проектирования дифракционных оптических покрытий. Полученные знания закрепляются в оптической лаборатории и дисплейном классе на примерах изучения конкретных дифракционных оптических элементов и многослойных покрытий со сложной геометрией.

Задачей учебно-методического комплекса магистерской программы «Оптика наноструктур» является обучение студентов навыкам самостоятельного анализа технических заданий на проектирование дифракционных оптических элементов и устройств. Они должны научиться выбирать из имеющихся в наличии алгоритмов и программ математического синтеза или разрабатывать их самостоятельно. В результате обучения обретут навыки ориентации в научной и бизнес информации с целью выбора нужной функции или нужного инструмента для реализации известной функции в области проектирования и создания дифракционных оптических наноструктур.

Цели и задача курса «Методы дифференциальных разностей расчета оптических покрытий»

Курс «Методы дифференциальных разностей расчета оптических покрытий» является составной частью магистерской программы «Оптика наноструктур». Магистерская программа «Оптика наноструктур» реализуется в рамках направления «Прикладная математика и информатика» и направления «Прикладная математика и физика», а возможно и других направлений. В составе магистерской программы «Оптика наноструктур» курс «Методы дифференциальных разностей расчета оптических покрытий» является обязательным, не привязанным к семестру. Для других магистерских программ этот курс может быть курсом по выбору без привязки к семестру или факультативным на усмотрение методической комиссии программы. Курс носит теоретический и практический характер.

Целью курса является подробное ознакомление студентов с устойчивыми современными методами численного решения систем обыкновенных дифференциальных уравнений, получающихся при применении модального Фурье метода к математической модели взаимодействия электромагнитного излучения в области светового диапазона с диэлектрическими структурами нанометровых размеров. Эта область знаний особенно быстро развивается в последние годы в связи с широким применением наноэлементов и тонких (менее одного микрометра толщиной) пленок, используемых в производстве жидко-кристаллических дисплеев, солнечных батарей на основе диэлектриков, фотоэмиссионных диодов, просветляющих покрытий, поляризаторов, миниатюрных лазеров, управляемых оптических элементов. Задачи оптики наноструктур практически не поддаются аналитическому решению, поэтому важным является не только освоение теоретического материала, но и изучение

специальных численных методов, используемых при решении данного класса задач, приобретение навыков создания программного обеспечения для численного моделирования различных оптических наноструктур.

Задачей курса «Методы дифференциальных разностей расчета оптических покрытий» является обучение студентов использованию специальных методов численного решения систем обыкновенных дифференциальных уравнений для моделирования и проектирования современных оптических устройств на основе тонкопленочных покрытий и дифракционных оптических элементов. Это позволит им при необходимости разрабатывать новое программное обеспечение. Безусловной задачей курса является также освоение существующего программного обеспечения, ориентированного на расчет и проектирование оптических покрытий. В результате обучения они получат умение и навыки правильно оценить сложность научно-исследовательских и конструкторских заданий на разработку дифракционных оптических элементов и устройств, аргументированно выбирать метод решения конструкторской задачи, а затем экономично и эффективно выполнять компьютерный дизайн требуемого дифракционного оптического покрытия или устройства.

Трудоемкость курса составляет 3 кредита; 2 часа лекций и 2 часа лабораторных занятий в дисплейном классе в неделю.

Инновационность курса.

Курс является инновационным по содержанию и по литературе, он включает в себя последние научные достижения в области решения задач дифракционной оптики, когда характерные размеры исследуемых объектов не превышают либо сравнимы с длиной волны оптического излучения. Эта область знаний интенсивно развивалась в последнее время,

но лишь недавно были созданы устойчивые алгоритмы и разработаны численные методы решения задач для многослойных решеток. Следует отметить, что для оптических однослойных и многослойных решеток с характерными размерами больше длины волны оптического излучения устойчивые методы решения известны с середины прошлого века. Сейчас алгоритмы решения оптических задач в субволновой области распространяются на объекты со сложной геометрией, такие как двумерные решетки с произвольным профилем, трехмерные решетки (фотонные кристаллы) и на анизотропные материалы. Они востребованы, поскольку позволяют создавать математические модели взаимодействия излучения с веществом в наномасштабах, а затем с их помощью проектировать новые эффективные устройства в высокотехнологичных областях медицины, энергетики, инфокоммуникаций и приборостроения.

