СОДЕРЖАНИЕ

Том 49, номер 11, 2009 год

Contents

Vol. 49, No. 11, 2009

A simultaneous English language translation of this journal is available from Pleiades Publishing, Ltd.
Distributed worldwide by Springer. *Computational Mathematica and Mathematical Physics* ISSN 0965-5425.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2009, том 49, № 11, с. 1907–1919

УДК 519.626

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ ОПТИМАЛЬНОЙ СТРАТЕГИИ ХИМИОТЕРАПИИ С УЧЕТОМ ДИНАМИКИ ЧИСЛА КЛЕТОК НЕОДНОРОДНОЙ ОПУХОЛИ

© 2009 г. А. В. Антипов, А. С. Братусь

(119991 Москва, Ленинские горы, МГУ, ВМК)

e-mail: aleksandr_antipo@mail.ru, applmath1miit@yandex.ru

Поступила в редакцию 03.09.2008 г.

Рассматривается математическая модель динамики численности клеток опухоли. Предполагается, что опухоль состоит из клеток двух типов: клеток, поддающихся воздействию химиотерапевтического средства, и клеток, которые этому воздействию не поддаются. Считается, что законы роста числа всех видов клеток задаются логистическими уравнениями. Мера воздействия химиотерапевтического средства на опухоль определяется функцией терапии. Рассматриваются два типа функций терапии: монотонно возрастающая функция и немонотонная функция, имеющая пороговое значение. В первом случае воздействие препарата на опухоль тем сильнее, чем больше его концентрация. Во втором случае имеется некоторая пороговая величина концентрации химиотерапевтического средства, при превышении которой интенсивность терапии падает. Также изучается случай, когда на суммарную величину используемого средства накладывается интегральное ограничение. Ранее близкая по постановке задача изучалась для случая линейной функции терапии при отсутствии ограничения на количество химиотерапевтического средства. С помощью принципа максимума Понтрягина найдены необходимые условия оптимальности, на основании которых сделаны важные выводы о характере оптимальной стратегии терапии. Численно найдены решения задачи оптимального управления, когда целью управления является минимизация общего числа клеток опухоли в случае монотонной и пороговой функций терапии, а также с учетом интегрального ограничения на количество химиотерапевтического средства. Библ. 12. Фиг. 9.

Ключевые слова: математическая модель оптимальной химиотерапии, задача оптимального управления, численные методы.

1. ПОСТАНОВКА ЗАДАЧИ

Рассмотрим модель динамики числа клеток изолированной неоднородной несосудистой опухоли в условиях действия на опухоль химиотерапевтического препарата, способного убивать клетки опухоли. Предполагается, что опухоль состоит из клеток двух типов: клеток, поддающихся воздействию химиотерапевтического средства, и клеток, которые этому воздействию не поддаются.

Пусть $y(t)$ – общее число злокачественных клеток в момент времени t, а $x(t)$ – количество клеток опухоли, не поддающихся химиотерапевтическому воздействию. Тогда $z(t) = y(t) - x(t)$ – количество чувствительных к лекарству (т.е. поддающихся воздействию препарата) клеток в момент t.

Рост числа клеток, не поддающихся воздействию, описывается уравнением

$$\frac{dx}{dt} = rx\left(1 - \frac{y}{\theta}\right) + \alpha r\left(1 - \frac{y}{\theta}\right)(y - x). \tag{1.1}$$

Первое слагаемое в правой части равенства (1.1) описывает естественный рост числа клеток за счет их деления. Второе слагаемое описывает прирост числа указанных клеток за счет того, что часть чувствительных к препарату клеток со временем становится невосприимчивой к его действию. Доля клеток, сначала поддававшихся химиотерапевтическому воздействию, но затем перешедших в класс нечувствительных к проводимой терапии, определяется коэффициентом α, где $0 < \alpha < 1$. Предполагается, что законы роста числа всех видов клеток задаются логистическими уравнениями. Константа $\theta > 0$ имеет смысл предельного количества клеток обоих видов, ко-

торое может содержать изолированная опухоль, а коэффициент $r > 0$ характеризует скорость роста числа клеток.

Для описания динамики числа клеток опухоли были предложены многочисленные модели (см., например, [4]). Следует отметить, что использование логистического закона роста достаточно хорошо согласуется с экспериментальными данными (см. [5]). Выбор значений параметров r и θ логистического закона представляет самостоятельную задачу и определяется конкретным видом опухоли.

Динамика общего числа злокачественных клеток описывается следующим уравнением:

$$\frac{dy}{dt} = ry\left(1 - \frac{y}{\theta}\right) - F(c)(y - x). \tag{1.2}$$

Здесь $c(t)$ – концентрация химиотерапевтического препарата в тканях опухоли, а $F(c)$ – функция, описывающая влияние химиотерапевтического воздействия на чувствительные к этому воздействию злокачественные клетки. Далее будем называть функцию $F(c)$ функцией терапии.

Первое слагаемое в правой части уравнения (1.2), как и раньше, описывает естественный рост числа клеток опухоли, подчиняющийся логистическому закону. Второе слагаемое характеризует убыль числа злокачественных клеток под воздействием химиотерапевтического средства.

Наконец, третье уравнение описывает изменение концентрации химиотерапевтического препарата в тканях опухоли с течением времени:

$$\frac{dc}{dt} = -\gamma c + u. \tag{1.3}$$

Здесь $u(t)$ – управление, характеризующее интенсивность поступления в ткани опухоли химиотерапевтического средства, γ – коэффициент диссипации.

Начальные условия имеют вид

$$x(0) = x_0, \quad y(0) = y_0, \quad c(0) = c_0 \tag{1.4}$$

и удовлетворяют неравенствам

$$0 < x_0 < y_0, \quad c_0 \geq 0.$$

Нашей целью является решение следующей задачи оптимального управления: найти такие время $T^* \in (0, +\infty)$ и функцию $u^* : [0, T^*] \longrightarrow \mathbb{R}$

$$u^* \in U = \{u \,|\, u \in L^{\infty}[0, T^*], 0 \leq u(t) \leq M, \dot{\forall} t \in [0, T^*]\}, \tag{1.5}$$

которые доставляют минимум функционалу

$$J(u(\cdot), T) = y(T). \tag{1.6}$$

Другими словами,

$$J(u^*(\cdot), T^*) = \min\{J(u(\cdot), T) \,|\, u(\cdot) \in L^{\infty}[0, T], T > 0; 0 \leq u(t) \leq M, \dot{\forall} t \in [0, T]\}.$$

Функционал представляет собой общее число клеток опухоли в момент окончания терапии.

Будем считать, что $F(c)$ – непрерывно дифференцируемая функция. Далее будут рассматриваться два типа функций терапии:

1) монотонно возрастающая функция терапии: $F(0) = 0$, $F(c) > 0$ при $c > 0$ и $F'(c) > 0$ при $c \geq 0$;

2) немонотонная функция терапии: $F(0) = 0$, $F(c) > 0$ при $c > 0$, $F'(c) > 0$ при $0 \leq c < \tilde{c}$, $F'(\tilde{c}) = 0$ и $F'(c) < 0$ при $c > \tilde{c}$.

Немонотонность функции терапии означает, что терапевтический эффект от воздействия уменьшается, если концентрация препарата превосходит некоторую величину $\tilde{c}$, которая является пороговой.

Наряду с поставленной задачей рассмотрим также задачу (1.1)–(1.6) с интегральным ограничением типа неравенства на фазовую переменную $c(t)$:

$$\int_0^T c(t)dt \leq C. \tag{1.7}$$

Здесь T – момент окончания терапии, C – заданное положительное число. Формула (17) накладывает ограничение на суммарное количество химиотерапевтического средства, которое может быть использовано в процессе лечения.

Ранее в [2] изучалась задача, аналогичная задаче (1.1)–(1.6) в случае линейной функции терапии $F(c) = c$. С помощью принципа максимума Понтрягина для нее были найдены необходимые условия оптимальности. Доказана релейность искомой функции оптимального управления. В простейшем случае линейного закона роста числа клеток (закона Мальтуса) доказана оптимальность стратегии лечения, заключающейся в наискорейшем увеличении концентрации химиотерапевтического средства на протяжении всего процесса лечения. Доказано, что такая стратегия является наилучшей в некотором классе допутимых релейных управлений.

Задачи терапии однородных опухолей исследовались в [6]–[9]. Отметим также близкие по смыслу работы по оптимальной стратегии иммунотерапии [10], [11].

Целью нашего исследования является построение оптимальной стратегии в случае нелинейного закона роста числа клеток и различных типов функций терапии, а также решение задачи с учетом ограничения (1.7).

2. СЛУЧАЙ МОНОТОННОЙ ФУНКЦИИ ТЕРАПИИ

Рассмотрим случай монотонной функции терапии.

На плоскости переменных x, y определим множество

$$\Omega = \{(x, y) \in \mathbb{R}^2 \,|\, 0 < x < y < \theta\}.$$

Лемма 1. *Пусть $u(t) \geq 0$ – существенно ограниченная функция, $(x(t), y(t), c(t))$ – соответствующее решение системы* (1.1)–(1.4) *на полупрямой* $[0, +\infty)$, *начальные данные x_0, y_0, c_0 которой удовлетворяют условию $(x_0, y_0) \in \Omega$. Тогда проекция данного решения на плоскость x, y – кривая $(x(t), y(t))$ – не будет иметь общих точек с границей множества Ω.*

Доказательство леммы в целом повторяет аргументы аналогичного утверждения из [1], и мы его не приводим. Отметим важное следствие из этой леммы. В любой момент времени t для фазовых переменных x и y выполняется условие

$$0 < x(t) < y(t) < \theta.$$

Напомним, что $y(t)$ – общее количество клеток в опухоли, $x(t)$ – количество клеток, не поддающихся воздействию химиотерапевтического препарата, θ – предельное количество клеток обоих видов.

Перейдем к решению поставленной задачи. Гамильтониан системы (1.1)–(1.3) имеет вид

$$H(x, y, c, \psi_1, \psi_2, \psi_3, u) = \psi_1\left(rx\left(1 - \frac{y}{\theta}\right) + \alpha r\left(1 - \frac{y}{\theta}\right)(y - x)\right) +$$

$$+ \psi_2\left(ry\left(1 - \frac{y}{\theta}\right) - F(c)(y - x)\right) + \psi_3(-\gamma c + u).$$

Сопряженная система и краевые условия для нее задаются следующими равенствами:

$$\begin{gathered}
\frac{d\psi_1}{dt} = -\left(r\psi_1\left(1 - \frac{y}{\theta}\right)(1 - \alpha) + \psi_2 F(c)\right), \\
\frac{d\psi_2}{dt} = -\left(\psi_1\left(-\frac{r}{\theta}x - \frac{r\alpha}{\theta}(y - x) + \alpha r\left(1 - \frac{y}{\theta}\right)\right) + \psi_2\left(r\left(1 - \frac{y}{\theta}\right) - \frac{r}{\theta}y - F(c)\right)\right), \\
\frac{d\psi_3}{dt} = \gamma\psi_3 + \psi_2 F'(c)(y - x), \\
\psi_1(T) = -\frac{\partial y(T)}{\partial x(T)} = 0, \quad \psi_2(T) = -\frac{\partial y(T)}{\partial y(T)} = -1, \quad \psi_3(T) = -\frac{\partial y(T)}{\partial c(T)} = 0.
\end{gathered} \tag{2.1}$$

Согласно принципу максимума Понтрягина, оптимальное управление $u^*(t)$ должно максимизировать гамильтониан в каждый фиксированный момент времени t. Учитывая, что гамильтониан линеен относительно управления, приходим к выводу, что оптимальное управление имеет вид

$$u^* = \begin{cases} M, & \text{если} \quad \psi_3 > 0, \\ 0, & \text{если} \quad \psi_3 < 0, \\ \text{неопределенно}, & \text{если} \quad \psi_3 = 0. \end{cases} \tag{2.2}$$

В последнем случае ($\psi_3 = 0$) будем говорить, что имеет место сингулярный закон управления системой.

Лемма 2. *Оптимальное управление не может быть сингулярным на интервале положительной длины.*

Доказательство. Предположим, что оптимальное управление u^* сингулярно на некотором интервале времени $\Delta \subset [0, T^*]$. Тогда на этом интервале $\psi_3 \equiv 0$. Но в таком случае из третьего уравнения сопряженной системы (2.1) следует, что

$$\psi_2(t)F'(c(t))[y(t) - x(t)] \equiv 0$$

на том же самом интервале. Согласно лемме 1, для всех t из интервала $[0, T^*]$ справедливо неравенство $y(t) - x(t) > 0$. Постоянная c_0 и функция $u(\cdot)$ неотрицательны по условию задачи, поэтому

$$c(t) = c_0 \exp(-\gamma t) + \int_0^t \exp[-\gamma(t-s)]u(s)ds \geq 0.$$

Согласно условиям, наложенным на функцию $F(\cdot)$, величина $F'(c(t)) > 0$ при любом t. Следовательно, $\psi_2 \equiv 0$ на множестве Δ. Поскольку гамильтониан тождественно равен нулю вдоль оптимальной траектории, приходим к выводу, что $\psi_1 \equiv 0$ на множестве Δ. Так как сопряженные уравнения линейны относительно сопряженных переменных, а коэффициенты при последних – ограниченные на сегменте $[0, T^*]$ функции, то выполнение на интервале положительной длины равенства $\psi_1 \equiv \psi_2 \equiv \psi_3 \equiv 0$ влечет за собой справедливость того же тождества при всех t из $[0, T^*]$. С другой стороны, $\psi_2(T^*) = 1$. Полученное противоречие говорит о неверности сделанного предположения и означает, что оптимальное управление u^* не может быть сингулярным на интервале положительной длины.

Лемма 3. *Если $u^*(t)$ – оптимальное управление, то существует $\varepsilon > 0$ такое, что $u^*(t) \equiv M$ при $t \in (T^* - \varepsilon, T^*)$.*

Доказательство. Из (2.1) следует, что

$$\frac{d\psi_3}{dt}(T^*) = -F'(c(T^*))[y(T^*) - x(T^*)] < 0.$$

Так как $\psi_3(t)$ – непрерывная функция и $\psi_3(T^*) = 0$, то можно утверждать, что $\psi_3(t) > 0$ в некоторой окрестности точки T^*. Тогда из представления (2.2) для оптимального управления будет следовать утверждение леммы.

Прямым следствием предыдущих двух лемм является

Лемма 4. *Оптимальным управлением является либо функция-константа $u^*(t) \equiv M$, либо релейная функция, поочередно принимающая значения 0 и M на интервалах времени положительной длины, причем на последнем из этих интервалов функция $u^*(t)$ тождественно равна постоянной M.*

Докажем, что в простейшем случае линейного закона роста числа клеток (закона Мальтуса) функция-константа $u^*(t) \equiv M$ является оптимальным управлением.

В этом случае система (1.1)–(1.4) имеет вид

$$\begin{gathered}\frac{dx}{dt} = rx + \alpha r(y-x),\\ \frac{dy}{dt} = ry - F(c)(y-x),\\ \frac{dc}{dt} = -\gamma c + u,\\ x(0) = x_0, \quad y(0) = y_0, \quad c(0) = c_0.\end{gathered} \tag{2.3}$$

Из третьего уравнения находим

$$c(t) = c_0\exp(-\gamma t) + \int_0^t \exp[-\gamma(t-s)]u(s)ds.$$

Отсюда следует, что если $u_1(t) \geq u_2(t)$, то $c_1(t) \geq c_2(t)$.

Пусть $z(t) = y(t) - x(t)$ (количество чувствительных к действию химиотерапевтического средства клеток). Тогда, вычитая из второго уравнения системы (2.3) первое, будем иметь

$$\frac{dz}{dt} = [r(1-\alpha) - F(c)]z \Rightarrow z(t) = z(0)\exp((1-\alpha)rt)\exp\left(-\int_0^t F(c(\tau))d\tau\right).$$

В силу монотонного возрастания функции $F(c)$, если $c_1(\tau) \geq c_2(\tau)$ при всех $\tau \in [0, t]$, то $z_1(\tau) \leq z_2(\tau)$ на том же самом интервале.

Пусть $u_1(t) \equiv M$, а $u_2(t)$ – какое-либо релейное управление со множеством значений $\{0, M\}$. Отвечающие данным управлениям решения системы (2.3) обозначим через $(x_1(t), y_1(t), c_1(t))$ и $(x_2(t), y_2(t), c_2(t))$ соответственно:

$$z_1(t) = y_1(t) - x_1(t), \quad z_2(t) = y_2(t) - x_2(t).$$

Поскольку $z_1(t) \leq z_2(t)$, то

$$\begin{gathered}\frac{dx_1}{dt} = rx_1 + \alpha r z_1 \leq rx_1 + \alpha r z_2,\\ \frac{dx_2}{dt} = rx_2 + \alpha r z_2,\\ x_1(0) = x_2(0) = x_0.\end{gathered}$$

Отсюда следует, что

$$x_1(\tau) \leq x_2(\tau) \quad \forall \tau \in [0, t].$$

Но тогда

$$y_1(\tau) = x_1(\tau) + z_1(\tau) \leq x_2(\tau) + z_2(\tau) = y_2(\tau) \quad \forall \tau \in [0, t].$$

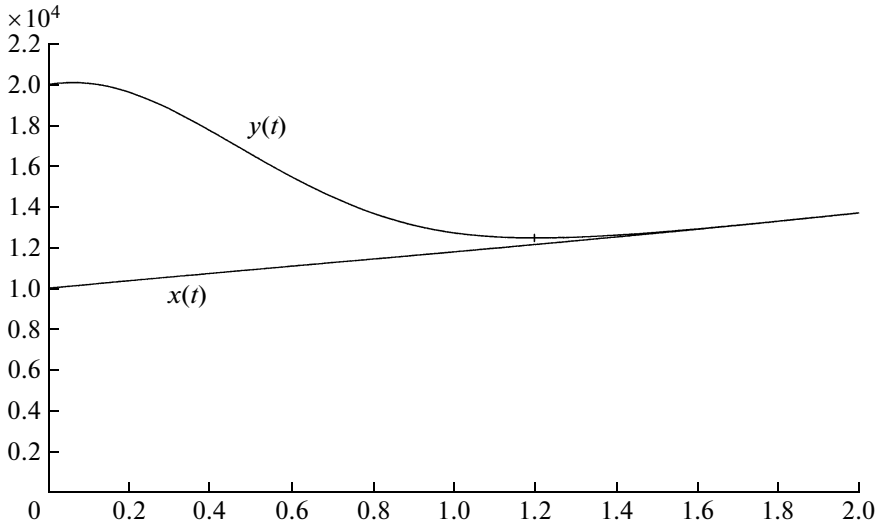
Таким образом, для всех t из $[0, +\infty)$ справедливо неравенство $y_1(t) \leq y_2(t)$. Последнее означает, что функция $u_1(t) \in M$ является оптимальным управлением.

Вернемся к изучению исходной системы (1.1)–(1.4) для двух частных случаев монотонной функции терапии $F(c)$:

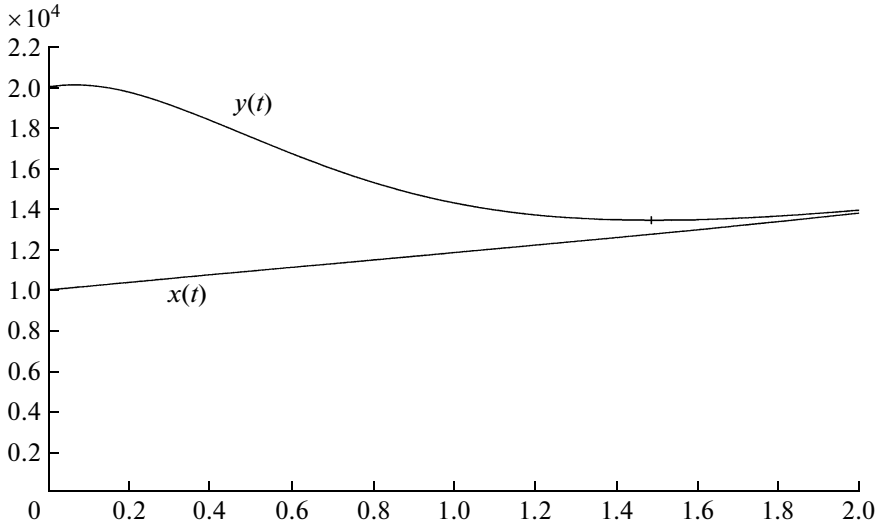
1) $F_1(c) = \beta c$;

2) $F_2(c) = \beta c/(1 + \delta c)$.

Обе функции монотонно возрастают, но $F_2(c) \longrightarrow \beta/\delta$ при $c \longrightarrow \infty$, что можно интерпретировать как достижение предельной эффективности химиотерапии, когда увеличение дозы препарата не дает большего положительного эффекта.



Фиг. 1. Функции $x(t)$ и $y(t)$ при $\alpha = 0.2$, $r = 0.15$, $\theta = 10^8$, $\gamma = 0.005$, $M = 1$, $x_0 = 10^4$, $y_0 = 2 \times 10^4$, $c_0 = 0$, $u(t) \equiv M$, $F(c) = 5c$.



Фиг. 2. Функции $x(t)$ и $y(t)$ при $\alpha = 0.2$, $r = 0.15$, $\theta = 10^8$, $\gamma = 0.005$, $M = 1$, $x_0 = 10^4$, $y_0 = 2 \times 10^4$, $c_0 = 0$, $u(t) \equiv M$, $F(c) = 5c/(1 + c)$.

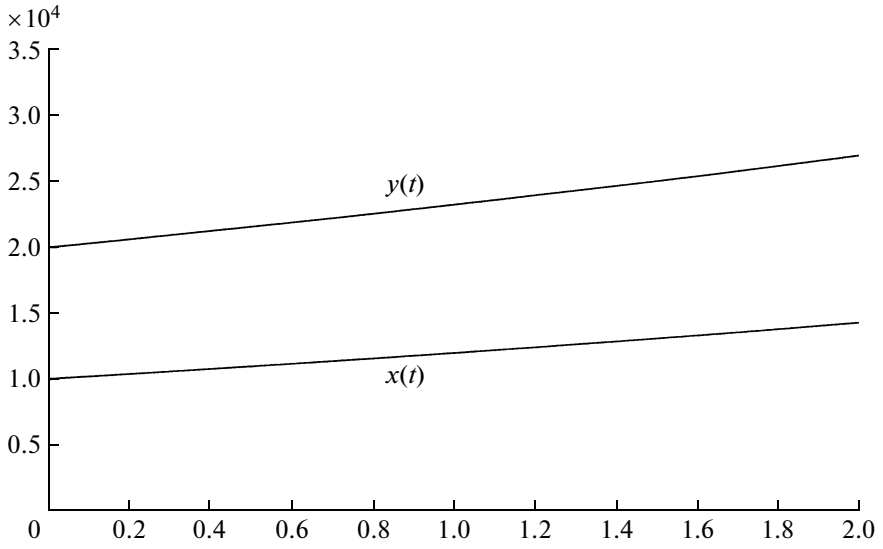
Для данного случая не удается аналитически доказать, что $u^*(t) \equiv M$, $0 \le t \le T^*$. Однако численные расчеты методом последовательных приближений подтверждают справедливость этого результата. Опишем схему применения данного метода (см. [12]).

а. Выберем некоторое допустимое начальное приближение функции управления $u^0(t)$. Интегрируя систему (1.1)–(1.4) в прямом времени методом Рунге–Кутты, находим $x^0(t)$, $y^0(t)$, $c^0(t)$. Определяем точку $T^0 = \operatorname{argmin} y^0(t)$.

б. Проинтегрируем сопряженную систему (2.1) с соответствующими краевыми условиями при $t = T^0$ в обратном времени от T^0 до 0. При этом мы будем считать, что $x = x^0$, $y = y^0$, $c = c^0$, $u = u^0$. После этого, зная функцию $\psi_3(t)$, по формуле

$$u^1 = \begin{cases} 0, & \text{если} \quad \psi_3 \le 0, \\ M, & \text{если} \quad \psi_3 > 0, \end{cases}$$

определим новое управление $u^1(t)$.



Фиг. 3. Функции $x(t)$ и $y(t)$ при $\alpha = 0.2$, $r = 0.15$, $\theta = 10^8$, $\gamma = 0.005$, $M = 1$, $x_0 = 10^4$, $y_0 = 2 \times 10^4$, $c_0 = 0$, $u(t) \equiv 0$, $F(c) = 5c/(1 + c)$.

в. Используя управление $u^1(t)$, повторяем шаги а и б и т.д.

Указанный метод уже за небольшое количество итераций приводит к функции оптимального управления $u^*(t) \equiv M$. На фиг. 1 и фиг. 2 показаны интегральные кривые системы (1.1)–(1.4) в случаях, когда $F = F_1$ и $F = F_2$ соответственно. В первом случае оптимальные значения времени окончания процесса и функционала равны

$$T^* \approx 1.2, \quad J^* \approx 12488,$$

а во втором

$$T^* \approx 1.485, \quad J^* \approx 13425.$$

(Величины T^* и J^* определяются как точка глобального минимума функции $y(t)$, соответствующей оптимальному управлению системой, значение $y(t)$ в этой точке.)

Из анализа фиг. 1 и фиг. 2 видно, что в обоих случаях траектории системы ведут себя идентично, причем $0 < x(t) < y(t)$ (утверждение леммы 1). Функция $x(t)$ монотонно растет, а функция $y(t)$ имеет четко выраженный минимум.

На фиг. 3 показаны графики функций $x(t)$ и $y(t)$ в случае, когда $c_0 = 0$ и $u(t) \equiv 0$ (т.е. когда отсутствует химиотерапевтическое воздействие на опухоль). В этом случае общая численность клеток опухоли и количество больных клеток, не поддающихся лечению, растут с экспоненциальной скоростью.

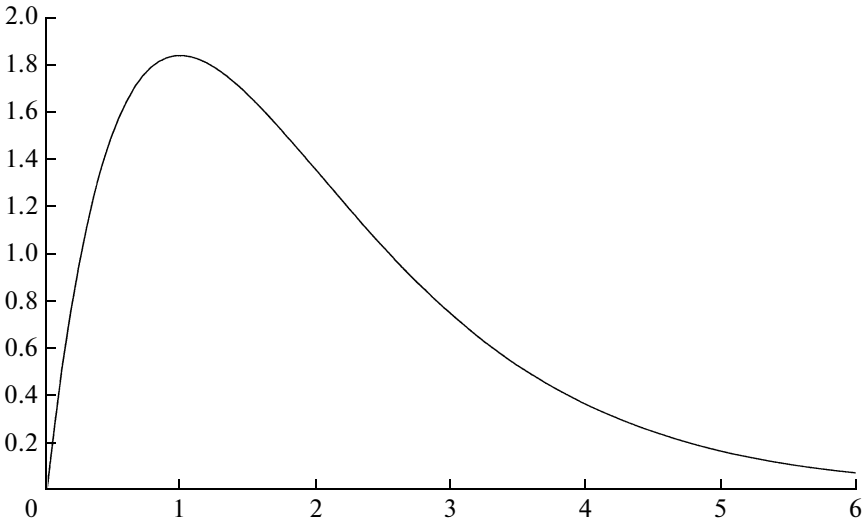
3. СЛУЧАЙ НЕМОНОТОННОЙ ФУНКЦИИ ТЕРАПИИ

Рассмотрим функцию терапии вида

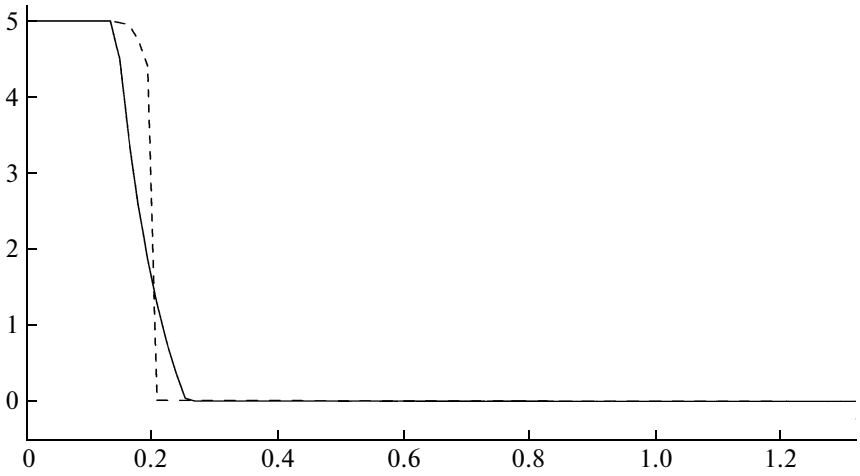
$$F(c) = \alpha c e^{-c}, \quad \sigma > 0.$$

График этой функции приведен на фиг. 4. Ее убывание при больших значениях концентрации химиотерапевтического средства можно интерпретировать как явление передозировки организма, когда избыточное количество препарата приносит вред. В этом случае требование положительности производной функции $F(c)$ не выполняется и, следовательно, несправедливы утверждения лемм 2, 3 и 4. Поэтому оптимальное управление не обязано быть релейной функцией.

На фиг. 5 изображены два графика оптимальной программы $u^*(t)$, полученных различными численными методами. Пунктир – график оптимальной программы, полученной модифицированным методом последовательных приближений, а сплошная линия – график программы, по-



Фиг. 4. Функция $F(c) = 5ce^{-c}$.



Фиг. 5. Функция $u^*(t)$ при $\alpha = 0.2$, $r = 0.15$, $\theta = 10^8$, $\gamma = 0.005$, $M = 5$, $x_0 = 10^4$, $y_0 = 2 \times 10^4$, $c_0 = 0$, $F(c) = 5ce^{-c}$.

лученной методом наискорейшего спуска. Обсудим особенности применения этих методов к поставленной задаче.

Модифицированный метод последовательных приближений отличается от уже описанного ранее метода лишь тем, что если на очередном шаге $J(u^{k+1}) \geq J(u^k)$, то процедура интегрирования системы (1.1)–(1.4) с управлением u^{k+1} заменяется интегрированием этой системы с управлением

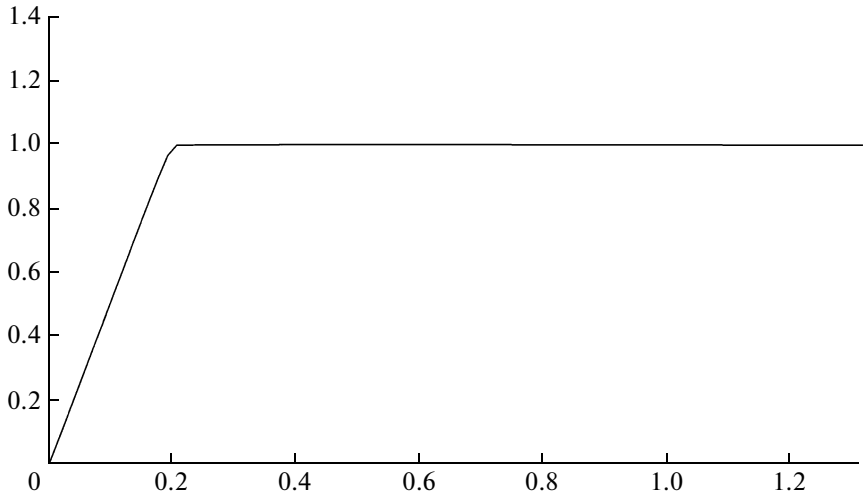
$$\tilde{u}^{k+1} = u^k + \frac{u^{k+1} - u^k}{h},$$

где h выбирается из условия $J(\tilde{u}^{k+1}) < J(u^k)$.

Указанная модификация метода последовательных приближений необходима для обеспечения его сходимости в случае нерелейных управлений.

Каждый шаг в методе наискорейшего спуска сводится к расчету очередного приближенного значения управления u по следующей формуле:

$$u^{k+1} = u^k + \kappa \frac{\partial H}{\partial u}(\psi^k, x^k, u^k),$$



Фиг. 6. Функция $c(t)$ при $\alpha = 0.2$, $r = 0.15$, $\theta = 10^8$, $\gamma = 0.005$, $M = 5$, $x_0 = 10^4$, $y_0 = 2 \times 10^4$, $c_0 = 0$, $F(c) = 5ce^{-c}$, $u = u^*$.

где H – гамильтониан системы, u^k – предыдущее приближение, x^k и ψ^k – соответствующие ему решения систем (1.1)–(1.4) и (2.1), к – шаг градиентного спуска. При этом на каждой итерации шаг к выбирается так, чтобы соответствующее управлению u^{k+1} значение функционала J было минимальным.

Если u^{k+1} не удовлетворяет наложенным на управление геометрическим ограничениям (другими словами, существуют значения t, для которых $u^{k+1}(t) \notin [0, M]$), то вместо управления u^{k+1} будем брать управление $\tilde{u}^{k+1}$ вида

$$\tilde{u}^{k+1}(t) = \begin{cases} 0, & \text{если} \quad u^{k+1}(t) < 0, \\ u^{k+1}(t), & \text{если} \quad 0 \le u^{k+1}(t) \le M, \\ M, & \text{если} \quad u^{k+1}(t) > M. \end{cases}$$

Как видно из фиг. 5, описанные выше методы приводят к очень похожим функциям управления. В частности, методом последовательных приближений были получены следующие оптимальные значения времени окончания терапии T^* и функционала J^*:

$$T^* \approx 1.335, \quad J^* \approx 13532,$$

а метод наискорейшего спуска дал результаты

$$T^* \approx 1.35, \quad J^* \approx 13539.$$

Из вида графиков на фиг. 5 следует, что оптимальная стратегия терапии заключается в следующем: до некоторого момента в ткань опухоли с максимально возможной интенсивностью вводится химиотерапевтическое средство, после чего происходит быстрое уменьшение дозы до нуля и подача препарата уже не возобновляется.

Интересно проследить, как с течением времени меняется концентрация химиотерапевтического средства в ткани опухоли при таком управлении. На фиг. 6 показан график функции $c(t)$, соответствующей найденной оптимальной программе. Сначала концентрация стремительно возрастает, но затем стабилизируется на уровне значения $\bar{c} = 1$. Легко проверить, что именно в этой точке достигает своего глобального максимума функция терапии $F(c)$. Т.е. оптимальная стратегия лечения заключается в скорейшей максимизации функции $F(c)$ и дальнейшем удержании ее значения на достигнутом максимальном уровне.

4. ЗАДАЧА С ФАЗОВЫМ ОГРАНИЧЕНИЕМ

Рассмотрим исходную задачу оптимального управления для системы (1.1)–(1.4), добавив к ее постановке следующее интегральное ограничение на фазовую переменную $c(t)$:

$$\int_0^T c(t)dt \le C,$$

где $T > 0$ – момент окончания процесса, C – заданное положительное число.

Для решения задачи введем вспомогательную функцию $g(t)$ по формуле

$$g(t) = \int_0^t c(s)ds.$$

В итоге получим следующую задачу оптимального управления:

$$\begin{gathered}
\frac{dx}{dt} = rx\left(1-\frac{y}{\theta}\right) + \alpha r\left(1-\frac{y}{\theta}\right)(y-x),\\
\frac{dy}{dt} = ry\left(1-\frac{y}{\theta}\right) - F(c)(y-x),\\
\frac{dc}{dt} = -\gamma c + u,\\
\frac{dg}{dt} = c,\\
x(0) = x_0, \quad y(0) = y_0, \quad c(0) = c_0, \quad g(0) = 0, \quad g(T) \le C,\\
T \text{ свободно}, \quad u(\cdot) \in L^{\infty}[0, T], \quad 0 \le u(t) \le M, \quad \dot\forall t \in [0, T],\\
J(u(\cdot), T) = y(T) \longrightarrow \min.
\end{gathered} \tag{4.1}$$

Введем в рассмотрение функцию штрафа

$$\phi(g(T)) = \begin{cases} 0, & \text{если} \quad g(T) \le C,\\ \lambda(g(T)-C)^2, & \text{если} \quad g(T) > C, \end{cases}$$

где λ – достаточно большое положительное число. Приближенное решение задачи оптимального управления (4.1) можно найти, решив задачу со свободным правым концом (т.е. без условия $g(T) \le C$), но с расширенным функционалом

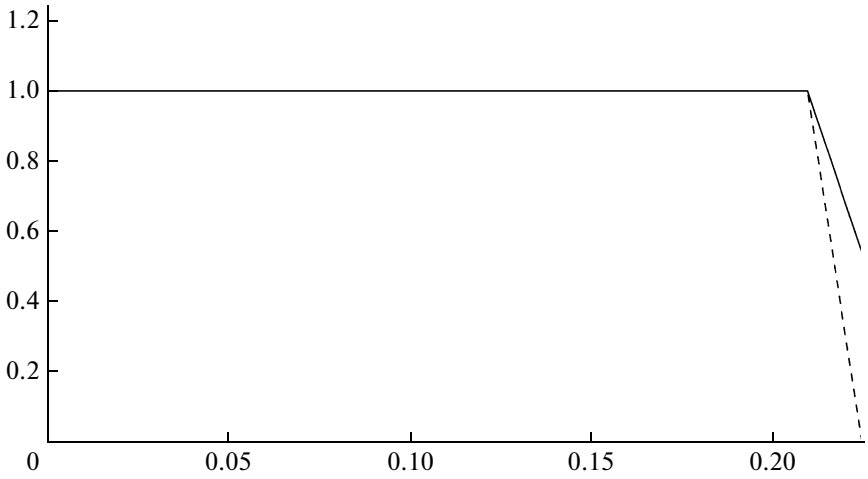
$$\bar J(u(\cdot), T) = y(T) + \phi(g(T)).$$

Запишем гамильтониан системы:

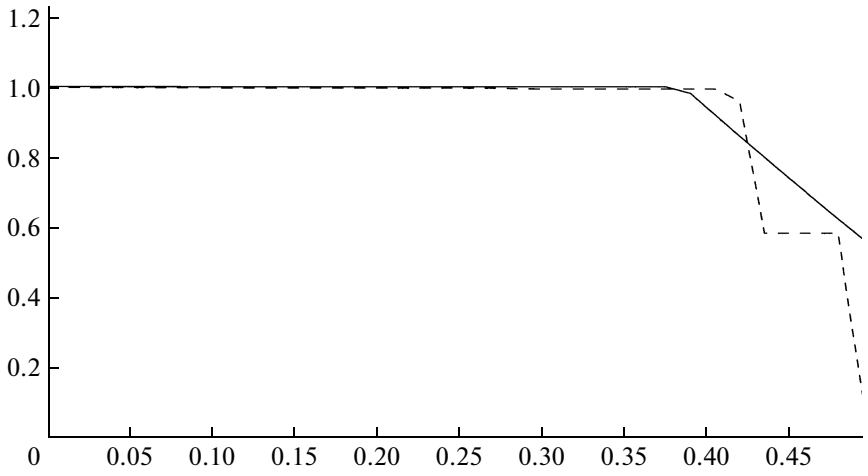
$$\begin{aligned}
H(x, y, c, g, \psi_1, \psi_2, \psi_3, \psi_4, u) = \psi_1\left(rx\left(1-\frac{y}{\theta}\right) + \alpha r\left(1-\frac{y}{\theta}\right)(y-x)\right) +\\
+ \psi_2\left(ry\left(1-\frac{y}{\theta}\right) - F(c)(y-x)\right) + \psi_3(-\gamma c + u) + \psi_4 c.
\end{aligned}$$

Сопряженная система и краевые условия для нее имеют вид

$$\begin{gathered}
\frac{d\psi_1}{dt} = -\left(r\psi_1\left(1-\frac{y}{\theta}\right)(1-\alpha) + \psi_2 F(c)\right),\\
\frac{d\psi_2}{dt} = -\left(\psi_1\left(-\frac{r}{\theta}x - \frac{r\alpha}{\theta}(y-x) + \alpha r\left(1-\frac{y}{\theta}\right)\right) + \psi_2\left(r\left(1-\frac{y}{\theta}\right) - \frac{r}{\theta}y - F(c)\right)\right),
\end{gathered}$$



Фиг. 7. Функция $u^*(t)$ при $C = 0.025$, $\lambda = 10^7$, $\alpha = 0.2$, $r = 0.15$, $\theta = 10^8$, $\gamma = 0.005$, $M = 1$, $x_0 = 10^4$, $y_0 = 2 \times 10^4$, $c_0 = 0$, $F(c) = 5c$.



Фиг. 8. Функция $u^*(t)$ при $C = 0.12$, $\lambda = 10^7$, $\alpha = 0.2$, $r = 0.15$, $\theta = 10^8$, $\gamma = 0.005$, $M = 1$, $x_0 = 10^4$, $y_0 = 2 \times 10^4$, $c_0 = 0$, $F(c) = 5c/(1 + c)$.

$$\frac{d\psi_3}{dt} = \gamma\psi_3 + \psi_2 F'(c)(y - x) - \psi_4,$$

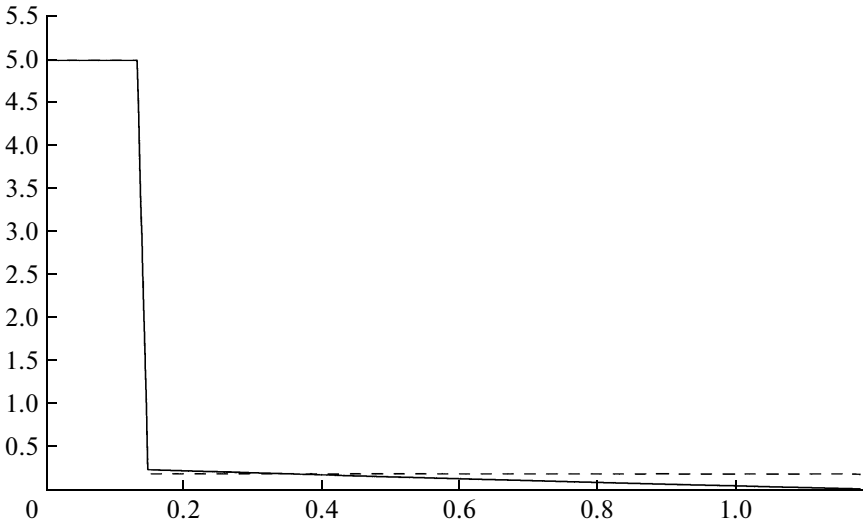
$$\frac{d\psi_4}{dt} = 0,$$

$$\psi_1(T) = 0, \quad \psi_2(T) = -1, \quad \psi_3(T) = 0, \quad \psi_4(T) = \begin{cases} 0, & \text{если} \quad g(T) \le C, \\ -2\lambda(g(T) - C), & \text{если} \quad g(T) > C. \end{cases}$$

Отсюда следует, что

$$\psi_4 \equiv \begin{cases} 0, & \text{если} \quad g(T) \le C, \\ -2\lambda(g(T) - C), & \text{если} \quad g(T) > C. \end{cases}$$

Оптимальное управление по-прежнему определяется формулой (2.2).



Фиг. 9. Функция $u^*(t)$ при $C = 0.9$, $\lambda = 10^7$, $\alpha = 0.2$, $r = 0.15$, $\theta = 10^8$, $\gamma = 0.005$, $M = 5$, $x_0 = 10^4$, $y_0 = 2 \times 10^4$, $c_0 = 0$, $F(c) = 5ce^{-c}$.

Будем решать задачу описанными выше методами последовательных приближений и наискорейшего спуска. На фиг. 7, 8 и 9 приведены графики приближенно найденных оптимальных программ для функций терапии $F_1(c) = \beta c$, $F_2(c) = \beta c/(1 + \delta c)$ и $F_3(c) = \sigma ce^{-c}$ соответственно. Пунктир – графики оптимальных программ, полученных модифицированным методом последовательных приближений, а сплошная линия – графики программ, полученных методом наискорейшего спуска. Для рассмотренных частных видов функций F_1, F_2, F_3 метод последовательных приближений дал следующие результаты:

$$T_1^* \approx 0.225, \quad \bar{J}_1^* \approx 19463,$$

$$T_2^* \approx 0.495, \quad \bar{J}_2^* \approx 17612,$$

$$T_3^* \approx 1.185, \quad \bar{J}_3^* \approx 13641.$$

Методом наискорейшего спуска получены значения

$$T_1^* \approx 0.225, \quad \bar{J}_1^* \approx 19463,$$

$$T_2^* \approx 0.495, \quad \bar{J}_2^* \approx 17616,$$

$$T_3^* \approx 1.2, \quad \bar{J}_3^* \approx 13639.$$

Из вида графиков на фиг. 7 и 8 следует, что для случая монотонной функции терапии оптимальная стратегия лечения заключается в следующем: сначала в опухоль с максимально возможной интенсивностью подается химиотерапевтическое средство, но в определенный момент начинается монотонное уменьшение дозы и заключительный этап терапии проходит при уже существенно меньшем расходе препарата.

Сравнивая графики, приведенные на фиг. 5 и 9, приходим к выводу, что в случае немонотонной функции терапии введение фазового ограничения (1.7) принципиально не меняет закон управления системой. Значит, наличие ограничения (1.7) существенно лишь для случая, когда функция терапии является монотонно возрастающей.

СПИСОК ЛИТЕРАТУРЫ

1. *Costa M.I.S., Boldrini J.L., Bassanezi R.C.* Optimal chemical control of populations developing drug resistance // IMA J. Math. Appl. Med. Biol. 1992. V. 9. P. 215–226.
2. *Costa M.I.S., Boldrini J.L., Bassanezi R.C.* Drug kinetics and drug resistance in optimal chemotherapy // Math. Biosciences. 1995. V. 125. P. 191–209.

3. *Costa M.I.S., Boldrini J.L., Bassanezi R.C.* Chemotherapeutic treatments involving drug resistance and level of normal cells as a criterion of toxicity // Math. Biosciences. 1995. V. 125. P. 211–228.
4. *Aranjo R.P., Mcelwain D.L.* A history of the study of solid tumour growth: the contribution of mathematical modeling // Bull. Math. Biol. 2004. V. 66. P. 1039–1091.
5. *Guiot C., Degiorgis P.G., Delsanto P.P. et al.* Does tumour growth follow a "universal low" // J. Theor. Biol. 2003. V. 225. P. 147–151.
6. *Byrne H.M.* A weakly nonlinear analysis of a model of avascular solid tumour growth // J. Math. Biol. 1999. V. 39. P. 151–181.
7. *Murray J.D.* Mathematical biology II: spatial models and biomedical applications. Berlin: Springer, 2003.
8. *Matzavinos A., Chaplain M., Kuznetsov V.* Mathematical modeling of the spatiotemporal response of cytotoxic T-lymphocytes to a solid tumour // Math. Medicine and Biol. 2004. V. 21. P. 1–34.
9. *Братусь А.С., Чумерина Е.С.* Синтез оптимального управления в задаче выбора лекарственного воздействия на растущую опухоль // Ж. вычисл. матем. и матем. физ. 2008. Т. 48. № 6. С. 946–966.
10. *Kirschner D., Panetta J.C.* Modelling immunotherapy of the tumour-immune interaction // J. Math. Biol. 1998. V. 37. P. 235–252.
11. *Burden T.N., Ernstberger J., Fister K.R.* Optimal control applied to immunotherapy // J. Discrete and Continuous Dynamical Systems. Ser. B. 2004. V. 4. P. 135–146.
12. *Моисеев Н.Н.* Элементы теории оптимальных систем. М.: Наука, 1975.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2009, том 49, № 11, с. 1920–1930

УДК 519.622

ПРОСТОЙ СПОСОБ ПОСТРОЕНИЯ ДВУХШАГОВЫХ МЕТОДОВ РУНГЕ–КУТТЫ

© 2009 г. Л. М. Скворцов

(105005 Москва, ул. 2-я Бауманская, 5, МГТУ им. Н.Э. Баумана)

e-mail: lm_skvo@rambler.ru

Поступила в редакцию 12.12.2008 г.
Переработанный вариант 04.05.2009 г.

Предложен способ построения двухшаговых методов Рунге–Кутты на основе одношаговых методов. Рассматриваются явные и диагонально неявные двухшаговые методы, имеющие второй или третий стадийный порядок. На тестовых задачах показано преимущество предложенных методов в сравнении с обычными одношаговыми методами. Библ. 15. Табл. 5.

Ключевые слова: двухшаговые методы Рунге–Кутты, стадийный порядок, явные методы, диагонально-неявные методы, жесткие системы уравнений.

1. ВВЕДЕНИЕ

Двухшаговые методы Рунге–Кутты (TSRK, см. [1]–[7]) обобщают обычные одношаговые методы, используя на очередном шаге интегрирования информацию, полученную не только на текущем, но и на предыдущем шаге. Благодаря этому двухшаговые методы могут иметь более высокий стадийный порядок, что позволяет повысить их точность, особенно при решении жестких и дифференциально-алгебраических задач. Общий класс двухшаговых методов Рунге–Кутты предложен в [1]. Мы рассмотрим один частный подкласс этого класса, отличающийся простотой построения таких методов и простотой их программной реализации.

Будем решать задачу Коши для системы обыкновенных дифференциальных уравнений (ОДУ)

$$\mathbf{y}' = \mathbf{f}(t, \mathbf{y}), \quad \mathbf{y}(t_0) = \mathbf{y}_0, \tag{1.1}$$

где $\mathbf{y}$ – вектор переменных, $\mathbf{f}$ – векторная функция, t – независимая переменная.

Рассмотрим двухшаговые методы Рунге–Кутты, задаваемые формулами

$$\mathbf{Y}_n^1 = \mathbf{y}_n, \quad \mathbf{F}_n^1 = \mathbf{f}(t_n, \mathbf{Y}_n^1), \tag{1.2а}$$

$$\mathbf{Y}_n^i = \mathbf{y}_n + h_n \sum_{j=1}^{s+1} (a_{ij}\mathbf{F}_{n-1}^j + b_{ij}\mathbf{F}_n^j), \quad \mathbf{F}_n^i = \mathbf{f}(t_n + c_i h_n, \mathbf{Y}_n^i), \quad i = 2, 3, \ldots, s, \tag{1.2б}$$

$$\mathbf{Y}_n^{s+1} = \mathbf{y}_n + h_n \sum_{j=1}^{s+1} b_{s+1,j}\mathbf{F}_n^j, \quad \mathbf{F}_n^{s+1} = \mathbf{f}(t_n + h_n, \mathbf{Y}_n^{s+1}), \quad \mathbf{y}_{n+1} = \mathbf{Y}_n^{s+1}, \tag{1.2в}$$

где h_n – размер очередного шага, а $\mathbf{Y}_n^i$ и $\mathbf{F}_n^i$, $i = 1, 2, \ldots, s+1$, – стадийные значения и их производные на этом шаге. Будем называть стадии (1.2б) *внутренними*, а стадию (1.2в) *заключительной*. Отметим, что стадия (1.2а) не требует вычислений, поскольку результат ее выполнения совпадает с результатом выполнения заключительной стадии предыдущего шага. Поэтому говорят, что такие методы обладают свойством FSAL (First Same As Last). Представим коэффициенты метода (1.2) в виде двух матриц и вектора

$$\mathbf{A} = [a_{ij}], \quad \mathbf{B} = [b_{ij}], \quad \mathbf{c} = [c_i], \quad i, j = 1, 2, \ldots, s+1,$$

при этом примем $c_1 = 0$, $c_{s+1} = 1$, $a_{1i} = a_{s+1,i} = b_{1i} = 0$, $i = 1, 2, ..., s+1$.

В предлагаемом способе построения двухшаговых методов вида (1.2) за основу берется обычный одношаговый метод Рунге–Кутты, задаваемый матрицей **B** и вектором **c**. Назовем его *исходным методом*. Матрицу **A** примем в виде

$$\mathbf{A} = \mathbf{d}\mathbf{g}^{\mathrm{T}}, \tag{1.3}$$

где векторы **d** и **g** определяются исходя из условий повышения стадийного порядка исходного метода и обеспечения необходимых свойств устойчивости. Такой подход позволяет повысить стадийный порядок исходного метода на 1 или на 2. В общем случае все коэффициенты двухшаговых методов зависят от соотношения размеров шагов $w = h_n/h_{n-1}$. В наших методах от w зависят только коэффициенты вектора **g**, что упрощает их реализацию с переменным шагом. Первый шаг выполняется исходным одношаговым методом, поэтому предложенные методы не нуждаются в специальной стартовой процедуре.

С учетом (1.3) формулы (1.2) можно записать в виде, более удобном для их реализации:

$$\mathbf{Y}_n^1 = \mathbf{y}_n, \quad \mathbf{F}_n^1 = \mathbf{f}(t_n, \mathbf{Y}_n^1), \quad \mathbf{u} = \sum_{j=1}^{s+1} g_j \mathbf{F}_{n-1}^j,$$

$$\mathbf{Y}_n^i = \mathbf{y}_n + h_n\left(d_i\mathbf{u} + \sum_{j=1}^{s+1} b_{ij}\mathbf{F}_n^j\right), \quad \mathbf{F}_n^i = \mathbf{f}(t_n + c_i h_n, \mathbf{Y}_n^i), \quad i = 2, 3, \ldots, s, \tag{1.4}$$

$$\mathbf{Y}_n^{s+1} = \mathbf{y}_n + h_n\sum_{j=1}^{s+1} b_{s+1,j}\mathbf{F}_n^j, \quad \mathbf{F}_n^{s+1} = \mathbf{f}(t_n + h_n, \mathbf{Y}_n^{s+1}), \quad \mathbf{y}_{n+1} = \mathbf{Y}_n^{s+1}.$$

2. УСЛОВИЯ ПОРЯДКА

Вывод условий порядка основывается на сравнении разложений в ряд Тейлора точного и численного решений. Часто используют предположения, позволяющие упростить условия порядка. При этом важным понятием является стадийный порядок, определяемый как наименьший порядок на всех стадиях. Условия порядка двухшаговых методов Рунге–Кутты рассматривались в [1]–[3]. Основываясь на этих работах, приведем некоторые условия порядка для методов вида (1.2) и (1.4).

Примем следующие обозначения $(s+1)$-мерных векторов:

$$\mathbf{e} = [1, \ldots, 1]^{\mathrm{T}}, \quad \mathbf{e}_{s+1} = [0, \ldots, 0, 1]^{\mathrm{T}}, \quad \mathbf{b} = [b_{s+1,1}, \ldots, b_{s+1,s+1}]^{\mathrm{T}}.$$

Предположим, что стадийный порядок исходного метода равен $\bar{q}$, т.е. выполняются условия

$$k\mathbf{B}\mathbf{c}^{k-1} = \mathbf{c}^k, \quad k = 1, 2, \ldots, \bar{q}, \quad (\bar{q}+1)\mathbf{B}\mathbf{c}^{\bar{q}} \neq \mathbf{c}^{\bar{q}+1} \tag{2.1}$$

(здесь и далее предполагается покомпонентное выполнение операции возведения вектора в степень). Примем также обычное для методов Рунге–Кутты предположение, что $\bar{q} \geq 1$, тогда $\mathbf{c} = \mathbf{B}\mathbf{e}$ и метод однозначно задается матрицей **B**. Пусть $\bar{p}$ – порядок исходного метода, причем будем предполагать, что $\bar{p} > \bar{q}$.

Рассмотрим двухшаговый метод (1.2), и пусть он имеет стадийный порядок q. Тогда должны выполняться равенства

$$k\left[\mathbf{A}\left(\frac{\mathbf{c}-\mathbf{e}}{w}\right)^{k-1} + \mathbf{B}\mathbf{c}^{k-1}\right] = \mathbf{c}^k, \quad k = 1, 2, \ldots, q. \tag{2.2}$$

Предполагая, что $q > \bar{q}$, и подставляя (1.3) в (2.2), при $k = \bar{q} + 1$ получаем

$$(\bar{q}+1)\mathbf{d}\mathbf{g}^{\mathrm{T}}\left(\frac{\mathbf{c}-\mathbf{e}}{w}\right)^{\bar{q}} = \mathbf{c}^{\bar{q}+1} - (\bar{q}+1)\mathbf{B}\mathbf{c}^{\bar{q}}.$$

Это равенство будет выполняться, если принять

$$\mathbf{d} = \mathbf{c}^{\bar{q}+1} - (\bar{q}+1)\mathbf{B}\mathbf{c}^{\bar{q}}, \tag{2.3}$$

$$\mathbf{g}^{\mathrm{T}}(\mathbf{c}-\mathbf{e})^{\bar{q}} = \frac{w^{\bar{q}}}{\bar{q}+1}. \tag{2.4}$$

Из (2.1), (2.2) имеем

$$\mathbf{g}^{\mathrm{T}}\mathbf{c}^{k} = 0, \quad k = 0, 1, \ldots, \bar{q}-1, \tag{2.5}$$

а из (2.4), (2.5) получим

$$\mathbf{g}^{\mathrm{T}}\mathbf{c}^{\bar{q}} = \frac{w^{\bar{q}}}{\bar{q}+1}. \tag{2.6}$$

Равенства (2.3), (2.5), (2.6) будем использовать как условия, обеспечивающие стадийный порядок $q = \bar{q} + 1$ метода (1.4). Для обеспечения стадийного порядка $q = \bar{q} + 2$ дополнительно должны выполняться условия

$$\mathbf{c}^{\bar{q}+2} - (\bar{q}+2)\mathbf{B}\mathbf{c}^{\bar{q}+1} = \alpha(\mathbf{c}^{\bar{q}+1} - (\bar{q}+1)\mathbf{B}\mathbf{c}^{\bar{q}}), \tag{2.7}$$

$$\mathbf{g}^{\mathrm{T}}\mathbf{c}^{\bar{q}+1} = \alpha\frac{w^{\bar{q}+1}}{\bar{q}+2} + w^{\bar{q}}, \tag{2.8}$$

где α – некоторая константа.

Если q – стадийный порядок метода (1.2), то его порядок (не ниже) $p = q$. Метод (1.2) имеет порядок $p = q + 1$, если

$$(q+1)\mathbf{b}^{\mathrm{T}}\mathbf{c}^{q} = 1. \tag{2.9}$$

Метод (1.2) имеет порядок $p = q + 2$, если выполняются условия (2.9) и

$$(q+2)\mathbf{b}^{\mathrm{T}}\mathbf{c}^{q+1} = 1, \tag{2.10}$$

$$(q+2)(q+1)\mathbf{b}^{\mathrm{T}}\left[\mathbf{A}\left(\frac{\mathbf{c}-\mathbf{e}}{w}\right)^{q} + \mathbf{B}\mathbf{c}^{q}\right] = 1. \tag{2.11}$$

Если $\bar{p} > \bar{q} + 1$, то (2.11) можно упростить. В этом случае из условий порядка одношагового метода имеем

$$\mathbf{b}^{\mathrm{T}}\mathbf{d} = \mathbf{b}^{\mathrm{T}}(\mathbf{c}^{\bar{q}+1} - (\bar{q}+1)\mathbf{B}\mathbf{c}^{\bar{q}}) = 0.$$

Подставляя это выражение и (1.3) в (2.11), получаем

$$(q+2)(q+1)\mathbf{b}^{\mathrm{T}}\mathbf{B}\mathbf{c}^{q} = 1. \tag{2.12}$$

Таким образом, метод (1.4) имеет порядок $p = q + 2$, если $\bar{p} > \bar{q} + 1$ и выполняются условия (2.9), (2.10), (2.12).

3. УСТОЙЧИВОСТЬ

Устойчивость методов решения ОДУ принято исследовать на модельном уравнении Далквиста $y' = \lambda y$, $y(t_0) = y_0$. Используя формулы (1.2) для решения этого уравнения, получаем

$$\mathbf{Y}_n = (\mathbf{e}\mathbf{e}_{s+1}^{\mathrm{T}} + z\mathbf{A})\mathbf{Y}_{n-1} + z\mathbf{B}\mathbf{Y}_n, \quad z = h_n\lambda. \tag{3.1}$$

Разрешая (3.1) относительно $\mathbf{Y}_n$ и учитывая (1.3), имеем

$$\mathbf{Y}_n = \mathbf{H}(z)\mathbf{Y}_{n-1}, \quad \mathbf{H}(z) = (\mathbf{I} - z\mathbf{B})^{-1}(\mathbf{e}\mathbf{e}_{s+1}^{\mathrm{T}} + z\mathbf{d}\mathbf{g}^{\mathrm{T}}), \tag{3.2}$$

где $\mathbf{I}$ – единичная матрица. Устойчивость разностного уравнения (3.2) определяется спектром матрицы $\mathbf{H}(z)$, т.е. корнями характеристического многочлена $P(\eta, z) = |\eta\mathbf{I} - \mathbf{H}(z)|$.

В общем случае $P(\eta, z)$, как многочлен от η, имеет $s + 1$ различных корней, зависящих от z, но при выборе матрицы $\mathbf{A}$ в виде (1.3) ранг матрицы $\mathbf{H}(z)$ равен 2, а в этом случае

$$P(\eta, z) = \eta^{s-1}[\eta^2 - p_1(z)\eta + p_0(z)]. \tag{3.3}$$

Таким образом, устойчивость предлагаемых методов определяется двумя нулями

$$\eta_{1,2}(z) = \frac{1}{2}(p_1(z) \pm \sqrt{p_1^2(z) - 4p_0(z)}) \tag{3.4}$$

многочлена (3.3). Согласно определению V.1.1 из [8], область устойчивости метода задается неравенствами

$$|\eta_1(z)| \le 1, \quad |\eta_2(z)| \le 1, \tag{3.5}$$

при этом оба нуля не должны одновременно равняться 1 или -1.

Чтобы получить выражения для $p_0(z)$, $p_1(z)$, представим матрицу $\mathbf{H}(z)$ в виде

$$\mathbf{H}(z) = \mathbf{X}\mathbf{Y}^{\mathrm{T}}, \quad \mathbf{X} = (\mathbf{I} - z\mathbf{B})^{-1}[\mathbf{e}\ z\mathbf{d}], \quad \mathbf{Y} = [\mathbf{e}_{s+1}\ \mathbf{g}].$$

Воспользовавшись тем, что матрицы $\mathbf{H}(z)$ и $\mathbf{V}(z) = \mathbf{Y}^{\mathrm{T}}\mathbf{X}$ имеют одинаковые ненулевые собственные значения, получим

$$\begin{gathered} p_1(z) = \nu_{11}(z) + \nu_{22}(z), \quad p_0(z) = \nu_{11}(z)\nu_{22}(z) - \nu_{12}(z)\nu_{21}(z), \\ \nu_{11}(z) = \mathbf{e}_{s+1}^{\mathrm{T}}(\mathbf{I} - z\mathbf{B})^{-1}\mathbf{e}, \quad \nu_{12}(z) = \mathbf{e}_{s+1}^{\mathrm{T}}(\mathbf{I} - z\mathbf{B})^{-1}z\mathbf{d}, \\ \nu_{21}(z) = \mathbf{g}^{\mathrm{T}}(\mathbf{I} - z\mathbf{B})^{-1}\mathbf{e}, \quad \nu_{22}(z) = \mathbf{g}^{\mathrm{T}}(\mathbf{I} - z\mathbf{B})^{-1}z\mathbf{d}. \end{gathered} \tag{3.6}$$

Заметим, что $\nu_{11}(z)$ является функцией устойчивости исходного одношагового метода. Соотношения (3.4)–(3.6) использовались нами для построения областей устойчивости двухшаговых методов.

4. ЯВНЫЕ МЕТОДЫ

Явные методы имеют $b_{ij} = 0$ при $j \ge i$. В этом случае исходный метод имеет первый стадийный порядок $\bar{q} = 1$ и может быть представлен в виде таблицы Бутчера

$$\begin{array}{c|ccc} 0 & & & \\ c_2 & b_{21} & & \\ \vdots & \vdots & \ddots & \\ c_s & b_{s1} & \dots & b_{s,s-1} \\ \hline & b_{s+1,1} & \dots & b_{s+1,s-1}\ b_{s+1,s} \end{array}$$

где s – число стадий, равное числу вычислений правой части на одном шаге интегрирования. В соответствии с (2.3) принимаем

$$\mathbf{d} = \mathbf{c}^2 - 2\mathbf{B}\mathbf{c}, \tag{4.1}$$

а из (2.5), (2.6) получаем

$$\mathbf{g}^{\mathrm{T}}\mathbf{e} = 0, \quad \mathbf{g}^{\mathrm{T}}\mathbf{c} = w/2. \tag{4.2}$$

Двухшаговые методы 2-го стадийного порядка можно построить на основе любого одношагового метода, порядок которого не ниже 2-го. Для этого достаточно задать $\mathbf{d}$ по формуле (4.1), а $\mathbf{g}$ определить из (4.2). При решении уравнений (4.2) используются только два коэффициента вектора $\mathbf{g}$, остальные можно задать нулевыми либо выбрать из других соображений. Например, приравняв нулю коэффициенты $g_2, \ldots, g_s$, получим

$$\mathbf{g} = \frac{w}{2}[-1, 0, \ldots, 0, 1]^{\mathrm{T}}. \tag{4.3}$$

Формулы (4.1), (4.3) задают наиболее простой способ построения явных двухшаговых методов.

Приведем конкретный метод. При $s = 2$ оптимальный по точности одношаговый метод 2-го порядка получаем, задав $c_2 = 2/3$. Построенный на его основе двухшаговый метод задается коэффициентами

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 \\ 2/3 & 0 & 0 \\ 1/4 & 3/4 & 0 \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} 0 \\ 2/3 \\ 1 \end{bmatrix}, \quad \mathbf{d} = \begin{bmatrix} 0 \\ 4/9 \\ 0 \end{bmatrix}, \quad \mathbf{g} = \frac{w}{2}\begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$$

и имеет порядок $p = 3$ и стадийный порядок $q = 2$. Длина его области устойчивости вдоль вещественной оси (при постоянном шаге) $l = 2.0$. Обозначим этот метод через TSRK23 (первая цифра s, вторая p).