В ходе проведения занятий по этому курсу разработчики предполагают использование традиционных методик преподавания, принятой в странах болонской системы образования, то есть с использованием кредитной системы оценки знаний.

Наряду с традиционными элементами преподавания математических методов решения прикладных задач, разработчики курса предполагают воспользоваться хорошо зарекомендовавшим себя опытом МФТИ и подобных вузов. Для этого в рамках подпрограммы «Оптика наноструктур» осуществляется закупка уникального измерительного и аналитического оборудования для выполнения измерений разнообразных характеристик оптических наноустройств с целью использования этого оборудования в учебном процессе и для проведения научно-исследовательских работ преподавателями, аспирантами и студентами.

По окончании магистратуры по направлению «Оптика наноструктур» выпускники Российского университета дружбы народов

станут конкурентно-способными специалистами в области проектирования современных оптических устройств, которые не будут испытывать затруднений при последующем трудоустройстве.

Данное направление научно-практических разработок сформировалось лишь в последние 10 – 15 лет. Поэтому наблюдается сильный дефицит учебно-методической литературы не только в России, но и во всем мире. Разрабатываемые в рамках инновационной программы «Оптика наноструктур» учебные пособия восполнят в некоторой степени этот пробел и составят основной список литературы для слушателей курсов. Вместе с ними следует использовать несколько учебников и монографий, вышедших в свет к настоящему времени и перечисленные в списке литературы. Курс базируется на публикациях научных статей мировых лидеров исследований в данной области в научной периодике, диссертационных работах их учеников, включающих работы по непосредственному моделированию, дизайну и последующему изготовлению лабораторных образцов оптических элементов и устройств. В список дополнительной и рекомендуемой литературы включены все научно-исследовательские публикации, положенные в основу предлагаемого курса.

В качестве практических заданий, курсовых работ и тем рефератов слушателям магистерской программы будут предложены актуальные проблемы и задачи, решение которых востребовано современным уровнем развития высокотехнологичных отраслей промышленности и научно-исследовательских лабораторий.

Структура курса (с указанием количества часов аудиторных/самостоятельной работы на темы)

Темы лекций

Тема 1. Обыкновенные дифференциальные уравнения, системы дифференциальных уравнений, линеаризация системы обыкновенных дифференциальных уравнений, нормы векторов и матриц, дифференцирование и интегрирование векторов и матриц, бесконечные суммы векторов и матриц, доказательство теоремы существования и единственности решения системы линейных обыкновенных дифференциальных уравнений (1 пара).

Тема 2. Системы линейных обыкновенных дифференциальных уравнений в нормальном виде, решение системы обыкновенных дифференциальных уравнений в виде матричной экспоненты, доказательство сходимости степенного ряда матричной экспоненты, свойства матричной экспоненты, вычисление матричной экспоненты в случае наличия кратных корней с использованием Жордановых клеток. (1 пара)

Тема 3. Системы линейных обыкновенных дифференциальных уравнений в нормальной форме с диагонализуемыми матрицами, метод диагонализации матрицы решения системы линейных обыкновенных дифференциальных уравнений. (1 пара)

Тема 4. Сведение системы уравнений Максвелла к бесконечной системе линейных обыкновенных дифференциальных уравнений методом разложения Фурье, гипотеза Релея, теорема Блоха-Флоке, граничные условия. (1 пара)

Тема 5. Дифференциальный метод решения системы обыкновенных дифференциальных уравнений, полученной «модальным методом Фурье»; Метод стрельбы с последующим решением системы линейных алгебраических уравнений для вычисления переходных матриц (2 пары)

Тема 6. Обсуждение сходства и различия FD-метода и RCWA-метода Фурье системы обыкновенных дифференциальных уравнений (1 пара)

Тема 7. Конечно-разностный метод Йе-Тафлава решения системы уравнений Максвелла, система разностных уравнений в первоначальной форме, условная устойчивость метода; Модифицированные системы разностных уравнений, предложенные Тафлавом, устойчивость метода; возможность решения задачи рассеяния электромагнитных волн в криволинейной системе координат, прмеры применения метода. (2 пары)

Тема 8. Решение FDTD -методом задачи взаимодействия электромагнитного излучения с ячейкой нематического жидкого кристалла (1 пара).