Двухшаговые методы 3-го стадийного порядка можно построить на основе одношаговых методов, коэффициенты которых удовлетворяют соотношению (2.7) при $\alpha = c_2$. В этом случае из (2.8) получаем

$$\mathbf{g}^{\mathrm{T}}\mathbf{c}^2 = w\left(1 + \frac{wc_2}{3}\right). \tag{4.4}$$

Вектор $\mathbf{d}$ определяем согласно (4.1), а $\mathbf{g}$ находим из (4.2), (4.4). При $s > 2$ число коэффициентов вектора $\mathbf{g}$ больше, чем необходимо для решения уравнений (4.2), (4.4). Поэтому мы задавали ненулевыми только три коэффициента этого вектора: g_1, g_k и g_{s+1}, а k выбирали так, чтобы область устойчивости была как можно больше. Решение уравнений (4.2), (4.4) относительно выбранных ненулевых коэффициентов имеет вид

$$g_1 = \frac{w}{6}\left(\frac{3 + 2wc_2}{c_k} - 3\right), \quad g_k = \frac{w}{6}\frac{(3 + 2wc_2)}{c_k(c_k - 1)}, \quad g_{s+1} = \frac{w}{6}\left(\frac{3 + 2wc_2}{1 - c_k} + 3\right).$$

Примем $s = 3$ и построим на основе одношагового метода 3-го порядка двухшаговый метод 4-го порядка, имеющий 3-й стадийный порядок. В этом случае исходный метод задается абсциссами c_2 и c_3. Из условия 4-го порядка заключительной стадии имеем

$$\begin{vmatrix} c_2 & c_3 & 1/2 \\ c_2^2 & c_3^2 & 1/3 \\ c_2^3 & c_3^3 & 1/4 \end{vmatrix} = 0, \tag{4.5}$$

а из (2.7) и условия 3-го порядка исходного метода получаем

$$c_2^2(c_3^3 - 3b_{32}c_2^2) - c_2^3(c_3^2 - 2b_{32}c_2) = 0, \quad b_{32} = \frac{2 - 3c_2}{6c_3(c_3 - c_2)}. \tag{4.6}$$

Уравнения (4.5), (4.6) имеют единственное решение $c_2 = 1/2$, $c_3 = 1$. Задав $g_3 = 0$, получим коэффициенты двухшагового метода

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 \\ -1 & 2 & 0 & 0 \\ 1/6 & 2/3 & 1/6 & 0 \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} 0 \\ 1/2 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{d} = \begin{bmatrix} 0 \\ 1/4 \\ -1 \\ 0 \end{bmatrix}, \quad \mathbf{g} = \frac{w}{6}\begin{bmatrix} 3 + 2w \\ -12 - 4w \\ 0 \\ 9 + 2w \end{bmatrix},$$

который обозначим через TSRK34. Длина области устойчивости этого метода $l = 2.372$.

При построении двухшаговых методов 3-го стадийного порядка в качестве исходного удобно выбрать метод, коэффициенты которого удовлетворяют условиям

$$2\sum_{j=2}^{i-1} b_{ij}c_j = c_i^2, \quad 3\sum_{j=2}^{i-1} b_{ij}c_j^2 = c_i^3, \quad i = 3, 4, \ldots, s + 1. \tag{4.7}$$

В этом случае достаточно скорректировать только 2-ю стадию, чтобы повысить стадийный порядок до 3-го. Поэтому вектор $\mathbf{d}$ имеет только один ненулевой коэффициент $d_2 = c_2^2$, что упрощает реализацию такого метода.

Примем $s = 4$ и воспользуемся соотношениями (4.7) для построения метода 4-го порядка. Задав также $g_2 = g_4 = 0$, получим метод TSRK44 с коэффициентами

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 0 & 0 \\ 1/8 & 3/8 & 0 & 0 & 0 \\ 1/2 & -3/2 & 2 & 0 & 0 \\ 1/6 & 0 & 2/3 & 1/6 & 0 \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} 0 \\ 1/3 \\ 1/2 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{d} = \begin{bmatrix} 0 \\ 1/9 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{g} = \frac{w}{18}\begin{bmatrix} 9 + 4w \\ 0 \\ -36 - 8w \\ 0 \\ 27 + 4w \end{bmatrix}.$$

Длина области устойчивости этого метода $l = 2.479$.

Аналогичный подход был применен при построении двухшагового метода 5-го порядка TSRK65 на основе метода Дорманда–Принса, коэффициенты которого приведены в [9]. В этом случае мы получили ненулевые коэффициенты векторов $\mathbf{d}$ и $\mathbf{g}$ в виде

$$d_2 = \frac{1}{25}, \quad g_1 = \frac{w}{18}(21 + 4w), \quad g_3 = -\frac{10w}{63}(15 + 2w), \quad g_7 = \frac{w}{42}(51 + 4w).$$

Длина области устойчивости полученного метода $l = 3.011$ (для исходного метода Дорманда–Принса $l = 3.307$).

Построенные методы были испытаны на ряде нежестких и умеренно жестких задач, при этом они оказывались, как правило, более эффективными, чем обычные методы Рунге–Кутты. Приведем некоторые результаты. Для сравнения приводим также результаты метода Богацки–Шампайна 3-го порядка (см. [10]) и метода Дорманда–Принса, которые обозначим через RK33 и RK65. Эти методы считаются лучшими явными методами Рунге–Кутты низкой и средней точности. В качестве показателя точности использовалось значение

$$\mathrm{scd} = -\lg\left(\max_i\left(\left|\frac{y_i - \tilde{y}_i}{y_i}\right|\right)\right), \tag{4.8}$$

Таблица 1

Метод	μ				
	1	10	20	40	80
RK33	4.91	4.59	4.34	4.03	3.78
TSRK23	5.39	5.40	5.38	5.29	4.96
TSRK34	6.72	7.24	7.08	6.73	6.78
RK65	7.70	6.07	5.17	4.04	–
TSRK65	9.13	8.28	7.52	6.49	–

Таблица 2

Метод	$\mathrm{Rtol} = 10^{-3}$		$\mathrm{Rtol} = 10^{-4}$		$\mathrm{Rtol} = 10^{-6}$		$\mathrm{Rtol} = 10^{-8}$	
	scd	Nf	scd	Nf	scd	Nf	scd	Nf
RK33	3.35	175	3.97	166	5.53	439	6.88	1243
TSRK23	3.68	109	4.46	125	6.04	221	7.83	843
TSRK34	4.88	176	6.34	182	7.77	171	9.90	496
RK65	4.02	211	5.29	253	6.47	535	8.83	1279
TSRK65	4.30	223	5.36	223	6.88	229	8.93	343

Таблица 3

Метод	$\mathrm{Rtol} = 10^{-3}$		$\mathrm{Rtol} = 10^{-4}$		$\mathrm{Rtol} = 10^{-6}$		$\mathrm{Rtol} = 10^{-8}$	
	scd	Nf	scd	Nf	scd	Nf	scd	Nf
RK33	0.74	856	1.36	1273	3.22	4927	5.19	21433
TSRK23	1.58	731	1.76	1135	3.88	4635	5.82	21347
TSRK34	0.14	630	2.23	914	4.72	2161	6.81	6324
RK65	0.53	979	1.93	1339	3.79	2647	5.67	5815
TSRK65	1.07	949	1.79	1039	4.62	1873	6.22	3301

где y_i – точное, а $\tilde{y}_i$ – численное решения по i-й компоненте в конечной точке интервала интегрирования. Вычислительные затраты оценивались числом вычислений правой части Nf.

При реализации методов с автоматическим выбором размера шага применялась вспомогательная формула, согласно которой на каждом шаге вычисляется вектор $\hat{\mathbf{y}}_{n+1}$. Оценку ошибки получаем как норму вектора $\hat{\mathbf{y}}_{n+1} - \mathbf{y}_{n+1}$. В методе TSRK23 принимали $\hat{\mathbf{y}}_{n+1} = \mathbf{y}_n + \frac{h}{2}(\mathbf{f}_n + \mathbf{f}_{n+1})$, а в методах TSRK34 и TSRK44 задавали $\hat{\mathbf{y}}_{n+1} = \mathbf{Y}_n^s$. В методе TSRK65 для вычисления $\hat{\mathbf{y}}_{n+1}$ применялась та же формула, что и в методе Дорманда–Принса. Во всех методах использовалась стандартная процедура управления длиной шага (см. [8]). Методы TSRK34 и TSRK44 показали близкие результаты, поэтому приводим результаты только первого из них.

Для исследования влияния жесткости на точность численного решения использовалась задача Капса

$$y_1' = -(\mu + 2)y_1 + \mu y_2^2, \quad y_2' = y_1 - y_2 - y_2^2, \\ y_1(0) = 1, \quad y_2(0) = 1, \quad 0 \le t \le 1, \tag{4.9}$$

решение которой $y_1(t) = e^{-2t}$, $y_2(t) = e^{-t}$ не зависит от параметра жесткости μ. Задача решалась при различных значениях μ с шагом $h = s/120$, что соответствует 120 вычислениям правой части. Результаты (значения scd) приведены в табл. 1, где прочерк соответствует расхождению численного решения. Эта же задача при $\mu = 100$ решалась с автоматическим выбором размера шага. Мы задавали допустимую относительную ошибку Rtol и принимали допустимую абсолютную ошибку как Atol = 0.01 × Rtol. Результаты при начальном шаге $h_0 = 0.01$ приведены в табл. 2.

Нежесткая задача BRUS из [9] решалась при Atol = 0.01 × Rtol и $h_0 = 0.01$, а результаты ее решения приведены в табл. 3. Результаты тестирования показали, что преимущество двухшаговых методов наиболее заметно при решении умеренно жестких задач и повышенных требованиях к точности.

5. ДИАГОНАЛЬНО НЕЯВНЫЕ МЕТОДЫ

Если в исходном методе задать $b_{ii} = \gamma > 0$, $i = 2, 3, \ldots, s+1$, и $b_{ij} = 0$ при $j > i$, то получим диагонально неявные методы Рунге–Кутты с явной первой стадией (методы ESDIRK, см. [11]–[14]), которые имеют таблицу Бутчера вида

$$\begin{array}{c|cccc} 0 & 0 & & & \\ c_2 & b_{21} & \gamma & & \\ \vdots & \vdots & \vdots & \ddots & \\ c_s & b_{s1} & b_{s2} & \ldots & \gamma \\ 1 & b_{s+1,1} & b_{s+1,2} & \ldots & b_{s+1,s}\ \gamma \\ \hline & b_{s+1,1} & b_{s+1,2} & \ldots & b_{s+1,s}\ \gamma \end{array} \qquad (5.1)$$

(в [12], [13] эти методы названы FSAL-DIRK). Формально метод (5.1) является $(s+1)$-стадийным, но реально на каждом шаге выполняются только s неявных стадий, поскольку явная стадия не требует вычислений. Среди неявных методов такие методы наиболее просто реализуются, однако их стадийный порядок не может быть выше второго, что препятствует их эффективному применению при высоких требованиях к точности.

Пусть исходный метод имеет 2-й стадийный порядок, т.е. удовлетворяет условиям (2.1) при $\bar{q} = 2$. В соответствии с (2.3) принимаем

$$\mathbf{d} = \mathbf{c}^3 - 3\mathbf{B}\mathbf{c}^2, \qquad (5.2)$$

а из (2.5), (2.6) получаем

$$\mathbf{g}^{\mathrm{T}}\mathbf{e} = 0, \quad \mathbf{g}^{\mathrm{T}}\mathbf{c} = 0, \quad \mathbf{g}^{\mathrm{T}}\mathbf{c}^2 = w^2/3. \qquad (5.3)$$

Двухшаговый метод 3-го стадийного порядка можно построить на основе любого метода ESDIRK 2-го стадийного порядка, порядок которого не ниже 3-го. Для этого достаточно задать $\mathbf{d}$ в виде (5.2), а коэффициенты вектора $\mathbf{g}$ определить так, чтобы выполнялись равенства (5.3).

Дополнительно потребуем, чтобы метод обладал $L(\alpha)$-устойчивостью. Для этого необходимо, чтобы при $z \longrightarrow \infty$ и любом w для коэффициентов характеристического многочлена (3.3) выполнялось $p_1(z) \longrightarrow 0$, $p_0(z) \longrightarrow 0$. Представим матрицу $\mathbf{B}$ в виде $\mathbf{B} = \left[\begin{array}{c|c} 0 & 0\ldots0 \\ \hline \tilde{\mathbf{b}} & \tilde{\mathbf{B}} \end{array}\right]$ и примем $\hat{\mathbf{B}} = \left[\begin{array}{c|c} 1 & 0\ldots0 \\ \hline \tilde{\mathbf{b}} & \tilde{\mathbf{B}} \end{array}\right]$, тогда

$$\hat{\mathbf{B}}^{-1} = \left[\begin{array}{c|c} 1 & 0\ldots0 \\ \hline -\tilde{\mathbf{B}}^{-1}\tilde{\mathbf{b}} & \tilde{\mathbf{B}}^{-1} \end{array}\right], \quad (\mathbf{I} - z\mathbf{B})^{-1} = \left[\begin{array}{c|c} 1 & 0\ldots0 \\ \hline (\tilde{\mathbf{I}} - z\tilde{\mathbf{B}})^{-1}z\tilde{\mathbf{b}} & (\tilde{\mathbf{I}} - z\tilde{\mathbf{B}})^{-1} \end{array}\right],$$

Таблица 4

Метод \ μ	10^0	10^1	10^2	10^3	10^4	10^5
ESDIRK54	5.82	5.53	5.20	5.93	6.75	7.07
TSDIRK54	6.05	6.17	6.43	6.90	7.20	7.25

Таблица 5

h	ESDIRK54		TSDIRK54	
	scd(y)	scd(z)	scd(y)	scd(z)
1/20	4.05	3.31	5.00	5.14
1/50	5.14	4.17	6.54	6.07
1/100	6.02	4.81	7.72	6.89
1/200	6.91	5.44	8.91	7.74

где $\tilde{\mathbf{I}}$ – единичная матрица размера $s \times s$. Учитывая, что $d_1 = c_1 = 0$, имеем

$$\begin{aligned} \lim_{z\to\infty}(\mathbf{I}-z\mathbf{B})^{-1}\mathbf{e} &= \hat{\mathbf{B}}^{-1}\mathbf{e}_1, \quad \mathbf{e}_1 = [1, 0, \ldots, 0]^{\mathrm{T}}, \\ \lim_{z\to\infty}(\mathbf{I}-z\mathbf{B})^{-1}z\mathbf{d} &= -\hat{\mathbf{B}}^{-1}\mathbf{d} = \hat{\mathbf{B}}^{-1}(3\mathbf{B}\mathbf{c}^2-\mathbf{c}^3) = 3\mathbf{c}^2-\hat{\mathbf{B}}^{-1}\mathbf{c}^3, \end{aligned} \tag{5.4}$$

а подставляя (5.4), (5.3) в (3.6), получаем

$$\begin{aligned} \nu_{11}(\infty) &= \mathbf{e}_{s+1}^{\mathrm{T}}\hat{\mathbf{B}}^{-1}\mathbf{e}_1, \quad \nu_{12}(\infty) = 3-\mathbf{e}_{s+1}^{\mathrm{T}}\hat{\mathbf{B}}^{-1}\mathbf{c}^3, \\ \nu_{21}(\infty) &= \mathbf{g}^{\mathrm{T}}\hat{\mathbf{B}}^{-1}\mathbf{e}_1, \quad \nu_{22}(\infty) = w^2-\mathbf{g}^{\mathrm{T}}\hat{\mathbf{B}}^{-1}\mathbf{c}^3. \end{aligned} \tag{5.5}$$

Теперь, используя формулы (5.5) и представление характеристического многочлена в виде (3.6), получаем в удобном виде необходимые условия $L(\alpha)$-устойчивости. Из условия $p_1(\infty) = 0$ получаем

$$\mathbf{e}_{s+1}^{\mathrm{T}}\hat{\mathbf{B}}^{-1}\mathbf{e}_1 = 0, \tag{5.6а}$$

$$\mathbf{g}^{\mathrm{T}}\hat{\mathbf{B}}^{-1}\mathbf{c}^3 = w^2, \tag{5.6б}$$

а чтобы было также и $p_0(\infty) = 0$, дополнительно должно удовлетворяться одно из следующих равенств:

$$\mathbf{e}_{s+1}^{\mathrm{T}}\hat{\mathbf{B}}^{-1}\mathbf{c}^3 = 3, \tag{5.7а}$$

$$\mathbf{g}^{\mathrm{T}}\hat{\mathbf{B}}^{-1}\mathbf{e}_1 = 0. \tag{5.7б}$$

Таким образом, необходимым условием $L(\alpha)$-устойчивости двухшагового метода является выполнение равенств (5.6) и одного из равенств (5.7).

Для одношаговых методов равенство (5.6а) является необходимым условием $L(\alpha)$-устойчивости, а условие (5.7а) обеспечивает повышение точности при решении жестких задач. В [14] приведены процедуры построения методов ESDIRK 4-го и 5-го порядков, удовлетворяющих этим условиям. Такие методы целесообразно использовать также и в качестве исходных при построении двухшаговых методов. В этом случае коэффициенты вектора $\mathbf{g}$ определяются из условий (5.3), (5.6б), а выполнение равенства (5.7б) является излишним.

С помощью приведенных в [14] формул для расчета коэффициентов методов ESDIRK, а также условий (5.3), (5.6б) были построены двухшаговые методы 4-го порядка. Наиболее удобный для реализации метод задается коэффициентами

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 1/4 & 1/4 & 0 & 0 & 0 & 0 \\ 55/196 & 2/49 & 1/4 & 0 & 0 & 0 \\ 17/96 & 7/12 & -49/96 & 1/4 & 0 & 0 \\ 5/48 & 13/24 & -49/48 & 9/8 & 1/4 & 0 \\ 1/6 & 0 & 0 & 2/3 & -1/12 & 1/4 \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} 0 \\ 1/2 \\ 4/7 \\ 1/2 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{d} = \begin{bmatrix} 0 \\ -1/16 \\ -61/686 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{g} = \frac{w^2}{6}\begin{bmatrix} 4 \\ 0 \\ 0 \\ -8 \\ 1 \\ 3 \end{bmatrix}$$

и является $L(\alpha)$-устойчивым при $\alpha = 89.58°$. Обозначим этот метод через TSDIRK54.

Приведем результаты решения с постоянным шагом трех тестовых задач. Для сравнения приводим результаты "оптимального" одношагового метода ESDIRK54, у которого при $\gamma = 1/4$ семь из девяти коэффициентов погрешности 5-го порядка равны 0 (остальные два коэффициента погрешности принципиально не могут быть нулевыми). Коэффициенты метода ESDIRK54 приведены в [13, формула (3.1)].

В качестве первого теста использовалась задача Капса (4.9), которая решалась при $h = 1/12$ и различных значениях параметра жесткости μ. Результаты (значения scd) приведены в табл. 4. При всех значениях μ двухшаговый метод имеет преимущество, наиболее заметное при умеренной жесткости ($\mu = 10^2$).

Второй тест (жесткая задача PLATE из [8]) содержит 80 уравнений и имеет комплексный спектр матрицы Якоби. При интегрировании на интервале [0, 7] с шагом $h = 1/24$ мы получили scd = 4.18 для метода ESDIRK54 и scd = 5.59 для метода TSDIRK54.

Двухшаговый метод оказался более точным также и при решении дифференциально-алгебраических уравнений (ДАУ) индексов 2 и 3. В качестве третьего теста выбрана система ДАУ индекса 2 из [15]:

$$y_1' = y_1 y_2^2 z^2, \quad y_2' = y_1^2 y_2^2 - 3y_2^2 z,$$

$$0 = y_1^2 y_2 - 1, \quad y_1(0) = y_2(0) = z(0) = 1, \quad 0 \le t \le 1.$$

Точное решение этих уравнений $y_1(t) = e^t$, $y_2(t) = e^{2t}$, $z(t) = e^{-2t}$. Результаты при различных значениях размера шага приведены в табл. 5, где scd(y) и scd(z) – значения (4.8), вычисленные, соответственно, для дифференциальных переменных y_1, y_2 и алгебраической переменной z.

Приведенные результаты показывают убедительное преимущество двухшагового метода, которое объясняется его более высоким стадийным порядком по сравнению с одношаговым методом.

СПИСОК ЛИТЕРАТУРЫ

1. *Jackiewich Z., Tracogna S.* A general class of two-step Runge–Kutta methods for ordinary differential equations // SIAM J. Numer. Analys. 1995. V. 32. № 5. P. 1390–1427.
2. *Butcher J.C., Tracogna S.* Order conditions for two-step Runge–Kutta methods // Appl. Numer. Math. 1997. V. 24. № 2–3. P. 351–364.
3. *Hairer E., Wanner G.* Order conditions for general two-step Runge–Kutta methods // SIAM J. Numer. Analys. 1997. V. 34. № 6. P. 2087–2089.
4. *Bartoszewski Z., Jackiewicz Z.* Construction of two-step Runge–Kutta methods of high order for ordinary differential equations // Numer. Algorithms. 1998. V. 18. № 1. P. 51–70.
5. *Tracogna S., Welfert B.* Two-step Runge–Kutta: theory and practice // BIT. 2000. V. 40. № 4. P. 775–799.
6. *Jackiewicz Z., Verner J.H.* Derivation and implementation of two-step Runge–Kutta pairs // Japan J. Ind. Appl. Math. 2002. V. 19. P. 227–248.
7. *Chollom J., Jackiewicz Z.* Construction of two-step Runge–Kutta methods with large regions of absolute stability // J. Comput. Appl. Math. 2003. V. 157. № 1. P. 125–137.
8. *Хайрер Э., Ваннер Г.* Решение обыкновенных дифференциальных уравнений. Жесткие и дифференциально-алгебраические задачи. М.: Мир, 1999.

9. *Хайрер Э., Нёрсетт С., Ваннер Г.* Решение обыкновенных дифференциальных уравнений. Нежесткие задачи. М.: Мир, 1990.
10. *Bogacki P., Shampine L.F.* A 3(2) pair of Runge–Kutta formulas // Appl. Math. Letts. 1989. V. 2. № 4. P. 321–325.
11. *Kværnø A.* Singly diagonally implicit Runge–Kutta methods with an explicit first stage // BIT. 2004. V. 44. № 3. P. 489–502.
12. *Скворцов Л.М.* Диагонально неявные FSAL-методы Рунге–Кутты для жестких и дифференциально-алгебраических систем // Матем. моделирование. 2002. Т. 14. № 2. С. 3–17.
13. *Скворцов Л.М.* Точность методов Рунге–Кутты при решении жестких задач // Ж. вычисл. матем. и матем. физ. 2003. Т. 43. № 9. С. 1374–1384.
14. *Скворцов Л.М.* Диагонально неявные методы Рунге–Кутты для жестких задач // Ж. вычисл. матем. и матем. физ. 2006. Т. 46. № 12. С. 2209–2222.
15. *Jay L.* Convergence of a class of Runge–Kutta methods for differential-algebraic systems of index 2 // BIT. 1993. V. 33. № 1. P. 137–150.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2009, том 49, № 11, с. 1931–1936

УДК 519.634

О РЕШЕНИИ КРАЕВЫХ ЗАДАЧ ДЛЯ УРАВНЕНИЯ ЛАПЛАСА НА КУСОЧНО-ОДНОРОДНОЙ ПЛОСКОСТИ С ПАРАБОЛИЧЕСКОЙ ТРЕЩИНОЙ (ЗАВЕСОЙ)

© 2009 г. С. Е. Холодовский

(672007 Чита, ул. Бабушкина, 129, ЗабГГПУ)

e-mail: hol47@yandex.ru

Поступила в редакцию 30.03.2009 г.
Переработанный вариант 13.05.2009 г.

Рассмотрены краевые задачи для уравнения Лапласа на кусочно-однородной плоскости, разделенной на две зоны сильно проницаемой трещиной или слабо проницаемой завесой в виде параболы, когда искомые потенциалы имеют заданные особые точки (источники, стоки и т.д.). Выведены формулы, выражающие искомые потенциалы через гармонические функции, имеющие заданные особые точки и описывающие аналогичные процессы на однородной плоскости. Библ. 11.

Ключевые слова: задачи математической физики в кусочно-однородных средах, параболическая трещина (завеса), метод свертывания разложений Фурье, уравнение Лапласа.

ВВЕДЕНИЕ

В связи с широким применением в технике композитных материалов, в том числе материалов с сильно и слабопроницаемыми пленками (трещинами и завесами), большой интерес имеют краевые задачи тепломассопереноса в кусочно-однородных средах с обобщенными условиями сопряжения на заданных линиях. Трещины и завесы, кроме уточненной модели неидеальных контактов разнородных сред, моделируют искусственные экраны, дренажи, проводники, которые используются для управления потоками.

В [1]–[3] разработан эффективный метод, позволяющий по известным решениям стандартных классических краевых задач строить решения серии усложненных задач с дополнительными условиями сопряжения, с более сложными уравнениями и граничными условиями и т.д. В данной статье указанный метод распространяется на криволинейные трещины и завесы в виде параболы, разделяющей кусочно-однородные зоны. В [4] рассмотрены аналогичные задачи при идеальном контакте двух сред (без трещины и завесы), при этом применяется метод отражения особых точек. В [5], [6] решены краевые задачи на плоскости с разрезами при неоднородных условиях сопряжения на сторонах разрезов, при этом задачи сводятся к решению системы интегральных уравнений Фредгольма II рода. Следует отметить, что наличие пленочных включений в силу малого их раскрытия и экстремальной проницаемости в них делает малоэффективным применение численных методов.

1. ПОСТАНОВКА ЗАДАЧИ

Рассмотрим установившиеся процессы теплопроводности, фильтрации, диффузии, электростатики на плоскости, разделенной параболической трещиной или завесой $x = ay^2 - b$ на зоны $D_1(x > ay^2 - b)$ и $D_2(x < ay^2 - b)$ постоянной проницаемости k_j в D_j, когда процесс индуцируется особыми точками (источниками, стоками и т.д.), заданными в D_j, где $a = (4l^2)^{-1}$, $b = l^2$; x, y – декартовы координаты.

Для вывода обобщенных условий сопряжения на трещине (завесе) рассмотрим параболические координаты ξ, η:

$$x = \xi^2 - \eta^2, \quad y = 2\xi\eta, \quad z = \zeta^2, \quad z = x + iy, \quad \zeta = \xi + i\eta, \quad \eta \geq 0, \tag{1}$$

$$\xi = \operatorname{sign}(y)\sqrt{\frac{\sqrt{x^2+y^2}+x}{2}}, \quad \eta = \sqrt{\frac{\sqrt{x^2+y^2}-x}{2}}, \tag{2}$$

где $\operatorname{sign}(\pm 0) = \pm 1$, при этом D_1 $(0 < \eta < l)$, D_2 $(l < \eta < \infty)$, $\xi \in \mathbb{R}$. Заменим трещину (завесу) слоем D_0 $(l < \eta < l + h)$ проницаемости k_0 при выполнении классических условий сопряжения на ∂D_0: $u_0 = u_j$, $k_0 \partial_\eta u_0 = k_j \partial_\eta u_j$ при $\eta = \eta_j$, где $j = 1, 2$; $\eta_1 = l$, $\eta_2 = l + h$, u_i – потенциалы в D_i. Отсюда, переходя к пределу при $h \longrightarrow 0$, $k_0 \longrightarrow \infty$, $k_0 h \longrightarrow A_2$ в случае трещины и при $h \longrightarrow 0$, $k_0 \longrightarrow 0$, $h k_0^{-1} \longrightarrow$ $\longrightarrow A_1 (k_1 k_2)^{-1}$ в случае завесы, аналогично работе [3] получаем условия сопряжения вида

$$\eta = l\colon u_2 - u_1 = A_1 k_2^{-1} \partial_\eta u_1, \quad k_2 \partial_\eta u_2 - k_1 \partial_\eta u_1 = A_2 \partial_{\eta\eta} u_1, \tag{3}$$

где одна из постоянных A_j равна нулю, т.е. имеет место либо трещина (при $A_1 = 0$), либо завеса (при $A_2 = 0$), $\partial_\eta = \partial/\partial\eta$, $\partial_{\eta\eta} = \partial^2/\partial\eta^2$.

Отметим, что краевые задачи вне криволинейных разрезов конечной длины на плоскости с неоднородным граничным условием (3) для уравнений Лапласа и Гельмгольца при $k_1 = k_2$ решены в [7], [8] в случае $A_2 = 0$ и в [9], [10] в случае, когда $A_1 = 0$ и вместо $\partial_{\eta\eta} u_1$ стоит u_1 (или u_2).

Пусть на плоскости ζ с декартовыми координатами ξ, η (1) определена гармоническая функция $f(\xi, \eta)$, которая имеет заданные особые точки в D_i и считается известной (функция $f(\xi, \eta)$ является потенциалом данного течения на однородной плоскости ζ). Далее предполагаем, что функция $f(\xi, \eta)$ в бесконечности может иметь особенность не выше полюса произвольного порядка, а в конечных точках – произвольные особенности.

Для потенциалов $u_i(\xi, \eta)$ в параболических координатах ξ, η задача имеет вид

$$\Delta u_i = 0, \quad (\xi, \eta) \in D_i, \quad i = 1, 2, \tag{4}$$

$$u_1(\xi, 0) = u_1(-\xi, 0), \quad \partial_\eta u_1(\xi, 0) = -\partial_\eta u_1(-\xi, 0) \tag{5}$$

при условиях сопряжения (3), причем в окрестности особых точек

$$u_i \sim f(\xi, \eta) \tag{6}$$

(уравнение (4) выполняется вне особых точек). Условия (5) выражают непрерывность потенциала и нормальной скорости на разрезе $L(x > 0, y = 0)$ плоскости течения z (см. (1) при $\eta = 0$), т.е. этим разрезом можно пренебречь.

В предельном случае при $l \longrightarrow 0$ параболическое включение D_1 вырождается в луч $L(x > 0, y = 0)$, т.е. имеют место динамические процессы на однородной плоскости проницаемости k с трещиной или завесой в виде луча L. В данном случае для потенциала $u(\xi, \eta)$ задача имеет вид

$$\Delta u = 0, \quad 0 < \eta < \infty, \quad u \sim f(\xi, \eta), \tag{7}$$

$$u(\xi, 0) - u(-\xi, 0) = 2A_1 \partial_\eta u(\xi, 0), \quad \partial_\eta u(\xi, 0) + \partial_\eta u(-\xi, 0) = 2A_2 \partial_{\eta\eta} u(\xi, 0), \tag{8}$$

где одна из постоянных A_j равна нулю. Условия (8) выводятся аналогично условиям (3) посредством замены трещины (завесы) слоем $D_0 (0 < \eta < h)$ проницаемости k_0 при выполнении условий сопряжения на ∂D_0 вида $u_0 = u$, $k_0 \partial_\eta u_0 = k \partial_\eta u$ при $\eta = h$ и $u_0(\xi, 0) = u_0(-\xi, 0)$, $\partial_\eta u_0(\xi, 0) = -\partial_\eta u_0(-\xi, 0)$ на разрезе $\eta = 0$ с последующим переходом к пределу при $h \longrightarrow 0$, $k_0 \longrightarrow \infty$, $k_0 h \longrightarrow A_2 k$ в случае трещины и при $h \longrightarrow 0$, $k_0 \longrightarrow 0$, $h k_0^{-1} \longrightarrow A_1 k^{-1}$ в случае завесы.

2. ОСНОВНЫЕ РЕЗУЛЬТАТЫ

2.1. Пусть особые точки гармонической функции $f(\xi, \eta)$ расположены в D_2, т.е. в условии (6) $i = 2$. В данном случае имеет место обтекание заданным потоком параболического включения D_1, экранированного трещиной или завесой. Выведем формулы, непосредственно выражающие ре-

шение задачи (3)–(6) через заданную функцию $f(\xi, \eta)$. Для вывода этих формул, следуя методу работ [1]–[3], в качестве промежуточного используем метод Фурье, который справедлив для более узкого класса функций $f(\xi, \eta)$ (на бесконечности должно выполняться необходимое условие $f \longrightarrow 0$). Полагая, что $f(\xi, l)$ разлагается в интеграл Фурье, представляем функцию $f(\xi, \eta)$ в полуплоскости $\eta \le l$, где она не имеет особых точек, в виде

$$f(\xi, \eta) = \int_0^\infty e^{\lambda(\eta - l)} f_i \sigma_i d\lambda, \tag{9}$$

где $\sigma_1 = \sin(\lambda\xi)$, $\sigma_2 = \cos(\lambda\xi)$, f_i – коэффициенты Фурье функции $f(\xi, l)$, по повторяющимся в одной части равенства индексам $i = 1, 2$ суммируем. Формула (9) выражает решение задачи Дирихле в полуплоскости $\eta \le l$ с граничной функцией $f(\xi, l)$, полученное методом Фурье. Умножая разложение функции $f(\xi, \eta - t + l)$ из (9) на $e^{-\gamma t} t^p$ и интегрируя по $t \in (0, \infty)$, с учетом формулы 2.3.3 (2) из [11] получаем равенство

$$\frac{1}{p!} \int_0^\infty e^{-\gamma t} t^p f(\xi, \eta - t + l) dt = \int_0^\infty \frac{e^{\lambda\eta} f_i \sigma_i}{(\lambda + \gamma)^{p+1}} d\lambda, \quad \gamma > 0, \quad \eta < 0, \quad p = 0, 1, \ldots. \tag{10}$$

Представим решение задачи (3)–(6) в виде разложений Фурье:

$$u_1 = \int_0^\infty [a_1 \sigma_1 \operatorname{sh}(\lambda\eta) + a_2 \sigma_2 \operatorname{ch}(\lambda\eta)] d\lambda, \quad u_2 = f(\xi, \eta) + \int_0^\infty e^{\lambda(l - \eta)} b_i \sigma_i d\lambda, \tag{11}$$

при этом функции (11) удовлетворяют уравнению (4) и условиям (5), (6) (при условии сходимости и дифференцируемости интегралов (11)). Из условий сопряжения (3) с учетом (9) находим $a_i = k_2 f_i d_i$, $b_i = (-1)^j (2k_j g_{ij} d_i - 1) f_i$, где $g_{i1} = c_i$, $g_{i2} = s_i$, $s_1 = c_2 = \operatorname{sh}(\lambda l)$, $s_2 = c_1 = \operatorname{ch}(\lambda l)$,

$$\begin{gathered} d_i = \frac{2e^{-\lambda l}}{A_j(\lambda + \gamma)[1 + (-1)^{i+j} q]}, \quad q = e^{-2\lambda l}\left(1 - \frac{\mu_\theta}{\lambda + \gamma}\right), \\ \gamma = \frac{k_1 + k_2}{A_j}, \quad \mu_i = \frac{2k_i}{A_j}, \end{gathered} \tag{12}$$

$j = 2$, $\theta = 1$ в случае трещины и $j = 1$, $\theta = 2$ в случае завесы $\eta = l$, при этом $|q| < 1$ при $0 \le \lambda < \infty$. Раскладывая дроби $[1 + (-1)^{i+j} q]^{-1}$ из (12) в геометрическую прогрессию и выделяя в потенциалах (11) выражения (10), окончательно решение задачи (3)–(6) приводим к виду

$$\begin{aligned} u_1 &= \mu_2 \sum_{n=0}^\infty \sum_{p=0}^n T_{np} \int_0^\infty e^{-\gamma t} t^p [f((-1)^n \xi, \eta - 2nl - t) + f((-1)^{n+1} \xi, -\eta - 2nl - t)] dt, \\ u_2 &= f(\xi, \eta) - (-1)^j f(\xi, -\eta + 2l) + \mu_j \sum_{n=0}^\infty \sum_{p=0}^n T_{np} \int_0^\infty e^{-\gamma t} t^p [f((-1)^{n+1} \xi, -\eta - 2nl - t) + \\ &\quad + (-1)^j f((-1)^n \xi, -\eta - 2l(n-1) - t)] dt, \end{aligned} \tag{13}$$

где $j = 2$, $T_{np} = (-1)^n (-\mu_1)^p (p!)^{-1} C_n^p$ в случае трещины и $j = 1$, $T_{np} = (-\mu_2)^p (p!)^{-1} C_n^p$ в случае завесы, C_n^p – биномиальные коэффициенты; ξ, η имеют вид (2).

В случае идеального контакта зон D_i при $A_j = 0$, $j = 1, 2$ (без трещины и завесы), аналогично получаем

$$u_1 = (1-\nu)\sum_{n=0}^{\infty}\nu^n[f((-1)^n\xi, \eta - 2nl) + f((-1)^{n+1}\xi, -\eta - 2nl)],$$
$$u_2 = f(\xi, \eta) - \nu f(\xi, -\eta + 2l) + (1-\nu^2)\sum_{n=0}^{\infty}\nu^n f((-1)^{n+1}\xi, -\eta - 2nl), \quad (14)$$

где $\nu = (k_1 - k_2)(k_1 + k_2)^{-1}$.

Формулы (13), (14), с одной стороны, проще формул (11) и формул работы [4]. Именно, решения (13) содержат по одной квадратуре от монотонных на бесконечности функций, в то время, как решения (11), полученные методом Фурье, содержат двукратные квадратуры от сильно осциллирующих функций. Последнее вызывает большие трудности при применении численных методов. С другой стороны, формулы из (13), (14) справедливы для более широкого класса функций $f(\xi, \eta)$. Можно показать, что для функции $f(\xi, \eta)$, имеющей в бесконечности полюс произвольного порядка, функции u_i из (13), (14) и их производные приводятся к сходящимся рядам типа $\sum_{n=0}^{\infty}|\nu|^n n^r$, где $r > 0$, $|\nu| < 1$ (14).

Например, пусть включение D_1 экранировано трещиной и $f(\xi, \eta) = \xi^2 - \eta^2$, что на плоскости z соответствует поступательному потоку вдоль оси x (вдоль оси симметрии параболического включения D_1). Из формул (13) потенциалы выражаются в конечном виде $u_1 = x$, $u_2 = x - 2k_2^{-1}[A_2 + l(k_1 - k_2)](\eta - l)$, где η имеет вид (2). Отсюда внутри включения D_1 имеет место поступательный поток. Вне включения, т.е. в D_2, наличие трещины в однородной среде ($A_2 > 0$, $k_1 = k_2$) и наличие более проницаемого включения D_1 без трещины ($A_2 = 0$, $k_1 > k_2$) приводит к увеличению компоненты скорости v_η, при этом трещина и более проницаемое включение "притягивают" поток. В указанных случаях при $A_2 = l(k_1 - k_2)$ влияние трещины (в однородной среде) и более проницаемого включения (без трещины) идентичны. Менее проницаемое включение D_1 без трещины ($A_2 = 0$, $k_1 < k_2$) уменьшает компоненту скорости v_η, т.е. "отталкивает" поток. В случае менее проницаемого включения D_1, экранированного трещиной, когда $A_2 = l(k_2 - k_1)$, указанные эффекты компенсируют друг друга, при этом на всей плоскости имеет место поступательный поток $u_1 = u_2 = x$.

В предельном случае трещины (завесы) в виде луча $L(x > 0, y = 0)$, представляя решение задачи (7), (8) в виде

$$u = f(\xi, \eta) + \int_0^{\infty} e^{-\lambda\eta} b_i \sigma_i d\lambda,$$

из условий (8) с учетом равенства (9) при $l = 0$ находим $b_1 = -f_1$, $b_2 = [2\gamma(\lambda + \gamma)^{-1} - 1]f_2$, $\gamma = A_2^{-1}$, в случае трещины и $b_1 = [1 - 2\gamma(\lambda + \gamma)^{-1}]f_1$, $b_2 = f_2$, $\gamma = A_1^{-1}$, в случае завесы L. Отсюда с учетом формулы (10) решение задачи (7), (8) в случае завесы $j = 1$ и трещины $j = 2$ получаем в виде

$$u = f(\xi, \eta) - (-1)^j f(\xi, -\eta) + \gamma\int_0^{\infty} e^{-\gamma t}[f(-\xi, \eta - t) + (-1)^j f(\xi, \eta - t)]dt.$$

2.2. Пусть особые точки гармонической функции $f(\xi, \eta)$ расположены внутри включения D_1, т.е. в условии (6) $i = 1$. Полагая, что функция $f(\xi, l) + f(-\xi, -l)$ разлагается в интеграл Фурье с коэффициентами Фурье f_i, представляем гармоническую функцию

$$F(\xi, \eta) = f(\xi, \eta) + f(-\xi, -\eta) \quad (15)$$

в полуплоскости $\eta \geq l$ (где функция F не имеет особых точек) в виде

$$F(\xi, \eta) = \int_0^{\infty} e^{\lambda(l-\eta)} f_i \sigma_i d\lambda. \quad (16)$$

Представляя решение задачи (3)–(6) в виде

$$u_1 = F(\xi, \eta) + \int_0^\infty [a_1\sigma_1 \mathrm{sh}(\lambda\eta) + a_2\sigma_2 \mathrm{ch}(\lambda\eta)]d\lambda, \quad u_2 = \int_0^\infty e^{\lambda(l-\eta)} b_i\sigma_i d\lambda, \tag{17}$$

из условий сопряжения (3) с учетом (16) находим $a_i = [k_1 - k_2 - (-1)^j A_j\lambda]f_i d_i$, $b_i = k_1 e^{l\lambda} f_i d_i$, где $j = 2$ и $j = 1$, соответственно, в случае трещины и завесы $\eta = l$; d_i определены в (12). Раскладывая дроби $[1 + (-1)^{i+j}q]^{-1}$ из (12) в ряды и выделяя в (17) выражения (10), приводим потенциалы (17) к виду

$$\begin{gathered} u_1 = F(\xi, \eta) + \sum_{n=1}^{\infty}\sum_{p=0}^{n} T_{np}[\Phi_p((-1)^n\xi, \eta + 2nl) + \Phi_p((-1)^{n+1}\xi, -\eta + 2nl)], \\ u_2 = \mu_1 \sum_{n=0}^{\infty}\sum_{p=0}^{n} T_{np}\Phi_{p+1}((-1)^n\xi, \eta + 2nl), \end{gathered} \tag{18}$$

где $T_{np} = (-1)^n(-\mu_1)^p C_n^p$ в случае трещины и $T_{np} = (-\mu_2)^p C_n^p$ в случае завесы,

$$\Phi_p(\xi, \eta) = \frac{1}{(p-1)!}\int_0^\infty e^{-\gamma t} t^{p-1} F(\xi, \eta + t)dt, \quad p = 1, 2, \ldots,$$

$\Phi_0(\xi, \eta) = F(\xi, \eta)$, $\eta > l$; γ, μ_i определены в (12), $F(\xi, \eta)$ – заданная гармоническая функция (15).

При отсутствии трещины и завесы ($A_1 = A_2 = 0$), когда особые точки функции $f(\xi, \eta)$ расположены внутри включения D_1, аналогично решение задачи (3)–(6) получаем в виде

$$\begin{gathered} u_1 = F(\xi, \eta) + \sum_{n=1}^{\infty} \nu^n [F((-1)^n\xi, \eta + 2nl) + F((-1)^{n+1}\xi, -\eta + 2nl)], \\ u_2 = (1 + \nu)\sum_{n=0}^{\infty} \nu^n F((-1)^n\xi, \eta + 2nl), \end{gathered} \tag{19}$$

где ν определена в (14).

Если особые точки, индуцирующие процесс, расположены в обеих зонах D_i, то потенциалы имеют вид суммы соответствующих функций (13), (14) и (18), (19).

СПИСОК ЛИТЕРАТУРЫ

1. *Холодовский С.Е.* Метод эффективного решения краевых задач с обобщенными условиями сопряжения // Обозрение прикл. и пром. матем. 2006. Т. 13. Вып. 6. С. 1128–1130.
2. *Холодовский С.Е.* Метод свертывания разложений Фурье в решении краевых задач с пересекающимися линиями сопряжения // Ж. вычисл. матем. и матем. физ. 2007. Т. 47. № 9. С. 1550–1556.
3. *Холодовский С.Е.* Метод рядов Фурье для решения задач в кусочно-неоднородных средах с прямолинейной трещиной (завесой) // Ж. вычисл. матем. и матем. физ. 2008. Т. 48. № 7. С. 1209–1213.
4. *Голубева О.В., Шпилевой А.Я.* О плоской фильтрации в средах с прерывно изменяющейся проницаемостью вдоль кривых второго порядка // Изв. АН СССР. Механ. жидкости и газа. 1967. № 2. С. 174–179.
5. *Крутицкий П.А., Сгибнев А.И.* Метод интегральных уравнений в обобщенной задаче о скачке для уравнения Лапласа вне разрезов на плоскости // Дифференц. ур-ния. 2002. Т. 38. № 9. С. 1199–1213.
6. *Крутицкий П.А., Прозоров К.В.* Обобщенная задача о скачке для уравнения Гельмгольца вне разрезов на плоскости // Дифференц. ур-ния. 2004. Т. 40. № 9. С. 1176–1189.
7. *Krutitskii P.A., Chikilev A.J., Krutitskaya N.Ch., Kolybasova V.V.* On a generalization of the Neumann problem for the Laplace equation outside cuts in a plane // Math. Meth. Appl. Sci. 2005. V. 28. P. 593–606.

8. *Крутицкий П.А., Колыбасова В.В.* Обобщение задачи Неймана для уравнения Гельмгольца вне разрезов на плоскости // Дифференц. ур-ния. 2005. Т. 41. № 9. С. 1155–1165.
9. *Krutitskii P.A.* The modified jump problem for the Laplace equation and singularities at the tips // J. Comput. Appl. Math. 2005. V. 183. P. 232–240.
10. *Krutitskii P.A.* The modified jump problem for the Helmholtz equation // Ann. dell Univ. Ferrara, Sez. VII, Sci. Matem. 2001. V. XLVII. P. 285–296.
11. *Прудников А.П., Брычков Ю.А., Маричев О.И.* Интегралы и ряды. М.: Наука, 1981.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2009, том 49, № 11, с. 1937–1952

УДК 519.633

ПЕРВЫЕ МОМЕНТНЫЕ ФУНКЦИИ РЕШЕНИЯ УРАВНЕНИЯ ТЕПЛОПРОВОДНОСТИ СО СЛУЧАЙНЫМИ КОЭФФИЦИЕНТАМИ

© 2009 г. В. Г. Задорожний, С. С. Хребтова

(394006 Воронеж, Университетская пл., 1, Воронежский гос. ун-т)

e-mail: zador@amm.vsu.ru

Поступила в редакцию 26.01.2009 г.
Переработанный вариант 28.04.2009 г.

Находятся математическое ожидание и вторая моментная функция решения линейного неоднородного уравнения теплопроводности со случайными коэффициентами. Библ. 7.

Ключевые слова: теплопроводность, вариационная производная, моментные функции.

ВВЕДЕНИЕ

Реальные процессы диффузии зависят от случайных факторов. Оценка погрешности при замене случайных коэффициентов их математическими ожиданиями может быть получена, если известны математическое ожидание и дисперсионная функция диффузионного процесса.

Мы находим математическое ожидание и вторую моментную функцию решения задачи Коши

$$\frac{\partial u(t,x)}{\partial t} = \varepsilon(t)\frac{\partial^2 u(t,x)}{\partial x_1^2} + \frac{\partial^2 u(t,x)}{\partial x_2^2} + \frac{\partial^2 u(t,x)}{\partial x_3^2} + f(t,x), \tag{1}$$

$$u(t_0,x) = u_0(x). \tag{2}$$

Здесь $t \in [t_0, t_1] = T \subset \mathbb{R}$, $x \in \mathbb{R}^3$, u: $T \times \mathbb{R}^3 \longrightarrow \mathbb{R}$ – искомая функция, $\varepsilon(t) > 0$ – случайный процесс, f: $T \times \mathbb{R}^3 \longrightarrow \mathbb{R}$ – случайный процесс, $u_0 : \mathbb{R}^3 \longrightarrow \mathbb{R}$ – случайный процесс, не зависящий от ε и f.

Обычно рассуждают так. Пусть ε, f, u_0 – какие-то реализации процессов. Можно выписать по известной формуле решение и найти математическое ожидание. Возникающие трудности можно понять на простом примере. Рассмотрим задачу (здесь $x \in \mathbb{R}$)

$$\frac{\partial u(t,x)}{\partial t} = \varepsilon(t)\frac{\partial^2 u(t,x)}{\partial x_1^2}, \quad u(0,x) = u_0(x),$$

где $\varepsilon > 0$ – случайный процесс, заданный плотностью распределения $p_\varepsilon(t, \eta)$, u_0 – заданная функция. Выпишем решение

$$u_\varepsilon(t,x) = \frac{1}{2\sqrt{\pi \int_0^t \varepsilon(s)\,ds}} \int_{-\infty}^{+\infty} \exp\left[-\frac{(x-\tau)^2}{4\int_0^t \varepsilon(s)\,ds}\right] u_0(\tau)\,d\tau.$$

Пусть $\varepsilon > 0$ – случайная величина с плотностью распределения $p_\varepsilon(\eta)$, тогда

$$M(u_\varepsilon(t)) = \int_{-\infty}^{+\infty} \frac{1}{2\sqrt{\pi \eta t}} \int_{-\infty}^{+\infty} \exp\left[-\frac{(x-\tau)^2}{4\varepsilon t}\right] u_0(\tau)\,d\tau\, p_\varepsilon(\eta)\,d\eta.$$

Если же $\varepsilon(t)$ – случайный процесс, заданный плотностью распределения $p_\varepsilon(t, \eta)$, то выражение

$$\mathbf{M}(u_\varepsilon(t)) = \int_{-\infty}^{+\infty} \frac{1}{2\sqrt{\pi \int_0^t \eta(s)ds}} \int_{-\infty}^{+\infty} \exp\left[-\frac{(x-\tau)^2}{4\int_0^t \eta(s)ds}\right] u_0(\tau)d\tau p_\varepsilon(t, \eta)d\eta$$

теряет смысл, поскольку интеграл от η является оператором, а не функцией от η.

Если случайный процесс задан мерой μ_ε на пространстве σ, представляющем собой реализации процесса ε, тогда приходим к необходимости вычисления сложного континуального интеграла

$$M(u_\varepsilon(t)) = \int_{\sigma} \frac{1}{2\sqrt{\pi \int_0^t \varepsilon(s)ds}} \int_{-\infty}^{+\infty} \exp\left[-\frac{(x-\tau)^2}{4\int_0^t \varepsilon(s)ds}\right] u_0(\tau)d\tau d\mu_\varepsilon.$$

Разрабатываются и другие методы нахождения моментных функций. Используют уравнения Колмогорова для плотности распределения процесса $u(t, x)$. Строят цепочки уравнений для моментных функций (см. [1]), которые иногда удается замкнуть и решить, строят асимптотические разложения по малым случайным возмущениям (см. [2]).

Есть и другие подходы к задаче нахождения моментных функций (см. [3], [4]). В работе [3] получены формулы первых двух моментных функций решения скалярного линейного дифференциального уравнения первого порядка.

В статье применяется следующий подход. Вводится [см. разд. 4] в рассмотрение вспомогательное отображение

$$y(t, x, v, \omega) = M(u(t, x)e(v, \omega)), \tag{3}$$

где

$$e(v(\cdot), \omega(\cdot)) = \exp\left(i\int_T \varepsilon(s)v(s)ds + i\int_T \int_{R^3} f(s, \tau)\omega(s, \tau)dsd\tau\right),$$

M – знак математического ожидания по функции распределения процессов ε и f. Для отображения y получается детерминированная задача, допускающая явное решение. При этом $M(u(t, x)) = y(t, x, 0, 0)$. Более того, дифференцированием легко находятся смешанные моментные функции

$$M(u(t, x)\varepsilon(s_1)\ldots\varepsilon(s_n)) = i^{-n}\frac{\delta^n y(t, x, 0, 0)}{\delta v(s_1)\ldots\delta v(s_n)}, \quad n = 1, 2, \ldots.$$

Пусть V – банахово пространство функций v: $T \longrightarrow \mathbb{R}$ с нормой $\|v(\cdot)\|_V$ и W – банахово пространство функций $\omega : T \times \mathbb{R}^3 \longrightarrow \mathbb{R}$. Пусть a, b – заданные числа. Обозначим через $\chi(a, b, \cdot)$ функцию, определяемую по следующему правилу: $\chi(a, b, s) = \operatorname{sign}(s-a)$ при s, лежащем в отрезке с концами a и b, и $\chi(a, b, s) = 0$ в противном случае. Будем предполагать, что при $a, b \in T$ функции $\chi(a, b, \cdot)$ принадлежат V и существует постоянная $m > 0$, для которой

$$\|v(\cdot)\| \le m\|v(\cdot)\|_{L_1} = m\int_T |v(t)|dt. \tag{4}$$

Пусть выборочные функции случайного процесса ε таковы, что $\int_T \varepsilon(t)v(t)dt$ является линейным ограниченным функционалом на V; аналогично, реализации процесса f определяют линей-

ный ограниченный функционал $\int_T \int_{\mathbb{R}^3} f(t,x)\omega(t,x)dxdt$ на W, через $\|x\|_3$ обозначается $(x_1^2 + x_2^2 + x_3^2)^{1/2}$ для $x \in \mathbb{R}^3$.

Будем предполагать что случайные процессы ε и f заданы характеристическим функционалом, т.е. известно, что

$$\varphi(v(\cdot), \omega(\cdot)) = \mathrm{Me}(v(\cdot), \omega(\cdot)).$$

Сначала изучим уравнения с вариационной производной.

1. УРАВНЕНИЕ ПЕРВОГО ПОРЯДКА С ВАРИАЦИОННОЙ ПРОИЗВОДНОЙ

Пусть X – банахово пространство функций на отрезке $T \subset \mathbb{R}$, x: $T \longrightarrow \mathbb{R}$ и y: $X \longrightarrow C$.

Определение 1 (см. [5]). Если дифференциал Фреше (см. [6]) $dy(x(\cdot), h)$ функционала y в точке $x_0(\cdot)$ имеет вид

$$dy(x_0(\cdot), h) = \int_T \varphi(t, x_0(\cdot))h(t)dt,$$

где интеграл понимается в смысле Лебега, то $\varphi : T \times X \longrightarrow C$ называется *вариационной (функциональной) производной функционала y* в точке $x_0(\cdot)$ и обозначается через $\dfrac{\delta y(x_0(\cdot))}{\delta x(t)}$.

Рассмотрим линейную неоднородную задачу

$$\frac{\partial y(t, x, v(\cdot))}{\partial t} = a_1(t)\frac{\delta y(t, x, v(\cdot))}{\partial v(t)} + a_2(t)y(t, x, v(\cdot)) + b(t, x, v(\cdot)), \tag{5}$$

$$y(t_0, x, v(\cdot)) = y_0(x, v(\cdot)) \tag{6}$$

относительно неизвестного отображения $y : T \times \mathbb{R}^3 \times V \longrightarrow C$ в некоторой окрестности $\Omega \subset V$ точки $v(\cdot) = 0$. Для доказательства теоремы 2 потребуется следующий результат (см. [5, с. 183]).

Теорема 1. *Если функция $a : T \longrightarrow C$ измерима и ограничена на T, $y : V \longrightarrow C$ имеет суммируемую по s вариационную производную $\delta y(a(\cdot)\chi(t_0, t, \cdot))/\delta v(s)$ и выполняется условие* (4), *то функция $f(t) = y(a(\cdot)\chi(t_0, t, \cdot))$ почти всюду на T дифференцируема, причем*

$$\frac{df(t)}{dt} = a(t)\frac{\delta y(a(\cdot)\chi(t_0, t, \cdot))}{\delta v(t)}.$$

Теорема 2. *Пусть функции $a_i : T \longrightarrow C$, $i = 1, 2$, измеримы и ограничены на T, выполнено условие* (4), *$y_0(x, v(\cdot) + a_1(\cdot)\chi(t_0, t, \cdot))$ имеет в Ω суммируемую вариационную производную, $b : T \times \mathbb{R}^3 \times V \longrightarrow C$ суммируемо на T по первой переменной, существует суммируемая по τ вариационная производная $\delta b(s, x, v(\cdot) + a_1(\cdot)\chi(s, t, \cdot))/\delta v(\tau)$, имеющая при $v(\cdot) \in \Omega$ суммируемую на T мажоранту $|\delta b(s, x, v(\cdot) + a_1(\cdot)\chi(t_0, t, \cdot))/\delta v(\tau)| \le m(s)$. Тогда*

$$\begin{aligned} y(t, x, v(\cdot)) = {} & \exp\left(\int_{t_0}^{t} a_2(s)ds\right) y_0(x, v(\cdot) + a_1(\cdot)\chi(t_0, t, \cdot)) + \\ & + \int_{t_0}^{t} \exp\left(\int_{s}^{t} a_2(\tau)d\tau\right) b(s, x, v(\cdot) + a_1(\cdot)\chi(s, t, \cdot))ds \end{aligned} \tag{7}$$

является решением задачи (5), (6) *в окрестности Ω.*

Доказательство. Так как $b(s, x, v(\cdot) + a_1(\cdot)\chi(t_0, t, \cdot))$ имеет в Ω вариационную производную по v, то оно и непрерывно в Ω по v. Далее, b суммируемо по первому переменному, тогда $b(s, x, v(\cdot) + a_1(\cdot)\chi(s, t, \cdot))$ суммируемо по s на T. Наличие суммируемой мажоранты $m(s)$ позволяет вычис-

лить вариационную производную от интеграла в (7) вариационным дифференцированием под знаком интеграла. Воспользовавшись теоремой 1, найдем

$$\frac{\partial y(t, x, v(\cdot))}{\partial t} = a_2(t)\exp\left(\int_{t_0}^{t} a_2(s)ds\right) y_0(x, v(\cdot) + a_1(\cdot)\chi(t_0, t, \cdot)) +$$

$$+ \exp\left(\int_{t_0}^{t} a_2(s)ds\right) a_1(t)\frac{\delta}{\delta v(t)} y_0(x, v(\cdot) + a_1(\cdot)\chi(t_0, t, \cdot)) + b(t, x, v(\cdot)) +$$

$$+ \int_{t_0}^{t} a_2(t)\exp\left(\int_{s}^{t} a_2(\tau)d\tau\right) b(s, x, v(\cdot) + a_1(\cdot)\chi(s, t, \cdot))ds +$$

$$+ \int_{t_0}^{t} \exp\left(\int_{s}^{t} a_2(\tau)d\tau\right) a_1(t)\frac{\delta b(s, x, v(\cdot) + a_1(\cdot)\chi(s, t, \cdot))}{\delta v(t)}ds.$$

Вычислим вариационную производную

$$\frac{\delta y(t, x, v(\cdot))}{\delta v(t)} = \exp\left(\int_{t_0}^{t} a_2(s)ds\right)\frac{\delta}{\delta v(t)} y_0(x, v(\cdot) + a_1(\cdot)\chi(t_0, t, \cdot)) +$$

$$+ \int_{t_0}^{t} \exp\left(\int_{s}^{t} a_2(\tau)d\tau\right)\frac{\delta b(s, x, v(\cdot) + a_1(\cdot)\chi(s, t, \cdot))}{\delta v(t)}ds.$$

Подставляя в (5), получаем верное равенство при $v(\cdot) \in \Omega$ почти при всех $t \in T$. Теорема доказана.

2. УРАВНЕНИЕ ТРЕТЬЕГО ПОРЯДКА С ВАРИАЦИОННОЙ ПРОИЗВОДНОЙ

Пусть $x \in \mathbb{R}^3$, $f: \mathbb{R}^3 \times \mathbb{R}^m \longrightarrow C$, ξ – вектор с координатами ξ_1, ξ_2, ξ_3, $F_x[f(x, y)](\xi)$ обозначает преобразование Фурье по переменному x, аналогичное обозначение используется для обратного преобразования Фурье, $\overset{x}{*}$ обозначает свертку (см. [7]) по переменному x.

Рассмотрим задачу Коши для дифференциального уравнения третьего порядка

$$\frac{\partial y(t, x, v(\cdot))}{\partial t} = -i\frac{\delta}{\delta v(t)}\frac{\partial^2}{\partial x_1^2} y(t, x, v(\cdot)) + \frac{\partial^2}{\partial x_2^2} y(t, x, v(\cdot)) + \frac{\partial^2}{\partial x_3^2} y(t, x, v(\cdot)) + b(t, x, v(\cdot)), \tag{8}$$

$$y(t_0, x, v(\cdot)) = y_0(x, v(\cdot)). \tag{9}$$

Здесь $t \in T \subset \mathbb{R}^3$, $x \in \mathbb{R}^3$, $b: T \times \mathbb{R}^3 \times V \longrightarrow C$ – заданное отображение, $y_0: \mathbb{R}^3 \times V \longrightarrow C$ задано, y – искомое отображение. В формулировке следующей теоремы у отображений y_0 и b опущены обозначения аргументов: соответственно, $x, v(\cdot) + i\xi_1^2\chi(t_0, t, \cdot)$ у y_0 и $s, x, v(\cdot) + i\xi_1^2\chi(s, t, \cdot)$ у b; $\delta(x)$ обозначает дельта-функцию.

Теорема 3. *Пусть выполнено условие* (4), *существует окрестность* Ω *точки* $v(\cdot) = 0$ *такая, что при всех* $v(\cdot) \in \Omega$ *функции*

$$|y_0|, \quad \left|\frac{\delta y_0}{\delta v(t)}\right|, \quad \left|F_{x_1}\left[\frac{\delta y_0}{\delta v(t)}\right](\xi)\right|, \quad |\xi_1 F_{x_1}[y_0](\xi)|, \quad \xi_1^2|F_{x_1}[y_0](\xi)|, \quad \left|\xi_1 F_{x_1}\left[\frac{\delta y_0}{\delta v(t)}\right](\xi)\right|,$$

$$\xi_1^2\left|F_{x_1}\left[\frac{\delta y_0}{\delta v(t)}\right](\xi)\right|, \quad |b|, \quad \left|\frac{\delta b}{\delta v(t)}\right|, \quad \left|F_{x_1}\left[\frac{\delta b}{\delta v(t)}\right](\xi)\right|, \quad |\xi_1||F_{x_1}[b](\xi)|, \quad \xi_1^2|F_{x_1}[b](\xi)|,$$

$$|\xi_1|\left|F_{x_1}\left[\frac{\delta b}{\delta v(t)}\right](\xi)\right|, \quad \xi_1^2\left|F_{x_1}\left[\frac{\delta b}{\delta v(t)}\right](\xi)\right|$$

ограничены при $t \in T$, $s \in T$ *суммируемыми на* $\mathbb{R}^3$ *функциями. Тогда решение задачи* (8), (9) *находится по формуле*

$$y(t, x, v(\cdot)) = \frac{1}{4\pi(t-t_0)}\exp\left[-\frac{x_2^2+x_3^2}{4(t-t_0)}\right]\delta(x_1) \overset{x}{*} F_{\xi_1}^{-1}[F_{x_1}[y_0(x, v(\cdot) + i\xi_1^2\chi(t_0, t, \cdot))](\xi_1)](x_1) +$$
$$+ \int_{t_0}^{t}\frac{1}{4\pi(t-s)}\exp\left[-\frac{x_2^2+x_3^2}{4(t-s)}\right]\delta(x_1) \overset{x}{*} F_{\xi_1}^{-1}[F_{x_1}[b(s, x, v(\cdot) + i\xi_1^2\chi(s, t, \cdot))](\xi_1)](x_1)ds. \tag{10}$$

Доказательство. Предположим, что для решения задачи (8), (9) существует преобразование Фурье по переменному x. Применив преобразование Фурье к (8) и (9), получим

$$\frac{\partial}{\partial t}F_x[y(t, x, v(\cdot))](\xi) = -i\frac{\delta}{\delta v(t)}(-i\xi_1)^2F_x[y(t, x, v(\cdot))](\xi) +$$
$$+ (-i\xi_2)^2F_x[y(t, x, v(\cdot))](\xi) + (-i\xi_3)^2F_x[y(t, x, v(\cdot))](\xi) + F_x[b(t, x, v(\cdot))](\xi),$$
$$F_x[y(t_0, x, v(\cdot))](\xi) = F_x[y_0(x, v(\cdot))](\xi).$$

Эта задача имеет вид (5), (6), при этом ξ является параметром. Воспользуемся формулой (7):

$$F_x[y(t, x, v(\cdot))](\xi) = \exp[-(\xi_2^2+\xi_3^2)(t-t_0)]F_x[y_0(x, v(\cdot) + i\xi_1^2\chi(t_0, t, \cdot))](\xi) +$$
$$+ \int_{t_0}^{t}\exp[-(\xi_2^2+\xi_3^2)(t-s)]F_x[b(s, x, v(\cdot) + i\xi_1^2\chi(s, t, \cdot))](\xi)ds.$$

Применяя обратное преобразование Фурье, получаем

$$y(t, x, v(\cdot)) = F_\xi^{-1}\{\exp[-(\xi_2^2+\xi_3^2)(t-t_0)]\}(x) \overset{x}{*} F_\xi^{-1}\{F_x[y_0(x, v(\cdot) + i\xi_1^2\chi(t_0, t, \cdot))](\xi)\}(x) +$$
$$+ \int_{t_0}^{t}F_\xi^{-1}\{\exp[-(\xi_2^2+\xi_3^2)(t-s)]\}(x) \overset{x}{*} F_\xi^{-1}\{F_x[b(s, x, v(\cdot) + i\xi_1^2\chi(s, t, \cdot))](\xi)\}(x)ds =$$
$$= \frac{1}{4\pi(t-t_0)}\exp\left[-\frac{x_2^2+x_3^2}{4(t-t_0)}\right]\delta(x_1) \overset{x}{*} F_\xi^{-1}\{F_x[y_0(x, v(\cdot) + i\xi_1^2\chi(t_0, t, \cdot))](\xi)\}(x) +$$
$$+ \int_{t_0}^{t}\frac{1}{4\pi(t-s)}\exp\left[-\frac{x_2^2+x_3^2}{4(t-s)}\right]\delta(x_1) \overset{x}{*} F_\xi^{-1}\{F_x[b(s, x, v(\cdot) + i\xi_1^2\chi(s, t, \cdot))](\xi)\}(x)ds.$$

Предположения теоремы о наличии суммируемых мажорант обеспечивают дифференцируемость под знаком интеграла по нужным нам переменным. Используя свойства преобразования Фурье, находим $F_\xi^{-1}\{F_x[y_0(x, v(\cdot) + i\xi_1^2\chi(t_0, t, \cdot))](\xi)\}(x) = F_{\xi_1}^{-1}\{F_{\xi_2\xi_3}^{-1}[F_{x_2x_3}[F_{x_1}[y_0(x, v(\cdot) + i\xi_1^2\chi(t_0, t, \cdot))](\xi_1)](\xi_2\xi_3)](x_2x_3)\}(x_1) = F_{\xi_1}^{-1}\{F_{x_1}[y_0(x, v(\cdot) + i\xi_1^2\chi(t_0, t, \cdot))](\xi_1)\}(x_1)$, аналогично получаем выражение $F_\xi^{-1}\{F_x[b(s, x, v(\cdot) + i\xi_1^2\chi(s, t, \cdot))](\xi)\}(x) = F_{\xi_1}^{-1}\{F_{\xi_2\xi_3}^{-1}[F_{x_2x_3}[F_{x_1}[b(s, x, v(\cdot) + i\xi_1^2\chi(s, t, \cdot))](\xi_1)](\xi_2\xi_3)](x_2x_3)\}(x_1) = F_{\xi_1}^{-1}\{F_{x_1}[b(s, x, v(\cdot) + i\xi_1^2\chi(s, t, \cdot))](\xi_1)\}(x_1)$. Подставляя эти выражения в предыдущее равенство, приходим к формуле (10).