Тема 9. Модальный метод Фурье для анизотропных пленок и покрытий (3 пары)

Тема 10. Единообразное изложение FD_метода и RCWA_метода в рамках метода Магнуса разложения операторной экспоненты. (1 пара)

Тема 11. Обзор материала за семестр, консультация перед итоговым контролем знаний (1 пара).

Темы семинарских и практических занятий

Программная реализация алгоритма решения системы обыкновенных дифференциальных уравнений методом Адамса.

Программная реализация алгоритма решения системы обыкновенных дифференциальных уравнений методом Рунге-Кутта четвертого порядка.

Программная реализация алгоритма решения системы обыкновенных дифференциальных уравнений с постоянными коэффициентами.

Сравнение программных реализаций разных алгоритмов решения систем обыкновенных дифференциальных уравнений с постоянными коэффициентами.

Программная реализация алгоритма Верле решения системы уравнений Максвелла.

Программная реализация алгоритма Йе-Тафлава решения системы уравнений Максвелла.

Тестирование программы на модельных примерах. Анализ результатов численных экспериментов.

Темы коллоквиумов и контрольных работ

Анализ методов решения систем обыкновенных дифференциальных уравнений.

Метод Йе-Тафлава решения системы уравнений Максвелла.

Описание методов Ньюмена и Нумерова решения систем обыкновенных дифференциальных уравнений.

Описание системы контроля знаний:

Общие правила выполнения контрольных заданий;

Требования к оформлению работы

Постановка задачи.

1. Краткая формулировка задачи.
2. Развернутая постановка задачи с указанием основных режимов работы и их сценариев.

Алгоритм решения.

1. Математическое описание алгоритма.
2. Структура алгоритма ядра программы (укрупненная блок-схема).

Тестирование.

1. Описание основных режимов тестирования алгоритма и программы и результатов работы программы.
2. Список возможных ошибок и аномалий, описание реакции программы на них.

Заключение.

Содержит общие комментарии и замечания исполнителя о выполненной работе.

Приложение.

Приложение должно содержать текст программы (полная распечатка или распечатка алгоритма ядра программы).

Работа должна быть представлена в виде распечатанного текста и на дискете (Word + Delphi и/или C++).

Рекомендации к составлению отчета

Оформление.

отчет по работе должен быть оформлен в форме Word-файла.

Содержание отчета.

Каждый пункт задания вычислительного эксперимента должен найти свое отражение в отчете.

Каждый раздел отчета должен содержать:

формулировку цели эксперимента

описание исходных данных - приближаемая функция, интервал и порядок приближения, метод приближения и т.п.

результаты эксперимента, представленные в форме таблиц, гистограмм и графиков

иллюстрационный материал в виде копий экрана с графиками зависимостей погрешности приближения, вида приближаемой функции и т.п.

выводы, следующие из результатов эксперимента в контексте его цели.

Шкала оценок, итоговые оценки (методика выставления)

Бально-рейтинговая методика оценки уровня знаний по обязательной дисциплине «Методы дифференциальных разностей расчета оптических покрытий», не привязанной к семестру

Порядок начисления баллов за семестр.

Контрольная работа № 1: 0 – 40 баллов

Теоретические вопросы: 0 – 10 баллов

Практические задания: 0 – 30 баллов

Контрольная работа № 2: 0 – 40 баллов

Теоретические вопросы: 0 – 10 баллов

Практические задания: 0 – 30 баллов

Контрольная работа № 3: 0 – 20 баллов

Теоретические вопросы: 0 – 20 баллов

Шкала бально-рейтинговой системы.

Баллы за семестр	Автоматическая оценка		Баллы за итоговый контроль знаний	Общая сумма баллов	Итоговая оценка
	Итоговая оценка	Дополнительные баллы			
78 – 80	зачет	по 5 баллов за каждый свыше 76**	0 – 20*	86 – 100	зачет
41 – 77	Нет	Нет	0 – 20 0 – 20	51 – 97 41 – 50	зачет незачет
< 41	незачет	Нет	Нет	Нет	незачет

* студент имеет право не проходить итоговый контроль знаний.

** дополнительные баллы начисляются автоматически:

за 86 баллов, набранных в семестре, начисляется дополнительно 6 баллов (общая сумма баллов – 92);

за 87 баллов – 12 баллов (99);

за 88 баллов – 18 баллов (106);

за 89 баллов – 24 балла (113);

за 90 баллов – 30 баллов (120).