Теорема доказана.

3. МАТЕМАТИЧЕСКОЕ ОЖИДАНИЕ РЕШЕНИЯ УРАВНЕНИЯ ТЕПЛОПРОВОДНОСТИ

Введем обозначение

$$y(t, x, v(\cdot), \omega(\cdot)) = M(u(t, x)e(v(\cdot), \omega(\cdot))),$$

где математическое ожидание вычисляется по функции распределения случайных процессов u_0, ε и f задачи (1), (2).

Умножим (1), (2) на $e(v(\cdot), \omega(\cdot))$ и найдем математическое ожидание по функции распределения процессов u_0, ε и f. Если существуют соответствующие производные отображения y, то последние равенства записываются в виде

$$\frac{\partial y(t, x, v(\cdot), \omega(\cdot))}{\partial t} = -i\frac{\delta}{\delta v(t)}\frac{\partial^2}{\partial x_1^2}y(t, x, v(\cdot), \omega(\cdot)) + \frac{\partial^2}{\partial x_2^2}y(t, x, v(\cdot), \omega(\cdot)) + \frac{\partial^2}{\partial x_3^2}y(t, x, v(\cdot), \omega(\cdot)) - i\frac{\delta\varphi(v(\cdot), \omega(\cdot))}{\partial\omega(t, x)}, \tag{11}$$

$$y(t_0, x, v(\cdot), \omega(\cdot)) = M(u_0(x))\varphi(v(\cdot), \omega(\cdot)), \tag{12}$$

где φ – характеристический функционал процессов ε и f. При этом использовано свойство независимости случайного процесса u_0 с ε и f.

Задача (11), (12) является детерминированной, но уравнение (11) нетрадиционно, так как содержит вариационное дифференцирование. Из вида y, естественно, приходим к следующему определению.

Определение 2. *Математическим ожиданием* решения задачи (1), (2) называется

$$\mathrm{M}u(t, x) = y(t, x, 0, 0), \tag{13}$$

где y – решение задачи (11), (12) в некоторой окрестности нулевой точки (0, 0) в $V \times W$. Если y является решением (11), (12) в смысле обобщенных функций, то (13) называется *обобщенным математическим ожиданием решения задачи* (1), (2).

Если положить $M(u_0(x))\varphi(v(\cdot), \omega(\cdot)) = y_0(x, v(\cdot), \omega(\cdot))$ и $-i\delta\varphi(v(\cdot), \omega(\cdot))/\delta\omega(t, x) = b(t, x, v(\cdot), \omega(\cdot))$, то задача (11), (12) при каждом фиксированном $\omega(\cdot)$ имеет вид (8), (9).

Теорема 4. *Пусть функция $M(u_0(x))$ суммируема на $\mathbb{R}^3$ и при каждом $\omega(\cdot)$ из некоторой окрестности нуля в W выполняются условия теоремы* 3, *тогда решение задачи* (11), (12) *находится по формуле*

$$y(t, x, v(\cdot), \omega(\cdot)) = \frac{1}{4\pi(t-t_0)}\exp\left[-\frac{x_2^2+x_3^2}{4(t-t_0)}\right]^{x_2x_3}_{*}\{M(u_0(x))\overset{x_1}{*}F^{-1}_{\xi_1}[\varphi(v(\cdot)+i\xi_1^2\chi(t_0, t, \cdot), \omega(\cdot))](x_1)\} - i\int_{t_0}^{t}\frac{1}{4\pi(t-s)}\exp\left[-\frac{x_2^2+x_3^2}{4(t-s)}\right]^{x_2x_3}_{*}F^{-1}_{\xi_1}\left\{F_{x_1}\left[\frac{\delta\varphi(v(\cdot)+i\xi_1^2\chi(s, t, \cdot), \omega(\cdot))}{\delta\omega(s, x)}\right](\xi_1)\right\}(x_1)ds. \tag{14}$$

Доказательство. Используя формулу (10), находим решение задачи (11), (12):

$$y = \frac{1}{4\pi(t-t_0)}\exp\left(-\frac{x_2^2+x_3^2}{4(t-t_0)}\right)\delta(x_1)\overset{x}{*}F^{-1}_{\xi_1}[F_{x_1}[M(u_0(x))\varphi(v(\cdot)+i\xi_1^2\chi(t_0, t, \cdot), \omega(\cdot))](\xi_1)](x_1) - i\int_{t_0}^{t}\frac{1}{4\pi(t-s)}\exp\left[-\frac{x_2^2+x_3^2}{4(t-s)}\right]\delta(x_1)\overset{x}{*}F^{-1}_{\xi_1}\left\{F_{x_1}\left[\frac{\delta\varphi(v(\cdot)+i\xi_1^2\chi(s, t, \cdot), \omega(\cdot))}{\delta\omega(s, x)}\right](\xi_1)\right\}(x_1)ds.$$

Так как обратное преобразование Фурье произведения функций преобразуется в свертку обратных преобразований Фурье от сомножителей, то из последнего равенства получаем (14). Теорема доказана.

Теорема 5. *При выполнении условий теоремы* 3 *математическое ожидание решения задачи* (1), (2) *находится по формуле*

$$Mu(t,x) = \frac{1}{4\pi(t-t_0)}\exp\left[-\frac{x_2^2+x_3^2}{4(t-t_0)}\right] \overset{x_2x_3}{*} [M(u_0(x)) \overset{x_1}{*} F^{-1}_{\xi_1}(\varphi(i\xi_1^2\chi(t_0,t,\cdot),0))(x_1)] -$$
$$- i\int_{t_0}^{t} \frac{1}{4\pi(t-s)}\exp\left[-\frac{x_2^2+x_3^2}{4(t-s)}\right] \overset{x_2x_3}{*} F^{-1}_{\xi_1}\left[F_{x_1}\left(\frac{\delta\varphi(i\xi_1^2\chi(s,t,\cdot),0)}{\delta\omega(s,x)}\right)(\xi_1)\right](x_1)ds. \tag{15}$$

Доказательство следует из (13) и (14).

4. ЧАСТНЫЕ СЛУЧАИ

Полученная формула (15) является довольно общей, не требуется даже независимости процессов ε и f.

1. *Случай независимых процессов* ε *и* f. При этом характеристический функционал $\varphi(v(\cdot), \omega(\cdot))$ является произведением характеристических функционалов $\varphi_\varepsilon(v(\cdot))$ и $\varphi_f(\omega(\cdot))$, определяющих процессы ε и f.

Теорема 6. *Пусть в задаче* (1), (2) *случайные процессы* u_0, ε *и* f *независимы, выполняется условие* (4), *характеристический функционал* $\varphi_\varepsilon : V \longrightarrow C$ *процесса* ε *имеет вариационную производную по* $v \in L_1(T)$, *функции* $M(u_0(x))$ *и* $M(f(t,x))$ *локально суммируемы. Тогда обобщенное математическое ожидание решения задачи* (1), (2) *находится по формуле*

$$Mu(t,x) = \frac{1}{4\pi(t-t_0)}\exp\left[-\frac{x_2^2+x_3^2}{4(t-t_0)}\right] \overset{x_2x_3}{*} \{M(u_0(x)) \overset{x_1}{*} F^{-1}_{\xi_1}[\varphi_\varepsilon(i\xi_1^2\chi(t_0,t,\cdot))](x_1)\} +$$
$$+ \int_{t_0}^{t} \frac{1}{4\pi(t-s)}\exp\left[-\frac{x_2^2+x_3^2}{4(t-s)}\right] \overset{x_2x_3}{*} \{F^{-1}_{\xi_1}[\varphi_\varepsilon(i\xi_1^2\chi(s,t,\cdot))](x_1) \overset{x_1}{*} Mf(s,x)\}ds. \tag{16}$$

Доказательство. Отметим, что $\dfrac{\delta\varphi_f(0)}{\delta\omega(t,x)} = iMf(t,x)$, $\varphi_f(0) = 1$; $\varphi(v(\cdot),\omega(\cdot)) = \varphi_\varepsilon(v(\cdot))\varphi_f(\omega(\cdot))$. Воспользовавшись формулами (13), (14), находим

$$Mu(t,x) = \frac{1}{4\pi(t-t_0)}\exp\left[-\frac{x_2^2+x_3^2}{4(t-t_0)}\right] \overset{x_2x_3}{*} \{M(u_0(x)) \overset{x_1}{*} F^{-1}_{\xi_1}[\varphi_\varepsilon(i\xi_1^2\chi(t_0,t,\cdot))\varphi_f(0)](x_1)\} -$$
$$- i\int_{t_0}^{t} \frac{1}{4\pi(t-s)}\exp\left[-\frac{x_2^2+x_3^2}{4(t-s)}\right] \overset{x_2x_3}{*} F^{-1}_{\xi_1}\left\{F_{x_1}\left[\varphi_\varepsilon(i\xi_1^2\chi(s,t,\cdot))\frac{\delta\varphi_f(0)}{\delta\omega(s,x)}\right](\xi_1)\right\}(x_1)ds =$$
$$= \frac{1}{4\pi(t-t_0)}\exp\left[-\frac{x_2^2+x_3^2}{4(t-t_0)}\right] \overset{x_2x_3}{*} \{M(u_0(x)) \overset{x_1}{*} F^{-1}_{\xi_1}[\varphi_\varepsilon(i\xi_1^2\chi(t_0,t,\cdot))](x_1)\} +$$
$$+ \int_{t_0}^{t} \frac{1}{4\pi(t-s)}\exp\left[-\frac{x_2^2+x_3^2}{4(t-s)}\right] \overset{x_2x_3}{*} F^{-1}_{\xi_1}[\varphi_\varepsilon(i\xi_1^2\chi(s,t,\cdot))F_{x_1}[Mf(s,x)](\xi_1)](x_1)ds.$$

Обратное преобразование Фурье под знаком интеграла выражается через свертку, откуда и следует (16). Теорема доказана.

2. *Случай гауссовского процесса* ε. Гауссовский случайный процесс определяется характеристическим функционалом

$$\varphi_\varepsilon(v(\cdot)) = \exp\left(i\int_T M\varepsilon(s)v(s)ds - \frac{1}{2}\iint_{TT} b(s_1,s_2)v(s_1)v(s_2)ds_1ds_2\right),$$

где $M\varepsilon(s)$ – математическое ожидание ε и $b(s_1, s_2) = M(\varepsilon(s_1)\varepsilon(s_2) - M\varepsilon(s_1)M\varepsilon(s_2)$ – ковариационная функция процесса ε.

Теорема 7. *Пусть гауссовский случайный процесс ε независим от процесса f, $M\varepsilon(t) > 0$, $M\varepsilon(\cdot) \in L_\infty(T)$, функция b: $T \times T \longrightarrow \mathbb{R}$ измерима и ограничена, функции $M(u_0(\cdot))$ и $M(f(t, \cdot))$ локально суммируемы. Тогда обобщенное математическое ожидание решения задачи* (1), (2) *вычисляется по формуле*

$$Mu(t, x) = \frac{1}{4\pi(t-t_0)}\exp\left[-\frac{x_2^2+x_3^2}{4(t-t_0)}\right] {}^{x_2x_3}\!* \left[\frac{1}{2\sqrt{\pi\int\limits_{t_0}^{t} M\varepsilon(s)\,ds}}\exp\left(-\frac{x_1^2}{4\int\limits_{t_0}^{t} M\varepsilon(s)\,ds}\right)\right] {}^{x_1}\!* \sum_{k=0}^{\infty}\frac{1}{2^k k!}\left(\int\limits_{t_0}^{t}\int\limits_{t_0}^{t} b(s_1, s_2)\,ds_1 ds_2\right)^k \times$$

$$\times \frac{\partial^{4k}}{\partial x_1^{4k}} M(u_0(x)) + \int\limits_{t_0}^{t}\frac{1}{4\pi(t-s)}\exp\left(-\frac{x_2^2+x_3^2}{4(t-s)}\right) {}^{x_2x_3}\!* \left[\frac{1}{2\sqrt{\pi\int\limits_{t_0}^{t} M\varepsilon(\tau)\,ds}}\exp\left(-\frac{x_1^2}{4\int\limits_{t_0}^{t} M\varepsilon(\tau)\,ds}\right)\right] {}^{x_1}\!* \sum_{k=0}^{\infty}\frac{1}{2^k k!} \times \tag{17}$$

$$\times\left(\int\limits_{s}^{t}\int\limits_{s}^{t} b(s_1, s_2)\,ds_1 ds_2\right)^k \frac{\partial^{4k}}{\partial x_1^{4k}} Mf(s, x)\,ds.$$

Доказательство. Функционал φ_ε имеет вариационную производную, и выражение (16) определено в обобщенном смысле. Известно (см. [7]), что

$$F_\xi^{-1}\left[\exp\left(-\xi_1^2\int\limits_{t_0}^{t} M\varepsilon(s)\,ds\right)\right](x) = \left(4\pi\int\limits_{t_0}^{t} M\varepsilon(s)\,ds\right)^{-1/2}\exp\left(-\frac{x_1^2}{4\int\limits_{t_0}^{t} M\varepsilon(s)\,ds}\right) \quad \text{при} \quad \int\limits_{t_0}^{t} M\varepsilon(s)\,ds > 0.$$

Последнее условие выполняется, так как $M\varepsilon(s) > 0$. Далее,

$$F_{\xi_1}^{-1}[\varphi_\varepsilon(i\xi_1^2\chi(t_0, t, \cdot))](x_1) = F_{\xi_1}^{-1}\left[\exp\left(-\xi_1^2\int\limits_{t_0}^{t} M\varepsilon(s)\,ds\right)\right](x_1) {}^{x}\!* F_{\xi_1}^{-1}\left[\exp\left(\frac{1}{2}\xi_1^4\int\limits_{t_0}^{t}\int\limits_{t_0}^{t} b(s_1, s_2)\,ds_1 ds_2\right)\right](x_1).$$

Введем обозначение $B(t_0, t) = \int_{t_0}^{t}\int_{t_0}^{t} b(s_1, s_2)\,ds_1 ds_2$, тогда

$$F_{\xi_1}^{-1}\left[\exp\left(\frac{1}{2}\xi_1^4 B(t_0, t)\right)\right](x_1) = \sum_{k=0}^{\infty}\frac{B^k(t_0, t)}{2^k k!}F_{\xi_1}^{-1}[\xi_1^4](x_1) = \sum_{k=0}^{\infty}\frac{B^k(t_0, t)}{2^k k!}\frac{d^{4k}}{dx_1^{4k}}\delta(x_1).$$

При этом

$$F_{\xi_1}^{-1}\left[\exp\left(\frac{1}{2}B(s, t)\xi_1^4\right)\right](x_1) {}^{x_1}\!* Mf(s, x) = \int\limits_R\left(\sum_{k=0}^{\infty}\frac{B^k(s, t)}{2^k k!}\frac{\partial^{4k}}{\partial x_1^{4k}}\delta(x-\eta)Mf(s, \eta)\right)d\eta =$$

$$= \sum_{k=0}^{\infty}\frac{B^k(t, s)}{2^k k!}\frac{\partial^{4k}}{\partial x_1^{4k}}Mf(s, x).$$

Подставляя эти соотношения в (16), находим (17). Теорема доказана.

5. ВСПОМОГАТЕЛЬНАЯ ЗАДАЧА КОШИ

Для нахождения второй моментной функции решения задачи (1), (2) поступим так же, как при нахождении математического ожидания. Введем вспомогательное отображение

$$z(t, t_1, x, \tau, v(\cdot), \omega(\cdot)) = M(u(t, x)u(t_1, \tau)e(v(\cdot), \omega(\cdot))).$$

Умножим уравнение (1) на $u(t_1, \tau)e(v(\cdot), \omega(\cdot))$ и усредним по функции распределения процессов ε, f и u_0. Формально это равенство записывается с использованием z и y в виде

$$\begin{aligned}\frac{\partial z(t, t_1, x, \tau, v(\cdot), \omega(\cdot))}{\partial t} &= -i\frac{\delta}{\delta v(t)}\frac{\partial^2}{\partial x_1^2}z(t, t_1, x, \tau, v(\cdot), \omega(\cdot)) + \frac{\partial^2}{\partial x_2^2}z(t, t_1, x, \tau, v(\cdot), \omega(\cdot)) + \\ &+ \frac{\partial^2}{\partial x_3^2}z(t, t_1, x, \tau, v(\cdot), \omega(\cdot)) - i\frac{\partial y(t_1, \tau, v(\cdot), \omega(\cdot))}{\partial \omega(t, x)}.\end{aligned} \tag{18}$$

К сожалению, из условия (2) не удается получить начальное значение $z(t_0, t_1, x, \tau, v(\cdot), \omega(\cdot))$. Однако если умножить (2) на $u(t_0, \tau)e(v(\cdot), \omega(\cdot))$ и усреднить по функции распределения процессов ε, f и u_0, то получим

$$z(t_0, t_0, x, \tau, v(\cdot), \omega(\cdot)) = M(u_0(x)u_0(\tau))\varphi(v(\cdot), \omega(\cdot)). \tag{19}$$

Из определения отображения z следует, что оно симметрично по переменным (t, x) и (t_1, τ), т.е. $z(t, t_1, x, \tau, v(\cdot), \omega(\cdot)) = z(t_1, t, \tau, x, v(\cdot), \omega(\cdot))$. Эта особенность позволяет найти отображение z.

Теорема 8. *Пусть выполнено условие* (4), $M(u_0(x))$, $M(u_0(\tau))$ *и* $M(u_0(x)u_0(\tau))$ *локально суммируемы, в некоторой окрестности точки* $(0, 0) \in V \times W$ *и существуют вариационные производные*

$$\frac{\delta\varphi(v(\cdot) + i\eta_1^2\chi(s, t_1, \cdot) + i\xi_1^2\chi(t_0, t, \cdot), \omega(\cdot))}{\delta\omega(s, \tau)}, \quad \frac{\delta\varphi(v(\cdot) + i\xi_1^2\chi(s, t, \cdot) + i\eta_1^2\chi(t_0, t_1, \cdot), \omega(\cdot))}{\delta\omega(s, x)},$$

$$\frac{\delta^2\varphi(v(\cdot) + i\eta_1^2\chi(\theta, t_1, \cdot) + i\xi_1^2\chi(s, t, \cdot), \omega(\cdot))}{\delta\omega(s, x)\delta\omega(\theta, \tau)}.$$

Тогда симметричное по переменным (t, x) и (t_1, τ) обобщенное решение задачи (18), (19) находится по формуле

$$z = \frac{1}{4\pi(t-t_0)}\exp\left[-\frac{x_2^2 + x_3^2}{4(t-t_0)}\right] {}^{x_2x_3}\!* \left\{\frac{1}{4\pi(t_1-t_0)} \times\right.$$

$$\times \exp\left[-\frac{\tau_2^2 + \tau_3^2}{4(t_1-t_0)}\right] {}^{\tau_2\tau_3}\!* \{M(u_0(x)u_0(\tau)) {}^{\tau_1}\!* F_{\xi_1}^{-1}[F_{\eta_1}^{-1}[\varphi(v(\cdot) + i\eta_1^2\chi(t_0, t_1, \cdot) + i\xi_1^2\chi(t_0, t, \cdot), \omega(\cdot))](\tau_1)](x_1)\} -$$

$$- i\frac{1}{4\pi(t-t_0)}\exp\left(-\frac{x_2^2 + x_3^2}{4(t-t_0)}\right) {}^{x_2x_3}\!* \left\{M(u_0(x)) {}^{x_1}\!* \int_{t_0}^{t_1}\frac{1}{4\pi(t_1-s)} \times\right.$$

$$\left.\left.\times \exp\left(-\frac{\tau_2^2 + \tau_3^2}{4(t_1-s)}\right) {}^{\tau_2\tau_3}\!* F_{\xi_1}^{-1}\left[F_{\eta_1}^{-1}\left[F_{x_1}\left[\frac{\delta}{\delta\omega(s, \tau)}\varphi(v(\cdot) + i\eta_1^2\chi(s, t_1, \cdot) + i\xi_1^2\chi(t_0, t, \cdot), \omega(\cdot))\right](\eta_1)\right](\tau_1)\right](x_1)ds\right\} - \tag{20}$$

$$- i\frac{1}{4\pi(t_1-t_0)}\exp\left(-\frac{\tau_2^2 + \tau_3^2}{4(t_1-t_0)}\right) {}^{\tau_2\tau_3}\!* \left\{M(u_0(\tau)) {}^{\tau_1}\!* \int_{t_0}^{t}\frac{1}{4\pi(t-s)}\exp\left(-\frac{x_2^2 + x_3^2}{4(t-s)}\right) {}^{x_2x_3}\!* F_{\eta_1}^{-1}\left[F_{\xi_1}^{-1}\left[F_{x_1}\left[\frac{\delta}{\delta\omega(s, x)}\varphi(v(\cdot) +\right.\right.\right.$$

$$+ i\eta_1^2\chi(t_0, t_1, \cdot) + i\xi_1^2\chi(s, t, \cdot), \omega(\cdot))\Big](\xi_1)\Big](x_1)\Big](\tau_1)ds\Big\} -$$

$$- \int_{t_0}^{t} ds \int_{t_0}^{t_1} \frac{1}{4\pi(t-s)} \exp\left(-\frac{x_2^2 + x_3^2}{4(t-s)}\right)^{x_2x_3} * \left\{\frac{1}{4\pi(t_1-\theta)} \exp\left[-\frac{\tau_2^2 + \tau_3^2}{4(t_1-\theta)}\right]^{\tau_2\tau_3} * F_{\xi_1}^{-1}\left[F_{x_1}\left[F_{\eta_1}^{-1}\left[F_{x_1} \times\right.\right.\right.\right.$$

$$\left.\left.\left.\left. \times \left[\frac{\delta^2\varphi(v(\cdot) + i\eta_1^2\chi(\theta, t_1, \cdot) + i\xi_1^2\chi(s, t, \cdot), \omega(\cdot))}{\delta\omega(s, x)\delta\omega(\theta, \tau)}\right](\eta_1)\right](\tau_1)\right](\xi_1)\right](x_1)\right\} d\theta.$$

Доказательство. Положим в (18) $t_1 = t_0$. Задача (18), (19) (при $t_1 = t_0$) для отображения $z(t, t_0, x, \tau, v(\cdot), \omega(\cdot))$ имеет вид задачи (11), (12). По формуле (14) находим

$$z(t, t_0, x, \tau, v(\cdot), \omega(\cdot)) = \frac{1}{4\pi(t-t_0)} \times$$

$$\times \exp\left(-\frac{x_2^2 + x_3^2}{4(t-t_0)}\right)^{x_2x_3} * (M(u_0(x)u_0(\tau))^{x_1} * F_{\xi_1}^{-1}[\varphi(v(\cdot) + i\xi_1^2\chi(t_0, t, \cdot), \omega(\cdot))](x_1)) -$$

$$- i\int_{t_0}^{t} \frac{1}{4\pi(t-s)} \exp\left(-\frac{x_2^2 + x_3^2}{4(t-s)}\right)^{x_2x_3} * F_{\xi_1}^{-1}\left[F_{x_1}\left(\frac{\delta y(t_0, \tau, v(\cdot) + i\xi_1^2\chi(s, t, \cdot), \omega(\cdot))}{\delta\omega(s, x)}\right)(\xi_1)\right](x_1)ds.$$

Так как z симметрично по переменным (t, x) и (t_1, τ), то

$$z(t_0, t_1, \tau, x, v(\cdot), \omega(\cdot)) = \frac{1}{4\pi(t_1-t_0)} \times$$

$$\times \exp\left(-\frac{x_2^2 + x_3^2}{4(t_1-t_0)}\right)^{x_2x_3} * (M(u_0(x)u_0(\tau))^{x_1} * F_{\xi_1}^{-1}[\varphi(v(\cdot) + i\xi_1^2\chi(t_0, t_1, \cdot), \omega(\cdot))](x_1)) -$$

$$- i\int_{t_0}^{t_1} \frac{1}{4\pi(t_1-s)} \exp\left(-\frac{x_2^2 + x_3^2}{4(t_1-s)}\right)^{x_2x_3} * F_{\xi_1}^{-1}\left[F_{x_1}\left(\frac{\delta y(t_0, \tau, v(\cdot) + i\xi_1^2\chi(s, t_1, \cdot), \omega(\cdot))}{\delta\omega(s, x)}\right)(\xi_1)\right](x_1)ds.$$

Тогда

$$z(t_0, t_1, x, \tau, v(\cdot), \omega(\cdot)) = \frac{1}{4\pi(t_1-t_0)} \times$$

$$\times \exp\left(-\frac{\tau_2^2 + \tau_3^2}{4(t_1-t_0)}\right)^{\tau_2\tau_3} * (M(u_0(x)u_0(\tau))^{\tau_1} * F_{\xi_1}^{-1}[\varphi(v(\cdot) + i\xi_1^2\chi(t_0, t_1, \cdot), \omega(\cdot))](\tau_1)) -$$

$$- i\int_{t_0}^{t_1} \frac{1}{4\pi(t_1-s)} \exp\left(-\frac{\tau_2^2 + \tau_3^2}{4(t_1-s)}\right)^{\tau_2\tau_3} * F_{\xi_1}^{-1}\left[F_{\tau_1}\left(\frac{\delta y(t_0, x, v(\cdot) + i\xi_1^2\chi(s, t_1, \cdot), \omega(\cdot))}{\delta\omega(s, \tau)}\right)(\xi_1)\right](\tau_1)ds.$$

Используя (12), находим начальное условие для уравнения (18):

$$z(t_0, t_1, x, \tau, v(\cdot), \omega(\cdot)) = \frac{1}{4\pi(t_1-t_0)} \times$$

$$\times \exp\left(-\frac{\tau_2^2 + \tau_3^2}{4(t_1-t_0)}\right)^{\tau_2\tau_3} * (M(u_0(x)u_0(\tau))^{\tau_1} * F_{\xi_1}^{-1}[\varphi(v(\cdot) + i\xi_1^2\chi(t_0, t_1, \cdot), \omega(\cdot))](\tau_1)) -$$

$$- iM(u_0(x))\int_{t_0}^{t_1} \frac{1}{4\pi(t_1-s)} \exp\left(-\frac{\tau_2^2 + \tau_3^2}{4(t_1-s)}\right)^{\tau_2\tau_3} * F_{\xi_1}^{-1}\left[F_{x_1}\left(\frac{\delta\varphi(v(\cdot) + i\xi_1^2\chi(s, t_1, \cdot), \omega(\cdot))}{\delta\omega(s, \tau)}\right)(\xi_1)\right](\tau_1)ds.$$

Уравнение (18) с этим начальным условием имеет вид задачи (8), (9), и по формуле (10) находим

$$z(t, t_1, x, \tau, v(\cdot), \omega(\cdot)) = \frac{1}{4\pi(t-t_0)}\exp\left(-\frac{x_2^2+x_3^2}{4(t-t_0)}\right)\delta(x_1) \overset{x}{*} F_{\xi_1}^{-1}\left[F_{x_1}\left[\frac{1}{4\pi(t_1-t_0)} \times\right.\right.$$

$$\times \exp\left(-\frac{\tau_2^2+\tau_3^2}{4(t_1-t_0)}\right) \overset{\tau_2\tau_3}{*} (M(u_0(x)u_0(\tau)) \overset{\tau_1}{*} F_{\eta_1}^{-1}[\varphi(v(\cdot) + i\eta_1^2\chi(t_0, t_1, \cdot) + i\xi_1^2\chi(t_0, t, \cdot), \omega(\cdot))](\tau_1)\left.\right](\xi_1)\left.\right](x_1)) -$$

$$- i\frac{1}{4\pi(t-t_0)}\exp\left(-\frac{x_2^2+x_3^2}{4(t-t_0)}\right)\delta(x_1) \overset{x}{*} F_{\xi_1}^{-1}\left[F_{x_1}\left[M(u_0(x))\int_{t_0}^{t_1}\frac{1}{4\pi(t_1-s)}\exp\left(-\frac{\tau_2^2+\tau_3^2}{4(t_1-s)}\right) \overset{\tau_2\tau_3}{*} F_{\eta_1}^{-1}\left[F_{x_1}\left[\frac{\delta}{\delta\omega(s,\tau)}\varphi(v(\cdot) +\right.\right.\right.\right.$$

$$\left.\left.\left.\left. + i\eta_1^2\chi(s, t_1, \cdot) + i\xi_1^2\chi(t_0, t, \cdot), \omega(\cdot))\right](\eta_1)\right](\tau_1)ds\right](\xi_1)\right](x_1) -$$

$$- i\int_{t_0}^{t}\frac{1}{4\pi(t-s)}\exp\left(-\frac{x_2^2+x_3^2}{4(t-s)}\right)\delta(x_1) \overset{x}{*} F_{\xi_1}^{-1}\left[F_{x_1}\left[\frac{\delta}{\delta\omega(s,x)}y(t_1, \tau, v(\cdot) + i\xi_1^2\chi(s, t, \cdot), \omega(\cdot))\right](\xi_1)\right](x_1)ds =$$

$$= \frac{1}{4\pi(t-t_0)}\exp\left(-\frac{x_2^2+x_3^2}{4(t-t_0)}\right) \overset{x_2x_3}{*} \left\{\frac{1}{4\pi(t_1-t_0)}\exp\left(-\frac{\tau_2^2+\tau_3^2}{4(t_1-t_0)}\right) \overset{\tau_2\tau_3}{*} \{M(u_0(x)u_0(\tau)) \overset{\tau_1}{*} F_{\xi_1}^{-1}[F_{\eta_1}^{-1}[\varphi(v(\cdot) +\right.\right.$$

$$\left.\left. + i\eta_1^2\chi(t_0, t_1, \cdot) + i\xi_1^2\chi(t_0, t, \cdot), \omega(\cdot))](\tau_1)](x_1)\}\right\} -$$

$$- i\frac{1}{4\pi(t-t_0)}\exp\left(-\frac{x_2^2+x_3^2}{4(t-t_0)}\right) \overset{x_2x_3}{*} \left\{M(u_0(x)) \overset{x_1}{*} \int_{t_0}^{t_1}\frac{1}{4\pi(t_1-s)}\exp\left(-\frac{\tau_2^2+\tau_3^2}{4(t_1-s)}\right) \overset{\tau_2\tau_3}{*} F_{\xi_1}^{-1}\left[F_{\eta_1}^{-1}\left[F_{x_1}\left[\frac{\delta}{\delta\omega(s,\tau)}\varphi(v(\cdot) +\right.\right.\right.\right.$$

$$\left.\left.\left.\left. + i\eta_1^2\chi(s, t_1, \cdot) + i\xi_1^2\chi(t_0, t, \cdot), \omega(\cdot))\right](\eta_1)\right](\tau_1)\right](x_1)ds\right\} -$$

$$- i\int_{t_0}^{t}\frac{1}{4\pi(t-s)}\exp\left(-\frac{x_2^2+x_3^2}{4(t-s)}\right) \overset{x_2x_3}{*} F_{\xi_1}^{-1}\left[F_{x_1}\left[\frac{\delta}{\delta\omega(s,x)}y(t_1, \tau, v(\cdot) + i\xi_1^2\chi(s, t, \cdot), \omega(\cdot))\right](\xi_1)\right](x_1)ds.$$

Согласно (14),

$$y(t_1, \tau, v(\cdot) + i\xi_1^2\chi(s, t, \cdot), \omega(\cdot)) = \frac{1}{4\pi(t_1-t_0)}\exp\left(-\frac{\tau_2^2+\tau_3^2}{4(t_1-t_0)}\right) \overset{\tau_2\tau_3}{*} (M(u_0(\tau))) \overset{\tau_1}{*} F_{\eta_1}^{-1}[\varphi(v(\cdot) +$$

$$+ i\eta_1^2\chi(t_0, t_1, \cdot) + i\xi_1^2\chi(s, t, \cdot), \omega(\cdot))](\tau_1) - i\int_{t_0}^{t_1}\frac{1}{4\pi(t_1-\theta)} \times$$

$$\times \exp\left(-\frac{\tau_2^2+\tau_3^2}{4(t_1-\theta)}\right) \overset{\tau_2\tau_3}{*} F_{\eta_1}^{-1}\left[F_{x_1}\left[\frac{\delta\varphi(v(\cdot) + i\eta_1^2\chi(\theta, t_1, \cdot) + i\xi_1^2\chi(s, t, \cdot), \omega(\cdot))}{\delta\omega(\theta, \tau)}\right](\eta_1)\right](\tau_1)d\theta.$$

Подставив это выражение и использовав свойства преобразования Фурье, найдем

$$z = \frac{1}{4\pi(t-t_0)}\exp\left(-\frac{x_2^2+x_3^2}{4(t-t_0)}\right)^{x_2x_3}_{*}\left\{\frac{1}{4\pi(t_1-t_0)}\exp\left(-\frac{\tau_2^2+\tau_3^2}{4(t_1-t_0)}\right)^{\tau_2\tau_3}_{*}\{M(u_0(x)u_0(\tau))^{\tau_1}_{*}F^{-1}_{\xi_1}[F^{-1}_{\eta_1}[\varphi(v(\cdot)\right.\right.$$

$$\left.\left.+i\eta_1^2\chi(t_0,t_1,\cdot)+i\xi_1^2\chi(t_0,t,\cdot),\omega(\cdot))](\tau_1)](x_1)\}\right\}-i\frac{1}{4\pi(t-t_0)}\exp\left(-\frac{x_2^2+x_3^2}{4(t-t_0)}\right)^{x_2x_3}_{*}\left\{M(u_0(x))^{x_1}_{*}\int_{t_0}^{t_1}\frac{1}{4\pi(t_1-s)}\times\right.$$

$$\left.\times\exp\left(-\frac{\tau_2^2+\tau_3^2}{4(t_1-s)}\right)^{\tau_2\tau_3}_{*}F^{-1}_{\xi_1}\left[F^{-1}_{\eta_1}\left[F_{x_1}\left[\frac{\delta}{\delta\omega(s,\tau)}\varphi(v(\cdot)+i\eta_1^2\chi(s,t_1,\cdot)+i\xi_1^2\chi(t_0,t,\cdot),\omega(\cdot))\right](\eta_1)\right](\tau_1)\right](x_1)ds\right\}-$$

$$-i\frac{1}{4\pi(t_1-t_0)}\exp\left(-\frac{\tau_2^2+\tau_3^2}{4(t_1-t_0)}\right)^{\tau_2\tau_3}_{*}\left\{M(u_0(\tau))^{\tau_1}_{*}\int_{t_0}^{t}\frac{1}{4\pi(t-s)}\exp\left(-\frac{x_2^2+x_3^2}{4(t-s)}\right)^{x_2x_3}_{*}F^{-1}_{\eta_1}\left[F^{-1}_{\xi_1}\left[F_{x_1}\left[\frac{\delta}{\delta\omega(s,x)}\varphi(v(\cdot)\right.\right.\right.\right.$$

$$\left.\left.\left.\left.+i\eta_1^2\chi(t_0,t_1,\cdot)+i\xi_1^2\chi(s,t,\cdot),\omega(\cdot)\right](\xi_1)\right](x_1)\right](\tau_1)ds\right\}-\int_{t_0}^{t}ds\int_{t_0}^{t_1}\frac{1}{4\pi(t_1-s)}\times$$

$$\times\exp\left(-\frac{x_2^2+x_3^2}{4(t-s)}\right)^{x_2x_3}_{*}\left\{\frac{1}{4\pi(t_1-\theta)}\exp\left(-\frac{\tau_2^2+\tau_3^2}{4(t_1-\theta)}\right)^{\tau_2\tau_3}_{*}F^{-1}_{\xi_1}\left[F_{x_1}\left[F^{-1}_{\eta_1}\left[F_{x_1}\times\right.\right.\right.\right.$$

$$\left.\left.\left.\left.\times\left[\frac{\delta^2\varphi(v(\cdot)+i\eta_1^2\chi(\theta,t_1,\cdot)+i\xi_1^2\chi(s,t,\cdot),\omega(\cdot))}{\delta\omega(s,x)\delta\omega(\theta,\tau)}\right](\eta_1)\right](\tau_1)\right](\xi_1)\right](x_1)\right\}d\theta.$$

В третьем от конца слагаемом переменные, по которым вычисляются обратные преобразования Фурье, переобозначим: η положим равным ξ и наоборот. При этом получим (20). Теорема доказана.

6. ВТОРАЯ МОМЕНТНАЯ ФУНКЦИЯ РЕШЕНИЯ ЗАДАЧИ (1), (2)

Определение 3. *Обобщенной второй моментной функцией* $M(u(t,x)u(t_1,\tau))$ задачи (1), (2) называется $z(t,t_1,x,\tau,0,0)$, где z – обобщенное симметричное по (t,x) и (t_1,τ) решение задачи (18), (19).

Теорема 9. *Пусть выполняются условия теоремы* 6, *тогда обобщенная вторая моментная функция решения задачи* (1), (2) *находится по формуле*

$$M(u(t,x)u(t_1,\tau)) = \frac{1}{4\pi(t-t_0)}\exp\left(-\frac{x_2^2+x_3^2}{4(t-t_0)}\right)^{x_2x_3}_{*}\left\{\frac{1}{4\pi(t_1-t_0)}\times\right.$$

$$\left.\times\exp\left(-\frac{\tau_2^2+\tau_3^2}{4(t_1-t_0)}\right)^{\tau_2\tau_3}_{*}\{M(u_0(x)u_0(\tau))^{\tau_1}_{*}F^{-1}_{\xi_1}[F^{-1}_{\eta_1}[\varphi(i\eta_1^2\chi(t_0,t_1,\cdot)+i\xi_1^2\chi(t_0,t,\cdot),0)](\tau_1)](x_1)\}\right\}-$$

$$-i\frac{1}{4\pi(t-t_0)}\exp\left(-\frac{x_2^2+x_3^2}{4(t-t_0)}\right)^{x_2x_3}_{*}\left\{M(u_0(x))^{x_1}_{*}\int_{t_0}^{t_1}\frac{1}{4\pi(t_1-s)}\exp\left(-\frac{\tau_2^2+\tau_3^2}{4(t_1-s)}\right)^{\tau_2\tau_3}_{*}F^{-1}_{\xi_1}\left[F^{-1}_{\eta_1}\left[F_{x_1}\times\right.\right.\right.$$

$$\times\left[\frac{\delta}{\delta\omega(s,\tau)}\varphi(i\eta_1^2\chi(s,t_1,\cdot)+i\xi_1^2\chi(t_0,t,\cdot),0)\right](\eta_1)\Big](\tau_1)\Big](x_1)ds\Bigg\}-i\frac{1}{4\pi(t-t_0)}\times$$

$$\times\exp\left(-\frac{\tau_2^2+\tau_3^2}{4(t_1-t_0)}\right)^{\tau_2\tau_3}_{*}\Bigg\{M(u_0(\tau))\overset{\tau_1}{*}\int_{t_0}^{t}\frac{1}{4\pi(t-s)}\exp\left(-\frac{x_2^2+x_3^2}{4(t-s)}\right)^{x_2x_3}_{*}F_{\eta_1}^{-1}\Big[F_{\xi_1}^{-1}\Big[F_{x_1}\times$$

$$\times\left[\frac{\delta}{\delta\omega(s,x)}\varphi(i\eta_1^2\chi(t_0,t_1,\cdot)+i\xi_1^2\chi(s,t,\cdot),0)\right](\xi_1)\Big](x_1)\Big](\tau_1)ds\Bigg\}-\int_{t_0}^{t}ds\int_{t_0}^{t_1}\frac{1}{4\pi(t-s)}\times$$

$$\times\exp\left(-\frac{x_2^2+x_3^2}{4(t-s)}\right)^{x_2x_3}_{*}\Bigg\{\frac{1}{4\pi(t_1-\theta)}\exp\left(-\frac{\tau_2^2+\tau_3^2}{4(t_1-\theta)}\right)^{\tau_2\tau_3}_{*}F_{\xi_1}^{-1}\Big[F_{x_1}\Big[F_{\eta_1}^{-1}\Big[F_{x_1}\times$$

$$\times\left[\frac{\delta^2\varphi(i\eta_1^2\chi(\theta,t_1,\cdot)+i\xi_1^2\chi(s,t,\cdot),0)}{\delta\omega(s,x)\delta\omega(\theta,\tau)}\right](\eta_1)\Big](\tau_1)\Big](\xi_1)\Big](x_1)\Bigg\}d\theta. \tag{21}$$

Доказательство. Воспользовавшись предыдущим определением и положив в (20) $v=0$, $\omega=0$, получим (21).

Теорема 10. *Пусть в задаче* (1), (2) *случайные процессы* u_0, ε *и* f *независимы, выполнено условие* (4), *характеристический функционал* φ_ε *процесса* ε *имеет в окрестности точки* $v=0$ *вариационную производную* $\dfrac{\delta\varphi(v(\cdot)+i\eta_1^2\chi(\theta,t_1,\cdot)+i\xi_1^2\chi(s,t,\cdot))}{\delta v(s)}$ *и функции* $M(u_0(x))$, $M(u_0(\tau))$, $M(u_0(x)u_0(\tau))$, $M(f(t,x)f(t_1,\tau))$ *локально суммируемы. Тогда вторая обобщенная моментная функция решения задачи* (1), (2) *имеет вид*

$$M(u(t,x)u(t_1,\tau))=\frac{1}{4\pi(t-t_0)}\exp\left(-\frac{x_2^2+x_3^2}{4(t-t_0)}\right)^{x_2x_3}_{*}\Bigg\{\frac{1}{4\pi(t_1-t_0)}\times$$

$$\times\exp\left(-\frac{\tau_2^2+\tau_3^2}{4(t_1-t_0)}\right)^{\tau_2\tau_3}_{*}\{M(u_0(x)u_0(\tau))\overset{\tau_1}{*}F_{\xi_1}^{-1}[F_{\eta_1}^{-1}[\varphi_\varepsilon(i\eta_1^2\chi(t_0,t_1,\cdot)+i\xi_1^2\chi(t_0,t,\cdot))](\tau_1)](x_1)\}\Bigg\}+$$

$$+\frac{1}{4\pi(t-t_0)}\exp\left(-\frac{x_2^2+x_3^2}{4(t-t_0)}\right)^{x_2x_3}_{*}\Bigg\{M(u_0(x))\overset{x_1}{*}\int_{t_0}^{t_1}\frac{1}{4\pi(t_1-s)}\exp\left(-\frac{\tau_2^2+\tau_3^2}{4(t_1-s)}\right)^{\tau_2\tau_3}_{*}\{F_{\xi_1}^{-1}[F_{\eta_1}^{-1}\times$$

$$\times[\varphi_\varepsilon(i\eta_1^2\chi(s,t_1,\cdot)+i\xi_1^2\chi(t_0,t,\cdot))](\tau_1)](x_1)\overset{x_1}{*}M(f(s,\tau))\}ds\Bigg\}+\frac{1}{4\pi(t_1-t_0)}\times$$

$$\times\exp\left(-\frac{\tau_2^2+\tau_3^2}{4(t_1-t_0)}\right)^{\tau_2\tau_3}_{*}\Bigg\{M(u_0(\tau))\overset{\tau_1}{*}\int_{t_0}^{t}\frac{1}{4\pi(t-s)}\exp\left(-\frac{x_2^2+x_3^2}{4(t-s)}\right)^{x_2x_3}_{*}\{F_{\eta_1}^{-1}[F_{\xi_1}^{-1}\times$$

$$\times[\varphi_\varepsilon(i\eta_1^2\chi(t_0,t_1,\cdot)+i\xi_1^2\chi(s,t,\cdot))](x_1)](\tau_1)\overset{x_1}{*}M(f(s,x))\}ds\Bigg\}+\int_{t_0}^{t}ds\int_{t_0}^{t_1}\frac{1}{4\pi(t-s)}\times \tag{22}$$

$$\times \exp\left(-\frac{x_2^2+x_3^2}{4(t-s)}\right)^{x_2x_3}_{*}\left\{\frac{1}{4\pi(t_1-\theta)}\exp\left(-\frac{\tau_2^2+\tau_3^2}{4(t_1-\theta)}\right)^{\tau_2\tau_3}_{*}\{F^{-1}_{\xi_1}[F^{-1}_{\eta_1}\times\right.$$

$$\left.\times[\varphi_\varepsilon(i\eta_1^2\chi(\theta,t_1,\cdot)+i\xi_1^2\chi(s,t,\cdot))](\tau_1)](x_1)^{\tau_1}_{*}M(f(s,x)f(\theta,\tau))\}\right\}d\theta.$$

Доказательство. Так как случайные процессы ε и f независимы, то $\varphi(v(\cdot),\ \omega(\cdot)) = \varphi_\varepsilon(v(\cdot))\varphi_f(\omega(\cdot))$, где $\varphi_\varepsilon(v(\cdot))$ и $\varphi_f(\omega(\cdot))$ – соответственно, характеристические функционалы ε и f. Далее,

$$\frac{\delta\varphi_f(0)}{\delta\omega(s,\tau)} = iMf(s,\tau),\quad \varphi_f(0)=1,\quad \frac{\delta^2\varphi_f(0)}{\delta\omega(t,x)\delta\omega(t_1,\tau)} = -M(f(t,x)f(t_1,\tau)).$$

Используя эти соотношения и свойства преобразования Фурье, получаем

$$-i\frac{1}{4\pi(t-t_0)}\exp\left(-\frac{x_2^2+x_3^2}{4(t-t_0)}\right)^{x_2x_3}_{*}\left\{M(u_0(x))^{x_1}_{*}\int_{t_0}^{t_1}\frac{1}{4\pi(t_1-s)}\exp\left(-\frac{\tau_2^2+\tau_3^2}{4(t_1-s)}\right)^{\tau_2\tau_3}_{*}F^{-1}_{\xi_1}\left[F^{-1}_{\eta_1}\left[F_{x_1}\times\right.\right.\right.$$

$$\left.\left.\left.\times\left[\varphi_\varepsilon(i\eta_1^2\chi(s,t_1,\cdot)+i\xi_1^2\chi(t_0,t,\cdot))\frac{\delta}{\delta\omega(s,\tau)}\varphi_f(0)\right](\eta_1)\right](\tau_1)\right](x_1)ds\right\} =$$

$$= \frac{1}{4\pi(t-t_0)}\exp\left(-\frac{x_2^2+x_3^2}{4(t-t_0)}\right)^{x_2x_3}_{*}\left\{M(u_0(x))^{x_1}_{*}\int_{t_0}^{t_1}\frac{1}{4\pi(t_1-s)}\exp\left(-\frac{\tau_2^2+\tau_3^2}{4(t_1-s)}\right)^{\tau_2\tau_3}_{*}F^{-1}_{\xi_1}[F^{-1}_{\eta_1}\times\right.$$

$$\left.\times[\varphi_\varepsilon(i\eta_1^2\chi(s,t_1,\cdot)+i\xi_1^2\chi(t_0,t,\cdot))F_{x_1}[M(f(s,\tau))](\eta_1)](\tau_1)](x_1)ds\right\} =$$

$$= \frac{1}{4\pi(t-t_0)}\exp\left(-\frac{x_2^2+x_3^2}{4(t-t_0)}\right)^{x_2x_3}_{*}\left\{M(u_0(x))^{x_1}_{*}\int_{t_0}^{t_1}\frac{1}{4\pi(t_1-s)}\exp\left(-\frac{\tau_2^2+\tau_3^2}{4(t_1-s)}\right)^{\tau_2\tau_3}_{*}(F^{-1}_{\xi_1}[F^{-1}_{\eta_1}\times\right.$$

$$\left.\times[\varphi_\varepsilon(i\eta_1^2\chi(s,t_1,\cdot)+i\xi_1^2\chi(t_0,t,\cdot))(\tau_1)](x_1)^{x_1}_{*}M(f(s,\tau)))ds\right\};$$

$$-i\frac{1}{4\pi(t_1-t_0)}\exp\left(-\frac{x_2^2+x_3^2}{4(t_1-t_0)}\right)^{\tau_2\tau_3}_{*}\left\{M(u_0(\tau))^{\tau_1}_{*}\int_{t_0}^{t}\frac{1}{4\pi(t-s)}\exp\left(-\frac{x_2^2+x_3^2}{4(t-s)}\right)^{x_2x_3}_{*}F^{-1}_{\eta_1}\left[F^{-1}_{\xi_1}\left[F_{x_1}\times\right.\right.\right.$$

$$\left.\left.\left.\times\left[\varphi_\varepsilon(i\eta_1^2\chi(t_0,t_1,\cdot)+i\xi_1^2\chi(s,t,\cdot))\frac{\delta}{\delta\omega(s,x)}\varphi_f(0)\right](\xi_1)\right](x_1)\right](\tau_1)ds\right\} =$$

$$= \frac{1}{4\pi(t_1-t_0)}\exp\left(-\frac{\tau_2^2+\tau_3^2}{4(t_1-t_0)}\right)^{\tau_2\tau_3} * \left\{ M(u_0(\tau))^{\tau_1} * \int_{t_0}^{t} \frac{1}{4\pi(t-s)}\exp\left(-\frac{x_2^2+x_3^2}{4(t-s)}\right)^{x_2x_3} * F_{\eta_1}^{-1}[F_{\xi_1}^{-1} \times \right.$$

$$\left. \times [\varphi_\varepsilon(i\eta_1^2\chi(t_0, t_1, \cdot) + i\xi_1^2\chi(s, t, \cdot))F_{x_1}[M(f(s, x))](\xi_1)](x_1)](\tau_1)ds \right\} =$$

$$= \frac{1}{4\pi(t_1-t_0)}\exp\left(-\frac{\tau_2^2+\tau_3^2}{4(t_1-t_0)}\right)^{\tau_2\tau_3} * \left\{ M(u_0(\tau))^{\tau_1} * \int_{t_0}^{t} \frac{1}{4\pi(t-s)}\exp\left(-\frac{x_2^2+x_3^2}{4(t-s)}\right)^{x_2x_3} * \{ F_{\eta_1}^{-1}[F_{\xi_1}^{-1} \times \right.$$

$$\left. \times [\varphi_\varepsilon(i\eta_1^2\chi(t_0, t_1, \cdot) + i\xi_1^2\chi(s, t, \cdot))](x_1)](\tau_1)^{x_1} * M(f(s, x))\}ds \right\};$$

$$-\int_{t_0}^{t} ds \int_{t_0}^{t_1} \frac{1}{4\pi(t-s)}\exp\left(-\frac{x_2^2+x_3^2}{4(t-s)}\right)^{x_2x_3} * \left\{ \frac{1}{4\pi(t_1-\theta)}\exp\left(-\frac{\tau_2^2+\tau_3^2}{4(t_1-\theta)}\right)^{\tau_2\tau_3} * F_{\xi_1}^{-1}\left[F_{x_1}\left[F_{\eta_1}^{-1}\left[F_{x_1} \times \right.\right.\right.\right.$$

$$\left.\left.\left.\left. \times \left[\varphi_\varepsilon(i\eta_1^2\chi(\theta, t_1, \cdot) + i\xi_1^2\chi(s, t, \cdot))\frac{\delta^2\varphi_f(0)}{\delta\omega(s, x)\delta\omega(\theta, \tau)}\right](\eta_1)\right](\tau_1)\right](\xi_1)\right](x_1) \right\}d\theta =$$

$$= \int_{t_0}^{t} ds \int_{t_0}^{t_1} \frac{1}{4\pi(t-s)}\exp\left(-\frac{x_2^2+x_3^2}{4(t-s)}\right)^{x_2x_3} * \left\{ \frac{1}{4\pi(t_1-\theta)}\exp\left(-\frac{\tau_2^2+\tau_3^2}{4(t_1-\theta)}\right)^{\tau_2\tau_3} * F_{\xi_1}^{-1}[F_{\eta_1}^{-1} \times \right.$$

$$\left. \times [\varphi_\varepsilon(i\eta_1^2\chi(\theta, t_1, \cdot) + i\xi_1^2\chi(s, t, \cdot))F_{x_1}[F_{\tau_1}[M(f(s, x)f(\theta, \tau))](\eta_1)](\tau_1)](\xi_1)](x_1) \right\}d\theta =$$

$$= \int_{t_0}^{t} ds \int_{t_0}^{t_1} \frac{1}{4\pi(t-s)}\exp\left(-\frac{x_2^2+x_3^2}{4(t-s)}\right)^{x_2x_3} * \left\{ \frac{1}{4\pi(t_1-\theta)}\exp\left(-\frac{\tau_2^2+\tau_3^2}{4(t_1-\theta)}\right)^{\tau_2\tau_3} * \{ F_{\xi_1}^{-1}[F_{\eta_1}^{-1} \times \right.$$

$$\left. \times [\varphi_\varepsilon(i\eta_1^2\chi(\theta, t_1, \cdot) + i\xi_1^2\chi(s, t, \cdot))](\tau_1)](x_1)^{\tau_1} * M(f(s, x)f(\theta, \tau))\} \right\}d\theta.$$

Подставляя эти соотношения в формулу (21) получаем (22). Теорема доказана.

7. ВТОРАЯ СМЕШАННАЯ МОМЕНТНАЯ ФУНКЦИЯ РЕШЕНИЯ ЗАДАЧИ (1), (2)

Полученная формула для отображения $y(t, x, v(\cdot), \omega(\cdot))$ позволяет вариационным дифференцированием находить смешанные моментные функции, например $M(u(t, x)\varepsilon(\tau)) = -i\dfrac{\delta y(t, x, 0, 0)}{\delta v(\tau)}$.

Теорема 11. *Пусть случайный процесс $\varepsilon(t)$ не зависит от процесса $f(t, x)$, $\varphi_\varepsilon(v(\cdot))$ – характеристический функционал процесса $\varepsilon(t)$ – имеет непрерывные по $v(\cdot)$ вариационные производные до второго порядка включительно, $M(u_0(x))$ суммируемо на $\mathbb{R}^3$, $M(f(t, x))$ суммируемо на $T\times\mathbb{R}^3$. Тогда*

$$M(u(t, x)\varepsilon(s)) = -i\frac{1}{4\pi(t-t_0)}\exp\left(-\frac{x_2^2+x_3^2}{4(t-t_0)}\right)^{x_2x_3} * \left(M(u_0(x))^{x_1} * F_{\xi_1}^{-1}\left[\frac{\delta\varphi_\varepsilon(i\xi_1^2\chi(t_0, t, \cdot))}{\delta v(\tau)}\right](x_1)\right) -$$

$$-i\int_{t_0}^{t} \frac{1}{4\pi(t-s)}\exp\left(-\frac{x_2^2+x_3^2}{4(t-s)}\right)^{x_2x_3} * \left\{ F_{\xi_1}^{-1}\left[F_{x_1}\left[\frac{\delta\varphi_\varepsilon(i\xi_1^2\chi(s, t, \cdot))}{\delta v(\tau)}\right](\xi_1)\right](x_1)^{x_1} * M(f(s, x)) \right\}ds.$$

Авторы благодарны А.В. Фурсикову за полезное обсуждение результатов.

СПИСОК ЛИТЕРАТУРЫ

1. *Фурсиков А.В.* О проблеме замыкания цепочки моментных уравнений в случае больших чисел Рейнольдса // Неклассич. ур-ния и ур-ния смешанного типа. Новосибирск: Ин-т матем. СО АН СССР, 1990. С. 231–250.
2. *Вентцель А.Д., Фрейндлин М.Н.* Флуктуации в динамических системах под действием малых случайных возмущений. М.: Наука, 1979.
3. *Адомиан Дж.* Стохастические системы. М.: Мир, 1987.
4. *Кляцкин В.И.* Стохастические уравнения и волны в случайно-неоднородных средах. М.: Наука, 1980.
5. *Задорожний В.Г.* Методы вариационного анализа. М.–Ижевск: НИЦ РХД, 2006.
6. *Колмогоров А.Н., Фомин С.В.* Элементы теории функций и функционального анализа. М.: Наука, 1972.
7. *Владимиров В.С.* Обобщенные функции в математической физике. М.: Наука, 1976.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2009, том 49, № 11, с. 1953–1969

УДК 519.634

АНАЛИЗ ВЫЧИСЛИТЕЛЬНЫХ СВОЙСТВ КВАЗИГАЗОДИНАМИЧЕСКОГО АЛГОРИТМА НА ПРИМЕРЕ РЕШЕНИЯ УРАВНЕНИЙ ЭЙЛЕРА

© 2009 г. Т. Г. Елизарова, Е. В. Шильников

(125047 Москва, Миусская пл., 4а, ИММ РАН)

e-mail: telizar@mail.ru; shiva@imamod.ru

Поступила в редакцию 19.01.2009 г.
Переработанный вариант 02.02.2009 г.

На примере численного решения задач Римана о распаде сильных разрывов и задачи о распространении звукового возмущения проведен анализ вычислительных свойств алгоритма, основанного на квазигазодинамических уравнениях. Показано, что данный алгоритм позволяет единообразно рассчитывать как задачи о распаде разрыва с большими перепадами плотности и давления, так и задачи о распространении акустических возмущений. Численно получены условия устойчивости и оценки точности и вычислительной сложности разностной схемы. Библ. 22. Илл. 16.

Ключевые слова: квазигазодинамический алгоритм, точность, устойчивость, вычислительная сложность, задача Римана, акустические возмущения, уравнения Эйлера.

ВВЕДЕНИЕ

Как правило, численные алгоритмы для моделирования газодинамических течений строятся на основе последовательного усложнения рассматриваемых моделей. Вначале предлагается эффективный метод решения упрощенных уравнений, например, уравнения переноса или уравнения Бюргерса, затем метод обобщается на одномерную систему уравнений газовой динамики, потом на систему, включающую в себя процессы диссипации, затем на многомерные задачи, неструктурированные сетки и так далее, в зависимости от поставленной цели. При построении кавазигазодинамических (КГД) уравнений был сразу предложен численный алгоритм для общего случая, а именно для расчета вязких нестационарных пространственных течений. При этом соответствующая система уравнений была выписана сразу в инвариантном виде. И только впоследствии алгоритм был редуцирован для численного моделирования двумерных и одномерных задач на равномерных сетках. Такой подход от общего к частному и предопределил тот факт, что расчетам одномерных течений до сих пор было уделено недостаточное внимание.

В данной работе на примере задачи Римана о распаде разрыва численно исследуются условия устойчивости и точности алгоритма, основанного на КГД-системе уравнений, и приводятся оценки его вычислительной сложности. Кроме того продемонстрированы решения задачи о распаде разрыва с большими значениями перепадов плотности и давления в начальный момент времени и задачи о распространении звукового возмущения. Данные примеры являются сложными тестами и демонстрируют широкий спектр применимости данного алгоритма. Представительная система тестовых расчетов, выполненных в рамках КГД-алгоритма, приведена также в [1].

Отметим, что ранее КГД-алгоритм (см., например, [2], [3]) и родственные ему кинетически-согласованные разностные схемы (см. [4]) успешно использовались для численного моделирования широкого круга течений вязкого сжимаемого газа как в двумерных, так и в пространственных приближениях.

1. СИСТЕМА КВАЗИГАЗОДИНАМИЧЕСКИХ УРАВНЕНИЙ

Система КГД-уравнений в традиционных обозначениях с учетом внешних сил F_i и внешних источников тепла Q согласно [2], [3] и [5] записывается в виде законов сохранения следующим образом:

$$\frac{\partial}{\partial t}\rho + \frac{\partial}{\partial x_i} j_{mi} = 0, \tag{1}$$

$$\frac{\partial}{\partial t}\rho u_i + \frac{\partial}{\partial x_j} j_{mj} u_i + \frac{\partial}{\partial x_i} p = \rho_* F_i + \frac{\partial}{\partial x_j}\Pi_{ji}, \tag{2}$$

$$\frac{\partial}{\partial t}\rho\left(\frac{u^2}{2} + \varepsilon\right) + \frac{\partial}{\partial x_i} j_{mi}\left(\frac{u^2}{2} + \varepsilon + \frac{p}{\rho}\right) + \frac{\partial}{\partial x_i} q_i = j_{mi} F_i + \frac{\partial}{\partial x_i}\Pi_{ij} u_j + Q, \tag{3}$$

где введены обозначения

$$j_{mi} = \rho(u_i - w_i), \tag{4}$$

$$w_i = \frac{\tau}{\rho}\left(\frac{\partial}{\partial x_j}\rho u_i u_j + \frac{\partial}{\partial x_i} p - \rho F_i\right), \tag{5}$$

$$\rho_* = \left(\rho - \tau\frac{\partial}{\partial x_k}\rho u_k\right), \tag{6}$$

$$\Pi_{ij} = \Pi_{\mathrm{NS}ij} + \tau\rho u_i\left(u_k\frac{\partial}{\partial x_k}u_j + \frac{1}{\rho}\frac{\partial}{\partial x_j}p - F_j\right) + \tau\delta_{ij}\left(u_k\frac{\partial}{\partial x_k}p + \gamma p\frac{\partial}{\partial x_k}u_k - (\gamma-1)Q\right), \tag{7}$$

$$q_i = q_{\mathrm{NS}i} - \tau\rho u_i\left(u_j\frac{\partial}{\partial x_j}\varepsilon + p u_j\frac{\partial}{\partial x_j}\frac{1}{\rho} - \frac{Q}{\rho}\right). \tag{8}$$

По повторяющимся индексам здесь и далее подразумевается суммирование. Уравнения связи имеют вид

$$p = \rho R T, \quad \varepsilon = \frac{p}{\rho(\gamma-1)}. \tag{9}$$

Тепловой поток и тензор вязких напряжений Навье–Стокса вычисляются по формулам

$$q_{\mathrm{NS}i} = -\kappa\frac{\partial}{\partial x_i} T, \tag{10}$$

$$\Pi_{\mathrm{NS}ij} = \mu\left(\frac{\partial}{\partial x_i}u_j + \frac{\partial}{\partial x_j}u_i - \frac{2}{3}\delta_{ij}\frac{\partial}{\partial x_k}u_k\right). \tag{11}$$

$$\mu = \mu_\infty\left(\frac{T}{T_\infty}\right)^\omega, \quad \kappa = \mu\frac{\gamma R}{(\gamma-1)\mathrm{Pr}}, \quad \tau = \frac{\mu}{p\mathrm{Sc}}, \tag{12}$$

где μ – коэффициент динамической вязкости, κ – коэффициент теплопроводности, τ – релаксационный параметр, имеющий размерность времени, γ – показатель адиабаты, Pr – число Прандтля, Sc – число Шмидта. Вопросы корректности и физической адекватности указанной системы и ее упрощенных вариантов исследовались в [2]–[7].

Заметим, что диссипативные слагаемые, образующие КГД-добавки с коэффициентом τ к потоку массы (5), тензору вязких напряжений (7) и тепловому потоку (8), обращаются в ноль в областях течения, где решение удовлетворяет стационарным уравнениям Эйлера.

Уравнение баланса энтропии s для КГД-системы имеет вид

$$\frac{\partial}{\partial t}\rho s + \frac{\partial}{\partial x_i} j_{mi} s = -\frac{\partial}{\partial x_i}\frac{q_i}{T} + \kappa\left(\frac{1}{T}\frac{\partial}{\partial x_i}T\right)^2 + \frac{\Phi}{T}, \tag{13}$$

где диссипативную функцию Φ удается представить выражением

$$\begin{aligned}\Phi = \frac{\Pi_{\mathrm{NS}ij}\Pi_{\mathrm{NS}ij}}{2\mu} + \tau\rho\left(u_k\frac{\partial}{\partial x_k}u_i + \frac{1}{\rho}\frac{\partial}{\partial x_i}p - F_i\right)^2 + \frac{\tau p}{\rho^2}\left(\frac{\partial}{\partial x_i}\rho u_i\right)^2 + \\ + \frac{\tau\rho}{\varepsilon}\left(u_i\frac{\partial}{\partial x_i}\varepsilon + \frac{p}{\rho}\frac{\partial}{\partial x_i}u_i - \frac{Q}{2\rho}\right)^2 + Q\left(1 - \frac{(\gamma-1)\tau Q}{4p}\right).\end{aligned} \tag{14}$$

Условие неотрицательности последнего слагаемого в (14) накладывает ограничение на величину параметра τ в зависимости от интенсивности внешних источников тепла[1)].

[1)] Окончательный вид двух последних слагаемых получен А.А. Злотником. Частное сообщение.

2. ЧИСЛЕННЫЙ АЛГОРИТМ РАСЧЕТА ОДНОМЕРНЫХ ТЕЧЕНИЙ

Для удобства численного решения система уравнений (1)–(3) приводится к безразмерному виду с использованием базовых значений плотности ρ_0, скорости звука $c_0 = \sqrt{\gamma R T_0}$ и длины L. Обезразмеривание не изменяет вида уравнений.

Будем решать КГД-уравнения для плоских одномерных течений без учета внешних сил и источников тепла ($Q = 0$, $F_i = 0$). При этом система (1)–(3) существенно упростится.

Введем равномерную сетку по координате x с шагом h и по времени с шагом Δt.

При численном решении уравнений Эйлера на основе системы (1)–(3) все диссипативные слагаемые, т.е. слагаемые с коэффициентами μ, κ и τ, рассматриваются как регуляризаторы. В этом случае релаксационный параметр и коэффициент вязкости и теплопроводности оказываются связанными между собой и в безразмерном виде вычисляются по формулам

$$\tau = \alpha\frac{h}{c}, \quad \mu = \tau p \mathrm{Sc}, \quad \kappa = \frac{\tau p \mathrm{Sc}}{\mathrm{Pr}(\gamma - 1)}, \tag{15}$$

где $c = \sqrt{\gamma p/\rho}$ – локальное значение скорости звука. При этом числа Pr и Sc рассматриваются как числовые коэффициенты для настройки искусственной вязкости, если такая настройка необходима.