Академическая этика, соблюдение авторских прав.

Все имеющиеся в тексте сноски тщательно выверены и снабжены «адресами». Авторы не включали в свою работу выдержки из работ других авторов без указания на это, не пересказывали чужих работ близко к тексту без отсылки к ним. Авторы также не использовали чужих идей без указания первоисточников. Это касается и источников, найденных в интернете. В необходимых случаях указан полный адрес сайта.

Программа курса «Методы дифференциальных разностей расчета оптических покрытий»

Аннотированное содержание курса.

Первый модуль трудоемкостью в 1 кредит составляют лекции на темы:

Обыкновенные дифференциальные уравнения, системы дифференциальных уравнений, линеаризация системы ОДУ, нормы векторов и матриц, дифференцирование и интегрирование векторов и матриц, бесконечные суммы векторов и матриц, доказательство теоремы существования и единственности решения системы линейных ОДУ. Системы линейных обыкновенных дифференциальных уравнений в нормальном виде, решение системы ОДУ в виде матричной экспоненты, доказательство сходимости степенного ряда матричной экспоненты, свойства матричной экспоненты, вычисление матричной экспоненты в случае наличия кратных корней с использованием Жордановых клеток. Системы линейных обыкновенных дифференциальных уравнений в нормальной форме с диагонализуемыми матрицами, метод диагонализации матрицы решения системы линейных ОДУ.

Далее система уравнений Максвелла сводится к бесконечной системе линейных ОДУ методом разложения Фурье с помощью гипотезы Релея, теорема Блоха-Флоке граничных условий для тангенциальных составляющих полей.

Дифференциальный метод решения системы ОДУ, полученной «Фурье модальным методом»; Метод стрельбы с последующим решением системы линейных алгебраических уравнений для вычисления переходных матриц

В завершение темы обсуждаются сходства и различия систем обыкновенных дифференциальных уравнений, полученных FD-методом и RCWA-методом.

К первому же кредиту относятся практические занятия в дисплейном классе в течение 14 академических часов, а также самостоятельные занятия по выполнению курсовых работ и написанию рефератов.

В конце этого модуля проводится промежуточный контроль знаний.

Второй модуль трудоемкостью в 1 кредит составляют теоретический материал по конечно-разностному методу Йе-Тафлава решения системы уравнений Максвелла, по системам разностных уравнений в первоначальной форме, условная устойчивость метода Йе. Далее рассматриваются модифицированные системы разностных уравнений, предложенные Тафлавом, устойчивость метода; возможность решения задачи рассеяния электромагнитных волн в криволинейной системе координат, прмеры применения метода.

Завершается модуль решением FDTD -методом задачи взаимодействия электромагнитного излучения с ячейкой нематического жидкого кристалла.

К второму кредиту относятся также практические занятия в дисплейном классе в течение 8 академических часов, и самостоятельные занятия по выполнению курсовых работ и написанию рефератов.

В конце этого модуля проводится промежуточный контроль знаний.

Третий модуль трудоемкостью в 1 кредит составляют лекции на тему:

Модальный метод Фурье для анизотропных пленок и покрытий. Единообразное изложение FD_метода и RCWA_метода в рамках метода Магнуса разложения операторной экспоненты.

Завершается модуль обзором материала за семестр и консультацией перед итоговым контролем знаний.

В конце этого модуля проводится итоговый контроль знаний.

Список обязательной и дополнительной литературы с указанием соответствия разделов источника (постранично) разделам читаемого курса

Список обязательной литературы.

1. Методы дифференциальных разностей расчета оптических покрытий. / Под ред. Ловецкого К.П. / - М.: - Изд. РУДН (готовится к печати).
2. Беллман Р Введение в теорию матриц.- М.: Наука, 1969, 368 с.
3. Тихонов А.Н., Васильева А.Б., Свешников А.Г. Дифференциальные уравнения. – М.: Изд. МГУ 1985, 232 с..
4. M. Neviere, E. Popov. Light Propagation in Periodic Media: Differential Theory and Design Marcel Dekker Inc, 2002, 432 p.