Значения всех газодинамических величин – скорости, плотности, давления – определяются в узлах сетки. Значения потоков определяются в полуцелых точках. Согласно [3], [6] выпишем явную по времени однородную разностную схему на пространственном шаблоне, включающем в себя три точки:

$$\hat{\rho}_i = \rho_i - \frac{\Delta t}{h}(j_{mi+1/2} - j_{mi-1/2}), \tag{16}$$

$$\widehat{\rho_i u_i} = \rho_i u_i + \frac{\Delta t}{h}[(\Pi_{i+1/2} - \Pi_{i-1/2}) - (p_{i+1/2} - p_{i-1/2}) - (j_{mi+1/2}u_{i+1/2} - j_{mi-1/2}u_{i-1/2})], \tag{17}$$

$$\hat{E}_i = E_i + \frac{\Delta t}{h}\Bigg[(\Pi_{i+1/2}u_{i+1/2} - \Pi_{i-1/2}u_{i-1/2}) - (q_{i+1/2} - q_{i-1/2}) - \\ - \left(\frac{j_{mi+1/2}}{\rho_{i+1/2}}(E_{i+1/2} + p_{i+1/2}) - \frac{j_{mi-1/2}}{\rho_{i-1/2}}(E_{i-1/2} + p_{i-1/2})\right)\Bigg], \tag{18}$$

$$p_i = (\gamma - 1)\left(E_i - \frac{\rho_i u_i^2}{2}\right).$$

Здесь $E_i = \rho_i u_i^2/2 + p_i/(\gamma - 1)$ – полная энергия единицы объема. Дискретный аналог потока массы $j_{mi+1/2}$ рассчитывается следующим образом:

$$w_{i+1/2} = \frac{\tau_{i+1/2}}{\rho_{i+1/2}}\frac{1}{h}(\rho_{i+1}u_{i+1}^2 + p_{i+1} - \rho_i u_i^2 - p_i), \tag{19}$$

$$j_{mi+1/2} = \rho_{i+1/2}(u_{i+1/2} - w_{i+1/2}). \tag{20}$$

Дискретное выражение для $\Pi_{i+1/2}$ вычисляется по формулам

$$\mu_{i+1/2} = \tau_{i+1/2}p_{i+1/2}\mathrm{Sc}, \tag{21}$$

$$\Pi_{\mathrm{NS}i+1/2} = \frac{4}{3}\mu_{i+1/2}\frac{u_{i+1} - u_i}{h}, \tag{22}$$

$$w^*_{i+1/2} = \tau_{i+1/2}\left(\rho_{i+1/2}u_{i+1/2}\frac{u_{i+1} - u_i}{h} + \frac{p_{i+1} - p_i}{h}\right), \tag{23}$$

$$R_{i+1/2} = \tau_{i+1/2}\left(u_{i+1/2}\frac{p_{i+1} - p_i}{h} + \gamma p_{i+1/2}\frac{u_{i+1} - u_i}{h}\right), \tag{24}$$

$$\Pi_{i+1/2} = (\Pi_{\mathrm{NS}} + uw^* + R)_{i+1/2}. \tag{25}$$

Вычисление $q_{i+1/2}$ происходит согласно соотношениям

$$R^q_{i+1/2} = (\tau\rho)_{i+1/2}\left(\frac{u_{i+1/2}}{\gamma-1}\frac{(p/\rho)_{i+1}-(p/\rho)_i}{h} + (pu)_{i+1/2}\frac{1/\rho_{i+1}-1/\rho_i}{h}\right), \tag{26}$$

$$q_{i+1/2} = -(\tau p)_{i+1/2}\frac{\gamma \mathrm{Sc}}{\mathrm{Pr}(\gamma-1)}\frac{(p/\rho)_{i+1}-(p/\rho)_i}{h} - u_{i+1/2}R^q_{i+1/2}. \tag{27}$$

Вопросы точности и устойчивости КГД-алгоритма анализируются далее на примере расчетов задачи Римана.

3. ОЦЕНКИ ТОЧНОСТИ ЧИСЛЕННОГО АЛГОРИТМА

Разностная схема (16)–(18) формально имеет порядок точности по пространству $O(\alpha h)$. Расчеты из [1] подтверждают, что уменьшение коэффициента α эквивалентно сгущению пространственной сетки.

Оценим точность вычислительной схемы на основании расчетов задачи о распаде разрыва. В качестве начальных условий зададим значения на левом (l) и правом (r) интервалах области расчета $(\rho_l, u_l, p_l) = (8, 0, 480)$, $(\rho_r, u_r, p_r) = (1, 0, 1)$ при $\gamma = 5/3$. Задача решается в области (0, 200), разрыв располагается в точке $x_0 = 100$. Диссипативные коэффициенты вычислялись для $\mathrm{Pr} = 2/3$, $\mathrm{Sc} = 1$, $\alpha = 0.4$. Шаг по времени определяется величиной $\beta = 0.5$.

На фиг. 1 и 2 приведено распределение плотности ρ и фрагмент этого распределения для момента времени $t_{\mathrm{fin}} = 4$. Расчет выполнен на последовательности сеток с шагами $h = 0.5, 0.2, 0.1$ и 0.02. Видна монотонная сходимость численного решения при сгущении пространственной сетки.

Оценим порядок точности разностной схемы (16)–(18), используя расчеты на последовательности сгущающихся сеток. Пусть u – точное решение задачи. В нашем случае – это автомодельное решение (см. [8]). Пусть u_1 – сеточное решение задачи на сетке с шагом h, u_2 – сеточное решение задачи на сетке с шагом $2h$. Тогда, согласно [9], [10], в случае если решение задачи достаточно гладкое, т.е. имеет производные вплоть до второго порядка, погрешность сеточного решения имеет вид

$$u_1 - u = A(2h)^n, \quad u_2 - u = Ah^n, \tag{28}$$

где n – порядок точности разностной схемы, A – некоторая константа, зависящая от производных решения. При этом порядок точности схемы определяется по формуле

$$n = \log_2\frac{u_1-u}{u_2-u}. \tag{29}$$

На фиг. 3 приведено автомодельное решение задачи в случае, когда начальное положение разрыва находится в точке $x_0 = 0$ (линия *1*), и порядок точности численного алгоритма n определяется согласно (29). Порядок точности вычислен для двух последовательностей сеток: $h = 0.05$ и $2h = 0.1$ – величина n_1 (кривая *2*), и $h = 0.025$ и $2h = 0.05$ – величина n_2 (кривая *3*).

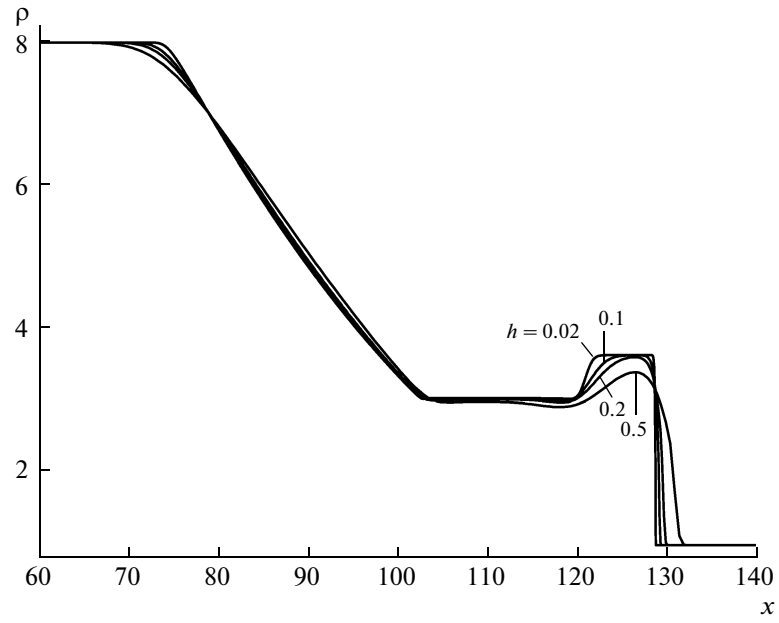
Оба расчета показывают, что в областях гладкого решения, т.е. в зоне волны разрежения $-35 < x < 5$, и в области за контактным разрывом $5 < x < 30$, эффективный порядок точности КГД-алгоритма составляет $0.5 \le n \le 2$. В тех областях, где течение близко к стационарному невязкому течению, решение задачи приближенно описывается стационарными уравнениями Эйлера, и эйлеровы комплексы, формирующие КГД-добавки, оказываются малыми. В этих зонах точность схемы возрастает.

В зонах разрывов решения и в точках, где разностное решение совпадает с точным ($u_1 = u$ или $u_2 = u$), использованный алгоритм оценки точности схемы некорректен. Последнее выражается в “пиках” и отрицательных значениях величины n.

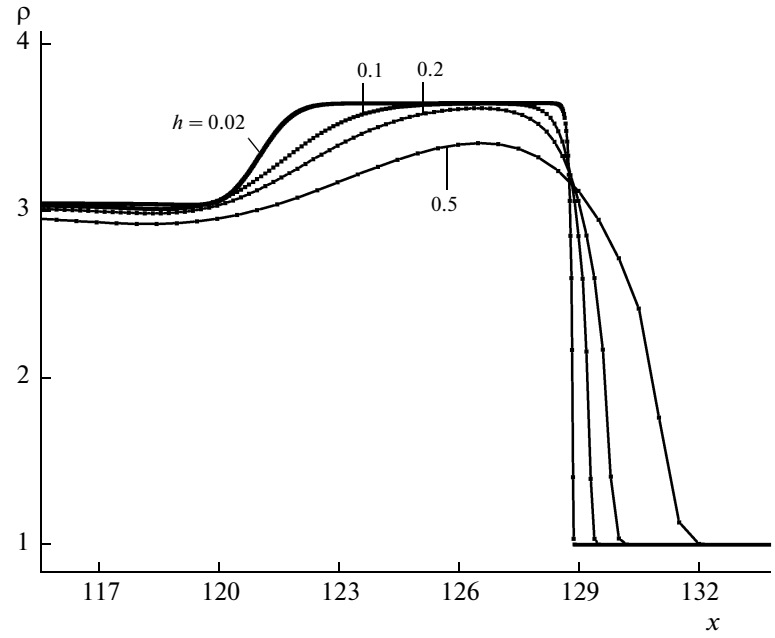
На фиг. 4 приведено распределение КГД-добавки $M = \rho w$ для уравнения неразрывности, которая в разностном виде, согласно (16), записывается как

$$M_i = \frac{1}{h}(\rho_{i+1/2}w_{i+1/2} - \partial(\rho w)/\partial x).$$

Здесь дискретные значения w вычисляются согласно (19). Расчет выполнен на сетках $h = 0.05$ (сплошная кривая) и $h = 0.025$ (штриховая линия). Из графика видно, что в областях гладкого решения величина добавки M_i мала, а в зоне между волной разрежения и контактным разрывом близка к нулю. Это приводит к повышению реального порядка точности разностной схемы в со-



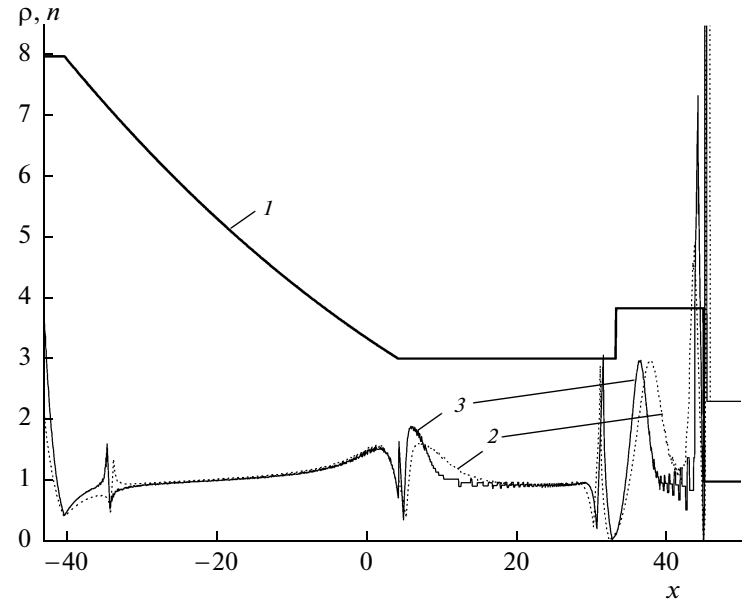
Фиг. 1.



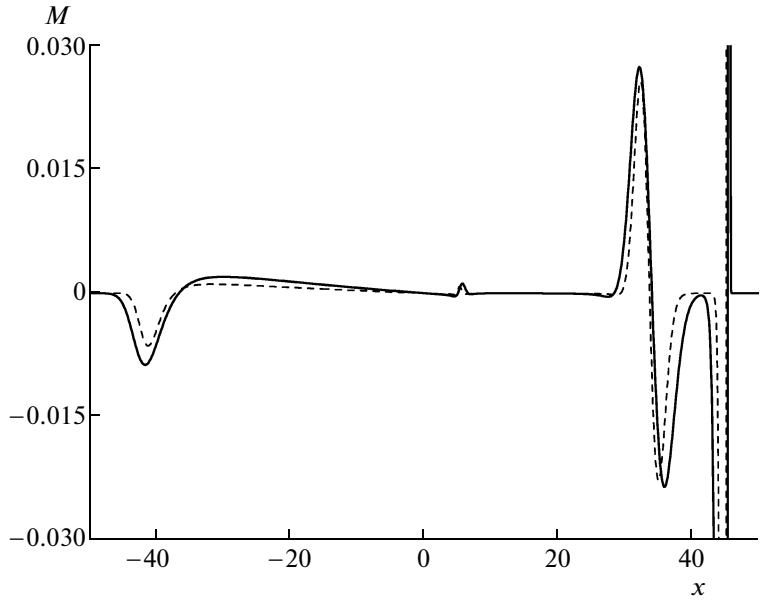
Фиг. 2.

ответствующих областях и коррелирует с распределениями величины n, приведенными на фиг. 3. Максимальные значения M_i достигаются в зоне ударной волны ($\max M_i \sim 40$, $\min M_i \sim 20$ на графике не изображены). Следующие по величине максимумы соответствуют области контактного разрыва, и два дополнительных малых по величине экстремума величины M_i возникают на звуковых точках, ограничивающих волну разрежения.

Тем самым величина КГД-диссипации, определяющая устойчивость разностного алгоритма, автоматически адаптируется к решению задачи и зависит от его локальных свойств.



Фиг. 3.



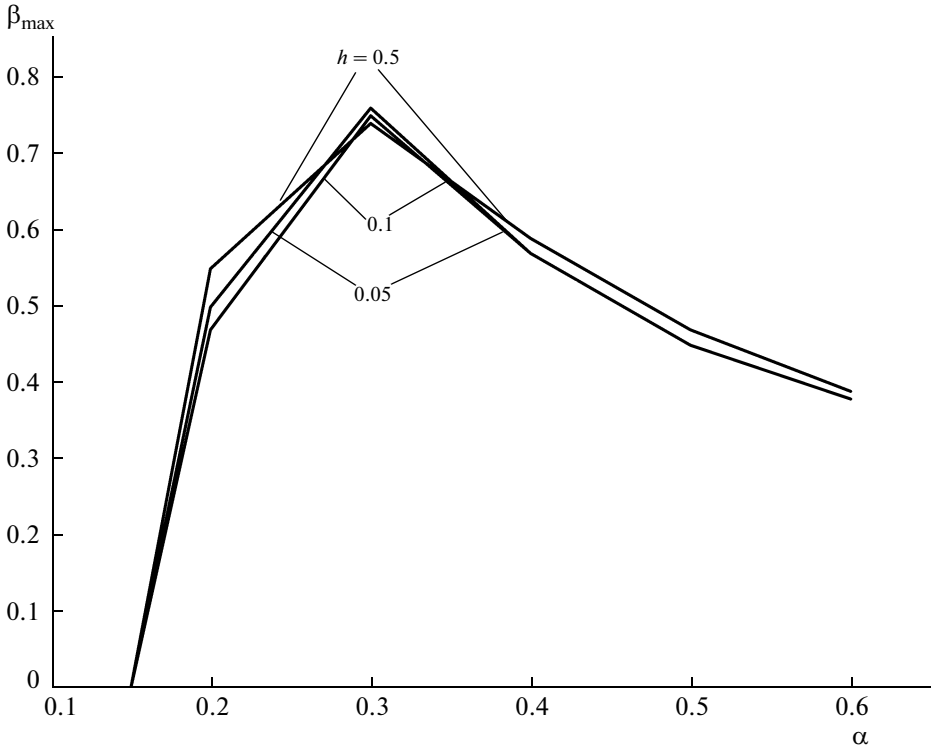
Фиг. 4.

3. ИССЛЕДОВАНИЕ УСТОЙЧИВОСТИ ЧИСЛЕННОГО АЛГОРИТМА

Численный алгоритм (16)–(18) представляет собой условно устойчивую явную по времени разностную схему. Как показывает практика численных расчетов и физические соображения, положенные в основу вывода КГД-уравнений на основании кинетических моделей, ограничение на временно́й шаг для этих алгоритмов определяется условием Куранта

$$\Delta t = \beta \min\left(\frac{h}{|u| + c}\right), \tag{30}$$

где $0 < \beta(\alpha) < 1$ – числовой коэффициент.



Фиг. 5.

В [7] методом энергетических неравенств было получено достаточное условие устойчивости для КГД-алгоритма и доказаны соответствующие теоремы. Рассматривалось одномерное течение в рамках уравнений Эйлера в акустическом приближении для одномерной по пространству разностной схемы с постоянным шагом. Полученное условие устойчивости Куранта имеет вид

$$\Delta t \leqslant \beta \frac{h}{c_*}, \tag{31}$$

где $c_* = \sqrt{\gamma \mathcal{R} T_*}$ – средняя по пространству скорость звука в начальный момент времени и коэффициент β определяется по формуле

$$\beta = \min(\beta_A, \beta_B, \beta_C), \quad \beta_A = \frac{A}{A^*}, \quad \beta_B = \frac{B}{B^*}, \quad \beta_C = \frac{C}{C^*}. \tag{32}$$

Значения A, B, C и A^*, B^*, C^* определяются величинами γ, Pr и значением коэффициента α, входящим в формулу для вычисления искусственной диссипации (15). Для Sc = 1 эти величины имеют вид

$$A = \frac{\alpha}{\gamma}, \quad B = \frac{\alpha}{\gamma}\left(\frac{4}{3} + \gamma\right), \quad C = \frac{\alpha}{\mathrm{Pr}(\gamma - 1)}, \tag{33}$$

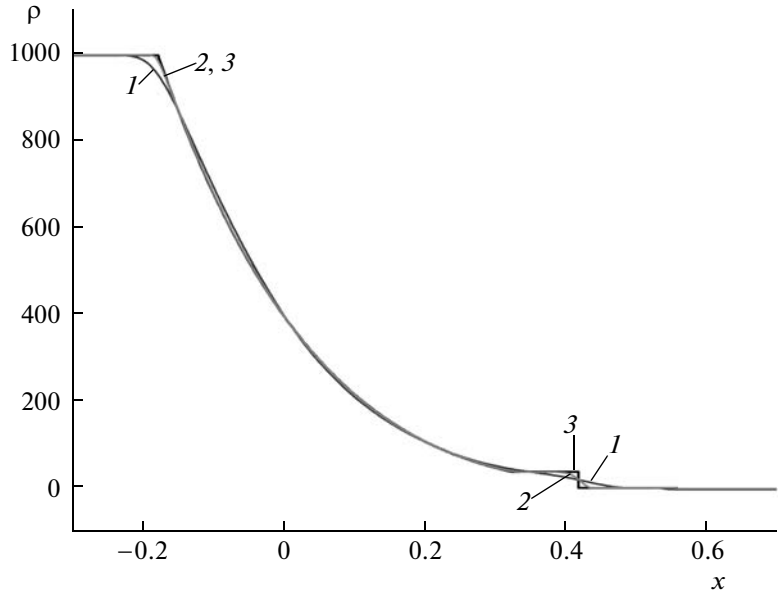
$$A^* = 2\gamma A^2 + 2(\gamma - 1)AC + \gamma A + B + \frac{1}{2\gamma}, \tag{34}$$

$$B^* = 2B^2 + A + \frac{B}{\gamma} + \frac{(\gamma - 1)C}{\gamma} + \frac{1}{2}, \tag{35}$$

$$C^* = (\gamma - 1)C(2A + 2C + 1). \tag{36}$$

Наиболее жестким является ограничение, определяемое величиной β_A. Для $\gamma = 5/3$, Pr = 2/3 при $\alpha = 0.5$ разностный алгоритм устойчив при $\beta \sim 0.12$ (см. [7]).

Далее на примере задачи, описанной выше, численно исследуется справедливость условия Куранта вида (30).



Фиг. 6.

Для каждого значения параметра α из промежутка 0.2–0.5 в формуле (15) варьировалась величина шага по времени – коэффициент β в формуле (30). При этом определялся максимально допустимый шаг по времени, соответствующий отсутствию осцилляций за фронтом ударной волны и энтропийного следа (провала), образующегося в профиле плотности за волной разрежения. Полученная таким образом зависимость максимально допустимого шага по времени приведена на фиг. 5 для шагов сетки $h = 0.5$, 0.1 и 0.05.

Расчеты показывают, что условие устойчивости схемы (16)–(18) имеет вид соотношения Куранта, в котором коэффициент β практически не зависит от шага пространственной сетки. Из графика наглядно видна зависимость шага по времени от параметра регуляризации. Для рассмотренного примера зависимость $\beta(\alpha)$ имеет максимум при $\alpha_{\max} = 0.3$, что соответствует $\beta = 0.7$. Использование в численных расчетах значений $\alpha > \alpha_{\max}$ представляется нецелесообразным. Для малых значений коэффициента $\alpha < 0.15$ численный алгоритм неустойчив.

Для других вычислительных задач зависимость $\beta_{\max}(\tau)$ может отличаться, но, как показывает практика численных расчетов, качественно отмеченные закономерности сохраняются.

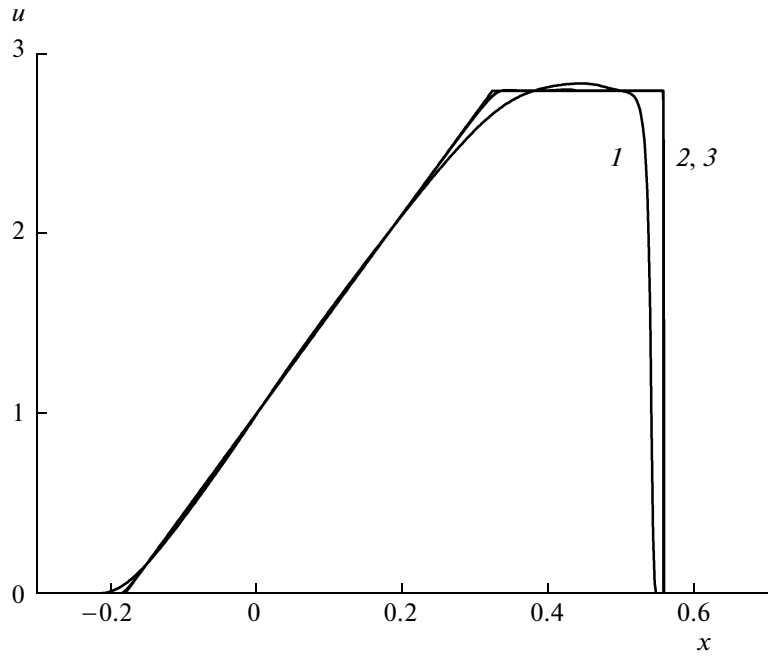
В заключение заметим, что полученное для линеаризованной задачи условие устойчивости (32) оказывается более жестким, чем условие устойчивости, реализуемое на практике.

5. ЗАДАЧА РИМАНА О СВЕРХСИЛЬНОМ РАЗРЫВЕ

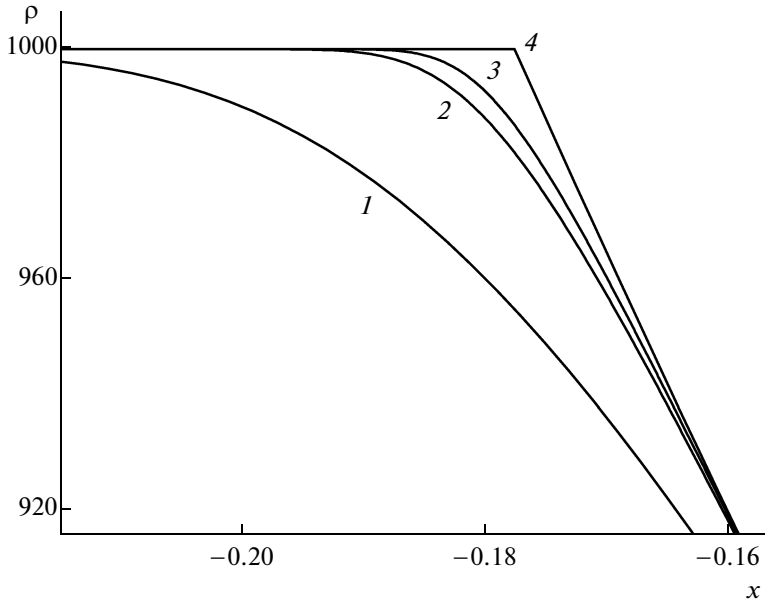
Для дальнейшей демонстрации возможностей КГД-алгоритма приведем решение задачи Римана о распаде разрыва с высокими перепадами плотности и давления. В качестве начальных условий зададим $(\rho_l, u_l, p_l) = (1000, 0, 1000)$, при $-0.3 < x < 0$, $(\rho_r, u_r, p_r) = (1, 0, 1)$ при $0 < x < 0.7$, $\gamma = 1.4$. Расчет ведется до времени $t_{\text{fin}} = 0.15$. На примере решения указанной задачи в [11] проведено детальное сопоставление возможностей восьми наиболее распространенных разностных схем высоких порядков точности. В частности, рассматривались схемы типа Годунова и различные варианты схем высокого порядка точности с расщеплением и коррекцией потока.

Полученные авторами оптимальные параметры расчета этой задачи для КГД-алгоритма составляют $\alpha = 0.3$, $\beta = 0.05$. Расчет выполнен при $\text{Pr} = 1$, $\text{Sc} = 1$.

Сходимость распределений плотности ρ и скорости u по сетке для всей области расчета показана на фиг. 6 и 7; здесь линия *1* – для $h = 0.002$, *2* – для $h = 0.0001$, *3* – точное решение. В силу больших перепадов газодинамических параметров, области решения вблизи звуковой точки, контактного разрыва и ударной волны приведены отдельно. Фрагменты распределений плотности и скорости в левой звуковой точке показаны на фиг. 8 и 9, те же величины в середине расчетной области изображены на фиг. 10, 11 и в зоне ударной волны фиг. 12 и 13. На фиг. 8–13 ли-



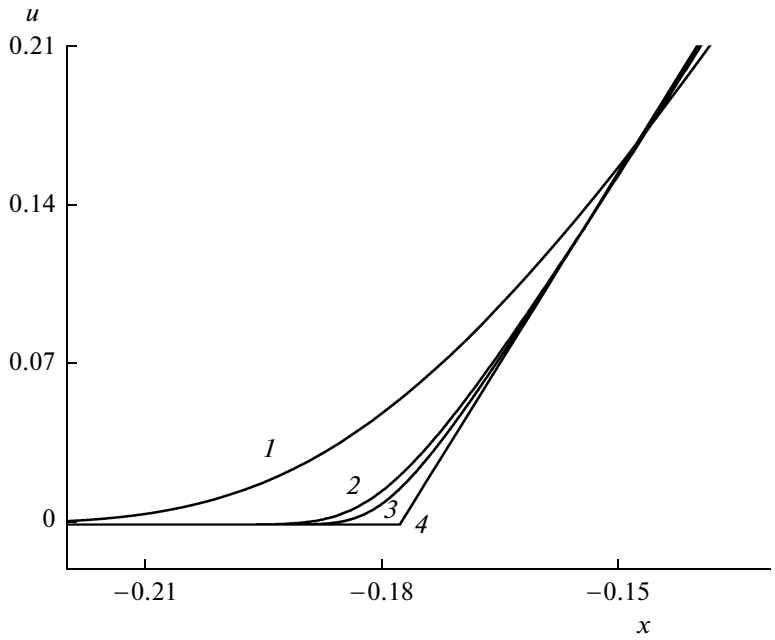
Фиг. 7.



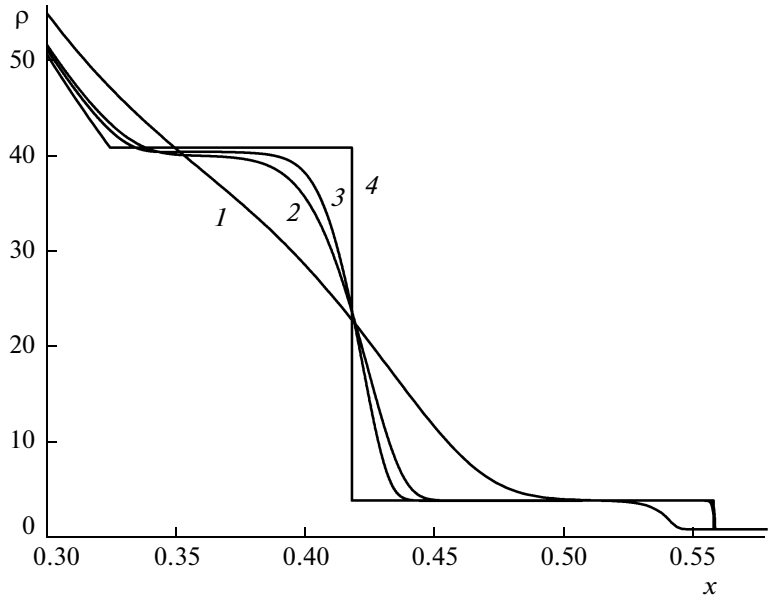
Фиг. 8.

ния *1* – для $h = 0.002$, *2* – для $h = 0.0002$, *3* – для $h = 0.0001$, *4* – точное решение. Из приведенных данных наглядно видна достаточно быстрая сходимость разностного решения к автомодельному при сгущении пространственной сетки.

Показательной характеристикой решения этой задачи является распределение скорости в ударной волне. Из сопоставления фиг. 13 с результатами из [11], полученными на такой же пространственной сетке $h = 0.002$, следует, что на этой сетке КГД-алгоритм близок по точности к



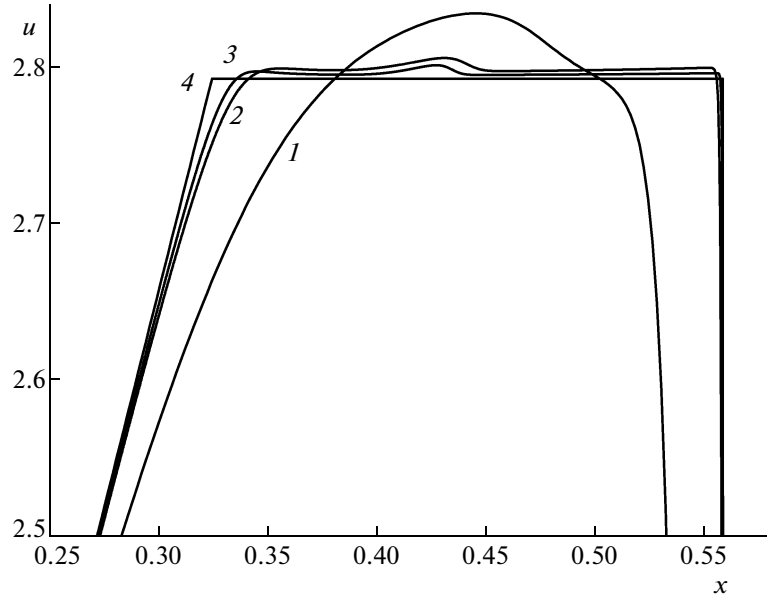
Фиг. 9.



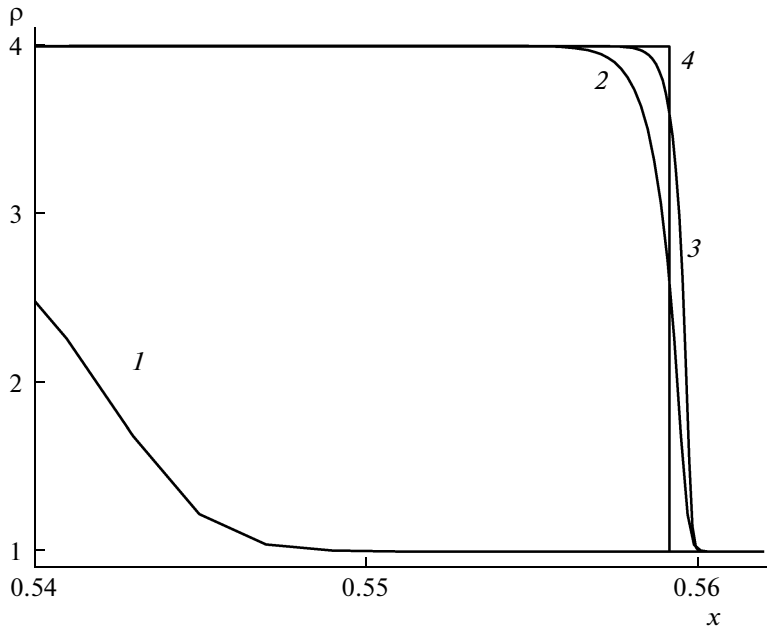
Фиг. 10.