5. A. Taflove, Computational Electrodynamics: The Finite Difference Time Domain Method. Norwood, MA: Artech House, 1995.
6. Hanyou Chu. Finite difference approach to optical scattering of gratings. // Proceedings of the SPIE, Volume 5188, pp. 358-370 (2003)/

Список дополнительной литературы и источников в интернете.

7. Методы компьютерной оптики/Под ред. В.А. Сойфера: Учеб. для вузов. — 2-е изд., испр. - М.: ФИЗМАТЛИТ, 2003. - 688 с.
8. Ярив А., Юх П. Оптические волны в кристаллах. М. Мир.1983.
9. Moharam, M. G., E. B. Grann, D. A. Pommet, and T. K. Gaylord, "Formulation for stable and efficient implementation of rigorous coupled-wave analysis of binary gratings," J. Opt. Soc. Am. A, Vol. 12, No. 5, 1068-1076, 1995.
10. Moharam, M. G., D. A. Pommet, E. B. Grann, and T. K. Gaylord, "Stable implementation of the rigorous coupled-wave analysis for surface-relief gratings: enhanced transmittance matrix approach," J. Opt. Soc. Am. A, Vol. 12, No. 5, 1077-1086, 1995.
11. L. Li, "Use of Fourier series in the analysis of discontinuous periodic structures," J. Opt. Soc. Am. A 13, 1870-1876 (1996)
12. P. Lalanne, "Improved formulation of the coupled-wave method for two-dimensional gratings," J. Opt. Soc. Am. A 14, 1592-1598 (1997)
13. P. Lalanne and G. M. Morris, "Highly improved convergence of the coupled-wave method for TM polarization," J. Opt. Soc. Am. A 13, 779-(1996)
14. Chr. Bohley. PhD. These Polarization Optics of Periodic Media, Neuchâtel, 2004

15. Koki Watanabe. PhD. Theses COUPLED-MODE ANALYSIS OF COUPLED WAVEGUIDES USING SINGULAR PERTURBATION TECHNIQUE, Fukuoka Institute of Technology, Japan, 2000
16. K. Watanabe, “Study on the Differential Theory of Lamellar Gratings Made of Highly Conducting Materials,” J. Opt. Soc. Am. A, Vol. 23, No. 1, pp. 69-72, (2006-01).

Темы рефератов, курсовых работ, эссе

Темы рефератов.

1. Метод Магнуса разложения операторной экспоненты.
2. Метод Нумерова решения системы обыкновенных дифференциальных уравнений.
3. Метод Ньюмарка решения системы обыкновенных дифференциальных уравнений.

Темы курсовых работ

1. Метод Йе решения системы уравнений Максвелла.
2. Метод Йе-Тафлава решения системы уравнений Максвелла.
3. Первый метод Ханью Чу решения системы обыкновенных дифференциальных уравнений.
4. Второй метод Ханью Чу решения системы обыкновенных дифференциальных уравнений.
5. Блочно-треугольный UL алгоритм решения системы линейных алгебраических уравнений

Темы курсовых работ с последующим продолжением в качестве магистерской диссертации.

6. Сравнение методов Ханью Чу с методом Рунге-Кутта.
7. Конкретизация метода Магнуса и сравнение с методами Ханью Чу.

Учебный тематический план курса УМК (календарный план, структурированный по видам учебных занятий)

Календарный план (20 недель) учебных занятий по обязательной дисциплине «Методы дифференциальных разностей расчета оптических покрытий», не привязанной к конкретному семестру магистратуры.

Виды и содержание учебных занятий				
Неделя	Лекции	Число часов	Лабораторные занятия	Число часов
1	Обыкновенные дифференциальные уравнения, системы дифференциальных уравнений, линеаризация системы ОДУ, нормы векторов и матриц, дифференцирование и интегрирование векторов и матриц, бесконечные суммы векторов и матриц, доказательство теоремы	2	Программная реализация алгоритма решения системы обыкновенных дифференциальных уравнений методом Адамса.	2

	существования и единственности решения системы линейных ОДУ.			
2	Системы линейных обыкновенных дифференциальных уравнений в нормальном виде, решение системы ОДУ в виде матричной экспоненты, доказательство сходимости степенного ряда матричной экспоненты, свойства матричной экспоненты, вычисление матричной экспоненты в случае наличия кратных корней с использованием Жордановых клеток.	2	Программная реализация алгоритма решения системы обыкновенных дифференциальных уравнений методом Адамса.	2
3	Системы линейных обыкновенных дифференциальных уравнений в нормальной форме с диагонализуемыми матрицами, метод диагонализации матрицы решения системы линейных ОДУ.	2	Программная реализация алгоритма решения системы обыкновенных дифференциальных уравнений методом Рунге-Кутта четвертого порядка.	2
4	Сведение системы уравнений Максвелла к бесконечной системе линейных ОДУ методом разложения	2	Программная реализация алгоритма решения системы обыкновенных дифференциальных уравнений методом Рунге-	2

	Фурье, гипотеза Релея, теорема Блоха-Флоке, граничные условия.		Кутта четвертого порядка.	
5	Дифференциальный метод решения системы ОДУ, полученной «Фурье модальным методом»; Метод стрельбы с последующим решением системы линейных алгебраических уравнений для вычисления переходных матриц.	2	Программная реализация алгоритма решения системы обыкновенных дифференциальных уравнений с постоянными коэффициентами.	2
6	Дифференциальный метод решения системы ОДУ, полученной «Фурье модальным методом»; Метод стрельбы с последующим решением системы линейных алгебраических уравнений для вычисления переходных матриц. Продолжение	2	Программная реализация алгоритма решения системы обыкновенных дифференциальных уравнений с постоянными коэффициентами.	2
7	Обсуждение сходства и различия FD-метода и RCWA-метода Фурье системы ОДУ.	2	Сравнение программных реализаций разных алгоритмов решения систем обыкновенных дифференциальных уравнений с постоянными коэффициентами.	2
8	Промежуточный контроль знаний (Контрольная работа №1)			2
9	Конечно-разностный метод Йе-Тафлава решения системы	2	Программная реализация алгоритма Верле решения системы уравнений	2

	уравнений Максвелла, система разностных уравнений в первоначальной форме, условная устойчивость метода; Модифицированные системы разностных уравнений, предложенные Тафлавом, устойчивость метода; возможность решения задачи рассеяния электромагнитных волн в криволинейной системе координат, прмеры применения метода.		Максвелла.	
10	Конечно-разностный метод Йе-Тафлава решения системы уравнений Максвелла, система разностных уравнений в первоначальной форме, условная устойчивость метода; Модифицированные системы разностных уравнений, предложенные Тафлавом, устойчивость метода; возможность решения задачи рассеяния электромагнитных волн в криволинейной системе координат, прмеры применения метода. Продолжение	2	Программная реализация алгоритма Йе-Тафлава решения системы уравнений Максвелла.	2

11	Конечно-разностный метод Йе-Тафлава решения системы уравнений Максвелла, система разностных уравнений в первоначальной форме, условная устойчивость метода; Модифицированные системы разностных уравнений, предложенные Тафлавом, устойчивость метода; возможность решения задачи рассеяния электромагнитных волн в криволинейной системе координат, прмеры применения метода. Окончание	2	Продолжение программной реализации алгоритма Йе-Тафлава решения системы уравнений Максвелла.	2
12	Решение FDTD - методом задачи взаимодействия электромагнитного излучения с ячейкой нематического жидкого кристалла.	2	Тестирование программы на модельных примерах.	2
13	Решение FDTD - методом задачи взаимодействия электромагнитного излучения с ячейкой нематического жидкого кристалла. Продолжение.	2	Анализ результатов численных экспериментов.	2
14	Промежуточный контроль знаний (Контрольная работа №2)			2

15	Модальный метод Фурье для анизотропных пленок и покрытий.	2	Демонстрация результатов выполнения курсовой работы №1.	2
16	Модальный метод Фурье для анизотропных пленок и покрытий. Продолжение.	2	Демонстрация результатов выполнения курсовой работы № 2.	2
17	Модальный метод Фурье для анизотропных пленок и покрытий. Окончание	2	Демонстрация результатов выполнения курсовой работы № 3.	2
18	Единообразное изложение FD_метода и RCWA_метода в рамках метода Магнуса разложения операторной экспоненты.	2	Демонстрация результатов выполнения курсовой работы № 4.	2
19	Заключительный обзор курса. Консультации по подготовке к итоговому контролю знаний.	2	Заключительный обзор курса. Консультации.	2
20	Итоговый контроль знаний			2