рассмотренным в [11] разностным схемам. При этом численное решение, полученное по КГД-алгоритму, располагается слева от автомодельного решения, в то время как все алгоритмы, изученные в [11], формируют профиль скорости, расположенный справа от автомодельного распределения. С уменьшением шага пространственной сетки точность КГД-алгоритма резко увеличивается и превосходит точность исследуемых в [11] методов.

Полученные данные, совместно с результатами из [1], позволяют сделать вывод о преимуществах КГД-алгоритмов при численном моделировании течений на подробных сетках, что является особенно актуальным в связи с широким использованием в настоящее время мощных вычислительных комплексов.



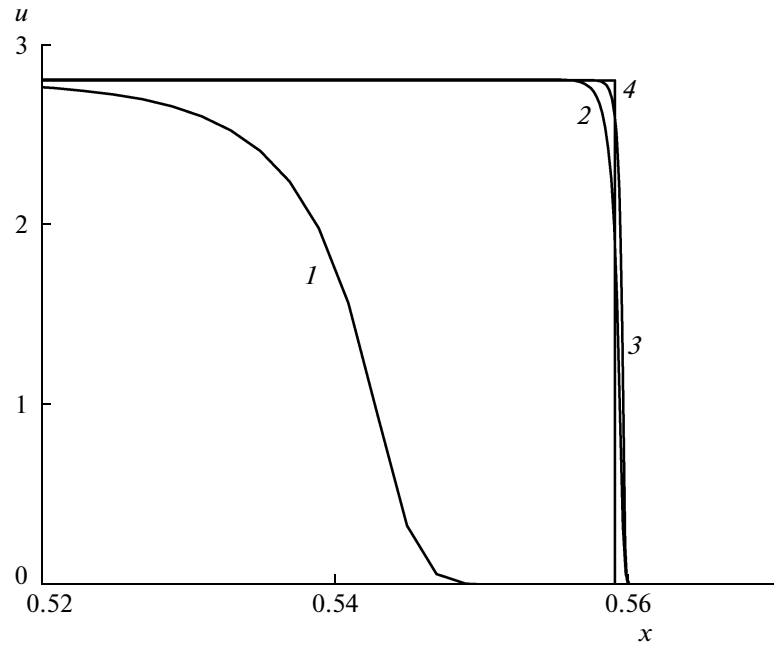
Фиг. 11.



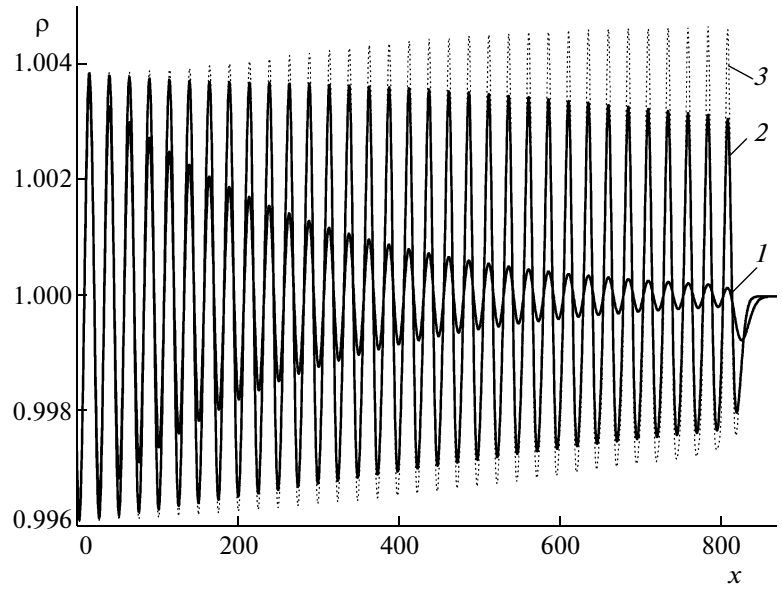
Фиг. 12.

6. ЗАДАЧА О РАСПРОСТРАНЕНИИ ЗВУКОВЫХ КОЛЕБАНИЙ

Теоретический анализ и опыт использования КГД-уравнений показывает, что используемые авторами численные алгоритмы, основанные на этой системе уравнений, эффективны при расчете нестационарных и пульсационных течений вязкого сжимаемого газа. Примерами расчета таких течений являются неустанавливающиеся течения, возникающие при сверхзвуковом обтекании торца с выступающей иглой (см. [12]), пульсационные течения в окрестности полого цилиндра (см. [13]), колебательные течения в кавернах (см. [14]), дозвуковые течения в следе за цилиндром, или дорожка Кармана (см. [15]), а также хаотические течения за обратным уступом, моделирующие турбулентные следы в донной области (см. [16]) и в следе за препятствием



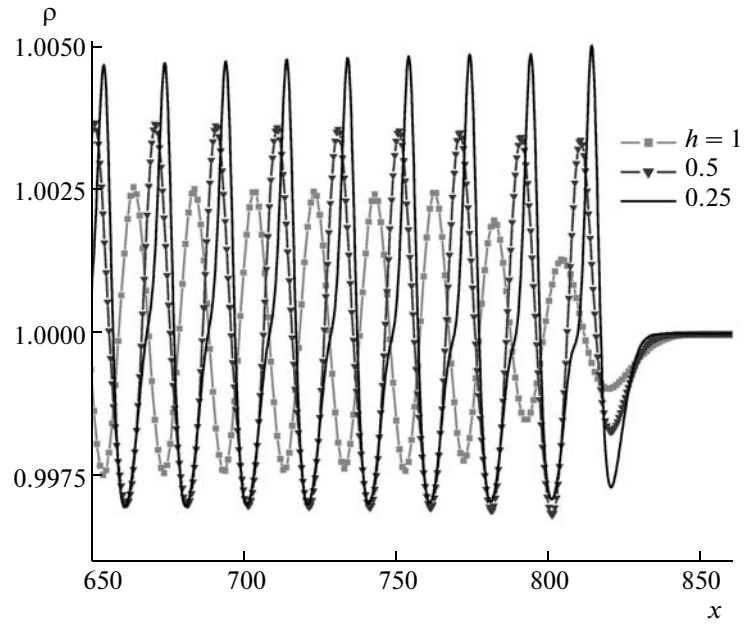
Фиг. 13.



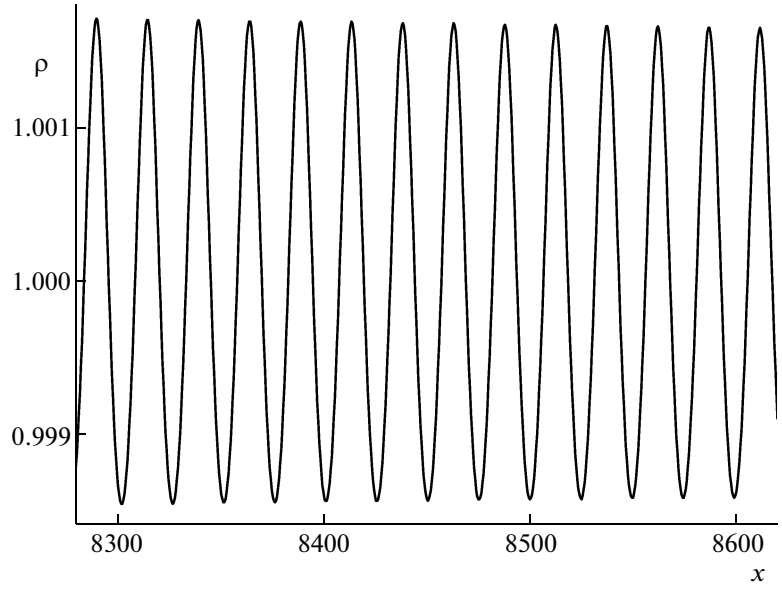
Фиг. 14.

(см. [17]). Такого рода течения сопровождаются генерацией звуковых колебаний, изучение которых является актуальной задачей аэроакустики. Традиционно в таких задачах применяются специальные численные методы высокого порядка точности, например, схемы четвертого порядка точности по времени и пространству (см. [18]), или схемы более высоких порядков, например, [19].

В силу сказанного выше представляет интерес изучение возможностей КГД-алгоритма для прямого и однородного численного моделирования как генерации акустических возмущений в пульсационных или турбулентных течениях, так и распространения этих возмущений вдали от зоны их зарождения. Возможность адекватного описания распространения звуковых волн с помощью КГД-модели в акустическом приближении показана в [20].



Фиг. 15.



Фиг. 16.

Далее приведены результаты расчета задачи о распространении слабых звуковых колебаний, выполненные в рамках КГД-алгоритма для уравнений Эйлера.

В качестве начальных условий используется невозмущенное поле течения $\rho_0 = 1$, $p_0 = 1$, $u_0 = 0$. Скорость распространения возмущения составляет $c_0 = \sqrt{\gamma}$.

Пусть граничное условие на левой границе области имеет вид гармонического возмущения:

$$u(t, 0) = -\frac{A_0}{\sqrt{\gamma}}\sin(2\pi c_0 t/\lambda),$$
$$\rho(t, 0) = 1 - A_0\sin(2\pi c_0 t/\lambda), \quad p(t, 0) = 1 - A_0\sin(2\pi c_0 t/\lambda). \tag{37}$$

На правой границе области зададим мягкие граничные условия, или условия сноса $\partial f/\partial x = 0$, где $f = (\rho, u, p)$. Длина волны звукового возмущения выбирается равной $\lambda = 20$, амплитуда A_0 варьируется от 0.1 до 0.005. Параметрами расчета являются значения $\alpha = 0.2$, $\beta = 0.4$, $\mathrm{Pr} = 1$. Для определенности выбрано $\gamma = 7/5$.

На фиг. 14 показаны результаты расчета этой задачи для малого возмущения, амплитуда которого составляет $A_0 = 0.005$, в зависимости от величины вязкости, т.е. для значений $\mathrm{Sc} = 1$ – линия *1*, для $\mathrm{Sc} = 0.1$ – линия *2* и для $\mathrm{Sc} = 0$ – линия *3*. Последнее соответствует невязкому течению. Расчеты выполнены на сетке с шагом $h = 0.5$. Увеличение коэффициентов вязкости и теплопроводности (увеличение числа Sc) приводит к росту затухания амплитуды колебаний. Однако частота колебаний при этом не искажается.

Влияние шага пространственной сетки на решение задачи показано на фиг. 15, где приведены фрагменты распределения плотности на момент времени $t = 700$ для $\mathrm{Sc} = 0$. Видно, что затухание амплитуды волны растет с ростом шага пространственной сетки. В расчетах на подробных сетках $h = 0.5$ и 0.25 наблюдается искажение формы исходного возмущения, что представляется естественным при описании распространения волны в рамках уравнений Эйлера (см. [21]). Из сравнения приведенных рисунков видно, что фазовая ошибка не зависит от величины вязкости, определяемой числом Sc, но зависит от величины шага сетки h.

Оценить затухание звуковых колебаний плотности на больших расстояниях от источника позволяет фиг. 16. Расчет выполнен при $\mathrm{Sc} = 0.01$ на сетке с шагом $h = 0.5$ до расстояния по x, которое составляет порядка 500 длин волны и соответствует времени $t = 8000$.

Результаты проведенных расчетов показывают, что КГД-алгоритмы, являющиеся схемами первого порядка точности по времени и пространству, позволяют моделировать распространение звуковых возмущений, в том числе и на больших расстояниях от источника. Последнее делает перспективным использование этого алгоритма в задачах аэроакустики.

7. ОЦЕНКА ВЫЧИСЛИТЕЛЬНОЙ СЛОЖНОСТИ АЛГОРИТМА

Время работы процессора является важной характеристикой при выборе метода численного решения конкретной задачи. Однако прямое сопоставление времени счета одной и той же задачи, реализованной с помощью разных численных алгоритмов, часто не проясняет ситуации, что обусловлено многообразием используемых процессорных элементов, операционных систем, оптимизирующих программ, а также уровнем оптимизации программируемых формул самим программистом. Поэтому оценка вычислительной сложности алгоритма и ее сопоставление с аналогичной характеристикой альтернативных численных подходов является важным аспектом при выборе того или иного метода вычисления.

Объективной характеристикой вычислительной сложности метода является число арифметических операций и время на передачу данных из памяти для их выполнения. Именно по таким данным проводятся оценки абсолютной производительности современных вычислительных комплексов, включающих как однопроцессорные одноядерные вычислительные модули, так и многопроцессорные многоядерные вычислители. Соответствующие оценки для сравнения разных алгоритмов можно выполнить и не прибегая к численной реализации метода.

Вычислительный элемент можно схематически представить как совокупность оперативной памяти (ОП), быстрой памяти (КЭШ) и самого процессорного элемента (ПЭ). Характерные данные, определяющие скорость работы современных вычислительных элементов, приведены, например, в [22] и цитированных в этой работе источниках.

Согласно современным представлениям, за один временно́й такт работы ПЭ выполняется от одной до четырех арифметических операций сложения/вычитания, умножения или деления. Скорость передачи слова между ПЭ и КЕШ составляет порядка 15 тактов. Время подготовки передачи одного слова из ОП в КЭШ, или так называемая латентность, составляет ~1000 тактов. Скорость передачи данных между ОП и ПЭ в случае, если эти данные выбираются не случайным образом, а в виде заранее подготовленной последовательности или массива, близка к скорости передачи слова между ПЭ и КЕШ.

В алгоритмах расчета газодинамических течений на регулярных сетках данные расположены в памяти регулярно, передаются для обработки большими массивами, и время латентности памяти не ограничивает производительность процессора. При этом если время обработки одного слова в ПЭ превосходит время его выборки из КЭШ, т.е. превосходит 15 тактов, то число арифметических операций, приходящихся на одну расчетную точку, непосредственно определяет вре-

мя работы процессора. На основе оценки указанных величин удается сопоставить вычислительную сложность разных алгоритмов.

При оценке вычислительной сложности КГД-алгоритма принята следующая модель: учитываются арифметические операции сложение/вычитание, умножение, деление и извлечение квадратного корня, осуществляемые в формулах (16)–(27), т.е. операции, составляющие непосредственное содержание данного математического алгоритма. Эти операции учитываются напрямую, без оптимизации. Операции, связанные с индексацией переменных, и другие вспомогательные операции не учитываются.

Рассмотрим решение первого уравнения – вычисление значения плотности ρ_i на новом слое: вычисление величины $w_{i+1/2}$ (по формуле (19)), которая представляет собой добавку к скорости $u_{i+1/2}$, требует 3 операции сложения, 6 умножений и 1 деление; вычисление значения $\tau_{i+1/2}$ (по формуле (15)) – одно умножение, одно деление и одно извлечение корня. Вычисление $j_{mi+1/2}$ (по формуле (20)) требует одно сложение и одно умножение. Вычисление плотности на новом временно́м слое ρ_i (по формуле (16)) требует два сложения, одно умножение и одно деление. Тем самым весь алгоритм нахождения плотности в одной точке сетки на новом слое требует 10 операций сложения, 17 умножения, 5 делений и 2 вычисления квадратного корня, т.е. всего 34 такта.

Для вычисления величины $\Pi_{i+1/2}$ (по схеме (21)–25)) требуется 9 сложений, 13 умножений и 5 делений. Т.е. вычисление нового значения $\widehat{\rho_i u_i}$ (17) требует 24 сложения, 30 умножений и 10 делений – всего 64 операции, или такта работы ПЭ.

При решении третьего уравнения схемы (18) вычисление теплового потока $q_{i+1/2}$ (по формулам (26), (27)) требует 14 сложений, 18 умножений и 20 делений. Расчет полной энергии на новом слое в точке i требует 22 сложений, 23 умножений и 22 делений – всего 67 тактов.

Тем самым весь алгоритм для трех уравнений в точке (16)–(18) требует 60 сложений, 70 умножений, 41 деления и 2 вычислений квадратного корня и расчет одного шага по времени для одной пространственной точки занимает ~180 тактов работы ПЭ. Если полагать, что за один такт выполняются четыре арифметические операции, то эта цифра уменьшается в 4 раза. Однако при расчете последовательности точек объем вычислений сокращается в 2 раза, если учесть, что все потоки для точки $i + 1/2$ вычисляются два раза – это поток справа при расчете точки i и поток слева при расчете точки $i + 1$. Оптимизация вычислительных формул на уровне написания программы позволяет уменьшить полученную цифру. В основном вычислительном цикле алгоритма отсутствуют логические операции, операции с удаленными точками шаблона и операции выбора одиночных слов из ОП, которые могли бы сильно замедлить работу вычислительного элемента в целом.

В КГД-алгортме при вычислении значения переменной на новом слое в точке i все перечисленные операции совершаются на основе четырех переменных u, ρ, p и E, с использованием трехточечного шаблона – значений этих переменных в точках $i-1$, i и $i+1$. При вычислении следующей $i + 1$ точки достаточно извлечь из памяти только 4 новых значения в точке $i + 2$, так как значения в точках i и $i + 1$ уже известны. Время их извлечения из памяти составляет $4 \times 15 = 60$ тактов, что не превышает времени перехода на следующий временно́й слой, составляющее ~90 тактов. Две эти величины определяют вычислительную сложность КГД-алгоритма.

Таким образом, время подкачки одного слова из КЭШ-памяти близко к времени его обработки в ПЭ, что делает КГД-алгоритм эффективным с точки зрения скорости работы вычислительного элемента.

Близкие соотношения времени подкачки данных из памяти и их обработки в процессоре получаются для КГД-алгоритмов в 3Д-формулировке, а также для реализации этого метода на неструктурированных сетках.

ЗАКЛЮЧЕНИЕ

Результаты численного моделирования задачи о распаде разрыва показывают, что в области гладких решений реальный порядок точности КГД-алгоритма колеблется от 0.5 до 2 в зависимости от локальных свойств решения. Искусственная диссипация, присущая КГД-алгоритму решения уравнений Эйлера, автоматически адаптируется к решению и близка к нулю в тех областях, где искомое решение не имеет особенностей.

Расчеты показывают, что КГД-алгоритм представляет собой условно устойчивую разностную схему с условием устойчивости Куранта. Основные параметры настройки алгоритма – это численный коэффициент α, входящий в параметр регуляризации τ, и коэффициент β, определяю-

щий шаг по времени. При этом зависимость $\beta(\alpha)$ имеет экстремум, соответствующий оптимальному для проведения расчетов шагу по времени.

Представленные расчеты задачи о распаде сверхсильного разрыва демонстрируют монотонную сходимость численного решения к автомодельному при сгущении пространственной сетки. Показано, что точность КГД-алгоритма на подробных сетках превосходит точность решения этой задачи, достигнутую на таких же сетках с помощью методов коррекции или расщепления потоков порядка точности.

Моделирование задачи о распространении слабых возмущений показывает, что при использовании достаточно подробных сеток КГД-алгоритм позволяет рассчитывать сотни периодов гармонического колебания. Тем самым этот алгоритм представляется перспективным для использования в задачах акустики наравне со схемами высоких порядков точности.

Проведенные оценки вычислительной сложности метода показывают, что сбалансированное соотношение времени выполнения арифметических операций и времени извлечения данных из памяти в расчете на один узел сетки обеспечивают высокую вычислительную эффективность метода при его реализации на современных вычислительных системах.

Полученные в работе оценки точности, устойчивости и вычислительной сложности КГД-алгоритма, выполненные для одномерных задач, являются ориентирами при выполнении практических расчетов многомерных задач на сетках различной структуры.

СПИСОК ЛИТЕРАТУРЫ

1. *Елизарова Т.Г., Шильников Е.В.* Возможности квазигазодинамического алгоритма для численного моделирования течений невязкого газа // Ж. вычисл. матем. и матем. физ. 2009. Т. 49. № 3. С. 549–566.
2. *Шеретов Ю.В.* Математическое моделирование течений жидкости и газа на основе квазигидродинамических и квазигазодинамических уравнений. Тверь: Тверской гос. ун-т, 2000.
3. *Елизарова Т.Г.* Квазигазодинамические уравнения и методы расчета вязких течений. М.: Научн. мир, 2007.
4. *Четверушкин Б.Н.* Кинетические схемы и квазигазодинамическая система уравнений. М.: Макс Пресс, 2004.
5. *Елизарова Т.Г., Хохлова А.А.* Квазигазодинамические уравнения для течений газа с внешними источниками тепла // Вестн. МГУ. Серия 3. Физика, астрономия. 2007. № 3. С. 10–13.
6. *Шеретов Ю.В.* О разностных аппроксимациях квазигазодинамических уравнений для осесимметричных течений // Применение функционального анализа в теории приближений. Тверь: Тверской гос. ун-т. 2001. С. 191–207.
7. *Шеретов Ю.В.* Анализ устойчивости модифицированной кинетически-согласованной разностной схемы в акустическом приближении // Применение функционального анализа в теории приближений. Тверь: Тверской гос. ун-т. 2004. С. 147–160.
8. *Рождественский Б.Л., Яненко Н.Н.* Системы квазилинейных уравнений. М.: Наука, 1978.
9. *Калиткин Н.Н.* Численные методы. М.: Наука, 1978.
10. *Самарский А.А.* Теория разностных схем. М.: Наука, 1989.
11. *Tang H., Liu T.* A note on the conservative schemes for Euler equations // J. Comput. Phys. 2006. № 218. P. 451–459.
12. *Антонов А.Н., Елизарова Т.Г., Павлов А.Н., Четверушкин Б.Н.* Математическое моделирование колебательных режимов при обтекании тела с иглой // Матем. моделирование. 1989. Т. 1. № 1. С. 14–23.
13. *Антонов А.Н., Елизарова Т.Г., Четверушкин Б.Н., Шеретов Ю.В.* Численное моделирование пульсационных режимов при сверхзвуковом обтекании полого цилиндра // Ж. вычисл. матем. и матем. физ. 1990. Т. 30. № 4. С. 548–556.
14. *Антонов М.А., Граур И.А., Косарев Л.В., Четверушкин Б.Н.* Численное моделирование пульсаций давления в трехмерных выемках // Матем. моделирование. 1996. Т. 8. № 5. С. 76–90.
15. *Elizarova T.G., Khokhlov A.A., Sheretov Yu.V.* Quasi-gasdynamic numerical algorithm for gas flow simulations // Internat. J. for Numer. Meth. in Fluids. 2008. V. 56. № 8. P. 1209–1215.
16. *Елизарова Т.Г., Никольский П.Н.* Численное моделирование ламинарно-турбулентного перехода в течении за обратным уступом // Вестн. МГУ. Серия 3. Физика. Астрономия. 2007. № 4. С. 14–17.
17. *Четверушкин Б.Н., Шильников Е.В.* Вычислительный и программный инструментарий для моделирования трехмерных течений вязкого газа на многопроцессорных системах // Ж. вычисл. матем. и матем. физ. 2008. Т. 48. № 2. С. 118–129.
18. *Tam C.K.W., Webb J.C.* Dispersion-relation-preserving finite difference schemes for computational acoustics // J. Comput. Phys. 1993. V. 107. P. 262–281.

19. *Абалакин И.В., Козубская Т.К.* Многопараметрическое семейство схем повышенной точности для линейного уравнения переноса // Матем. моделирование. 2007. Т. 19. № 7. С. 56–66.
20. *Шеретов Ю.В.* Анализ задачи о распространении звука для линеаризованных КГД-систем // Применение функционального анализа в теории приближений. Тверь: Тверской гос. ун-т. 2001. С. 178–191.
21. *Ландау Л.Д., Лифшиц Е.М.* Гидродинамика. М.: Наука, 1986.
22. *Williams S., Shalf J., Oliker L. et al.* Scientific computing kernels on the Cell processor // Internat. J. Parallel Program. 2007. V. 35. № 3. P. 263–298.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2009, том 49, № 11, с. 1970–1987

УДК 519.634

СИММЕТРИЧНЫЕ РАЗНОСТНЫЕ СХЕМЫ ПОКОМПОНЕНТНОГО РАСЩЕПЛЕНИЯ И ЭКВИВАЛЕНТНЫЕ ИМ СХЕМЫ ПРЕДИКТОР-КОРРЕКТОР ДЛЯ РЕШЕНИЯ МНОГОМЕРНЫХ ЗАДАЧ ГАЗОВОЙ ДИНАМИКИ МЕТОДОМ ГОДУНОВА

© 2009 г. О. А. Макотра, Н. Я. Моисеев, И. Ю. Силантьева, Т. В. Топчий, Н. Л. Фролова

(456770 Снежинск, Челябинская обл., а.я. 245,
ФГУП РФЯЦ-ВНИИТФ им. акад. Е.И. Забабахина)
e-mail: nyamoiseyev@vniitf.ru

Поступила в редакцию 14.11.2008 г.
Переработанный вариант 23.03.2009 г.

Рассмотрен подход к повышению точности численных решений многомерных задач газовой динамики в схемах Годунова. Основная идея подхода заключается в построении симметричных разностных схем расщепления по пространственным переменным с последующим преобразованием их к эквивалентным схемам предиктор-корректор. Показано, что одним из источников ошибок аппроксимации в схемах Годунова является вычисление "больших" величин из решения одномерной задачи о плоском распаде произвольного разрыва на границе двух соседних ячеек. Предложена реконструкция "больших" величин, которая позволила устранить отмеченный источник ошибок аппроксимации. Шаг интегрирования по времени в модифицированных схемах согласован с выбором шага в одномерных схемах и на равномерных по пространству разностных сетках, в 2 и 3 раза больших, чем в классических схемах Годунова для решения двумерных и трехмерных задач соответственно. Результаты расчетов тестовых задач подтвердили выводы о повышении точности решений в модифицированных схемах. Библ. 24. Фиг. 2.

Ключевые слова: метод расщепления по пространственным переменным, схемы Годунова.

1. ВВЕДЕНИЕ

Одним из распространенных методов численного решения уравнений газовой динамики является метод Годунова (см. [1]), свойства которого хорошо известны из [2], [3]. Разностные схемы, построенные на основе этого метода, являются консервативными, имеют первый порядок аппроксимации по времени и по пространству и для расчетов сильных разрывов типа ударных волн (УВ) не требуют введения дополнительной искусственной вязкости. Численные решения, как правило, монотонные. Разностные схемы являются схемами типа предиктор-корректор (см. [2]–[5]) и конструируются единообразно для решения как одномерных, так и многомерных задач. На шаге предиктора вычисляются "большие" величины из решения одномерной задачи о плоском распаде произвольного разрыва, возникающего на границе двух соседних ячеек. На шаге корректора вычисляются основные величины из интегральных законов сохранения массы, импульса и полной энергии. Однако такая универсальность имеет свои недостатки. Так, если решать одномерную задачу по двумерным или трехмерным методикам с применением равномерных по пространству разностных сеток (квадратных и кубических), то шаг интегрирования по времени по сравнению с шагом в одномерной методике будет в два и три раза меньше соответственно. Следовательно, точность решений будет ниже, чем точность решений по одномерной методике. Поэтому вопрос повышения точности решений многомерных задач методом Годунова остается актуальным.

В работах [6]–[9] для повышения точности решений предлагается решать задачу о распаде произвольного разрыва со многими начальными состояниями. Так, если рассматривать решение двумерных задач на четырехугольных сетках, то дополнительно решаются задачи о распаде разрыва в узлах разностной сетки. Эти задачи будут задачами с четырьмя начальными состояниями. Обобщение подхода к решению пространственных задач неочевидно и потребует еще больших затрат по времени счета, чем в решениях двумерных задач.

С другой стороны, вопрос учета влияния возмущений, приходящих из узлов разностной сетки, исследовался в 1956 г. К.В. Брушлинским, который построил схему для уравнений акустики с точным решением в углах по функционально-инвариантным решениям С.Л. Соболева. Сравнение решений по этой схеме и по "грубой" схеме с решением одномерной задачи о плоском распаде произвольного разрыва показало, что решения по схемам практически совпадают (см. стр. 56 в [2]). Поэтому вопрос применения для более широкого класса задач в численных методах решений задачи о распаде разрыва со многими начальными состояниями, по всей видимости, требует дополнительных исследований.

В работе [10] рассматривается подход к повышению точности решений двумерных задач. Согласно этому подходу при вычислении потоков через грани ячеек из решения задачи о распаде разрыва, предварительно находятся поправки к начальным данным с привлечением матриц собственных векторов и матриц собственных значений, соответствующих матрицам в записи гиперболических систем уравнений в недивергентной форме. Выбор значений поправок зависит от параметров течения.

В работе [11] при анализе подходов к повышению точности метода Годунова для решения трехмерных задач сделан вывод, что, по всей видимости, схемы расщепления (см. [12]–[18]) являются наиболее эффективными как по производительности, так и по реализации.

В работе [12] на примерах решений линейных уравнений переноса с постоянными коэффициентами показано, что применение метода покомпонентного расщепления к явным схемам типа Годунова приводит к повышению точности численных решений. Однако при переходе к таким схемам расщепления возникают новые проблемы, связанные с некоммутативностью операторов расщепления, с реализацией граничных условий (см. [18]), а также с организацией счета (см. [19]).

В настоящей работе предложен подход к повышению точности численных решений многомерных задач газовой динамики в явных схемах типа Годунова на основе построения симметричных разностных схем покомпонентного расщепления (см. [15], [20], [21]) с последующим преобразованием этих схем к эквивалентным схемам предиктор-корректор. Вследствие такого перехода выписываются выражения для реконструкции больших величин в исходной схеме Годунова. Реконструкция проводится перед шагом корректора после вычисления больших величин методом Годунова. В модифицированных явных схемах выбор шагов интегрирования по времени согласован с выбором в одномерных схемах и при числах Куранта, равных единице, условие сдвига выполняется. Проведен анализ аппроксимации разностных схем Годунова без расщепления и с покомпонентным расщеплением для решения уравнений переноса и акустики с постоянными коэффициентами. Показано, что одним из источников ошибок аппроксимации в схемах Годунова являются большие величины давлений, которые вычислены из решения одномерной задачи о плоском распаде произвольного разрыва, возникающего на границе двух соседних ячеек. Исследования свойств разностных схем проводились методом дифференциальных приближений (ДП) (см. [22]) с применением системы MAPLE (см. [23]).

2. ОСНОВЫ ПОДХОДА ДЛЯ РЕШЕНИЯ ЛИНЕЙНЫХ УРАВНЕНИЙ ПЕРЕНОСА

2.1. Уравнения переноса с двумя пространственными переменными

Рассмотрим задачу Коши для решения линейного уравнения переноса с двумя пространственными переменными

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} + b\frac{\partial u}{\partial y} = 0, \quad a > 0, b > 0 - \text{const}, \tag{2.1}$$

с начальными данными $u(0, x, y) = u(x, y)$. Построим в пространстве (t, x, y) разностную сетку с шагами τ, h_x, h_y по времени и по пространству соответственно. Для простоты рассуждений рассматриваем прямоугольную и равномерную по пространственным переменным разностную сетку. Обозначим через $u^{j,k}$, $u_{j,k}$ значения функции в центрах ячеек в моменты времени $t_n + \tau$, t_n соответственно. Символами $U_{j+1/2,k}$, $U_{j-1/2,k}$ и $U_{j,k+1/2}$, $U_{j,k-1/2}$ обозначим большие величины, которые относятся к "вертикальным" и к "горизонтальным" граням ячеек соответственно. Разностную схему предиктор-корректор для решения уравнения (2.1) запишем в виде

$$\begin{gathered} u^{j,k} = u_{j,k} - c_x(U_{j+1/2,k} - U_{j-1/2,k}) - c_y(U_{j,k+1/2} - U_{j,k-1/2}), \\ c_x = a\tau/h_x, \quad c_y = b\tau/h_y. \end{gathered} \tag{2.2}$$

Здесь c_x, c_y – числа Куранта. В методе Годунова большие величины $U_{j+1/2,k}$, $U_{j-1/2,k}$, $U_{j,k+1/2}$, $U_{j,k-1/2}$ вычисляются из уравнений

$$U_{j+1/2,k} = u_{j,k}, \quad U_{j-1/2,k} = u_{j-1,k}, \quad U_{j,k+1/2} = u_{j,k}, \quad U_{j,k-1/2} = u_{j,k-1} \tag{2.3}$$

и относятся к моментам времени $t_n + \tau_{j+1/2,k}$, $t_n + \tau_{j,k+1/2}$, где $\tau_{j+1/2,k} = 0.5h_x/a$, $\tau_{j,k+1/2} = 0.5h_y/b$ (см. [12], [24]). Запишем схему (2.2) в операторной форме:

$$u^{j,k} = [E - \tau\Lambda_1 - \tau\Lambda_2]u_{j,k}.$$

Здесь $\Lambda_1 = a(U_{j+1/2,k} - U_{j-1/2,k})/h_x$, $\Lambda_2 = b(U_{j,k+1/2} - U_{j,k-1/2})/h_y$ – это разностные операторы, которые аппроксимируют дифференциальные операторы $a\dfrac{\partial}{\partial x}$, $b\dfrac{\partial}{\partial y}$ соответственно. В [12] показано, что схема (2.2), (2.3) при числах Куранта, равных единице, не удовлетворяет условию сдвига: $u^{j,k} = u_{j-1,k-1}$. Если решение находить методом покомпонентного расщепления (см. [15])

$$u^{j,k} = (E - \tau\Lambda_2)(E - \tau\Lambda_1)u_{j,k} = [E - \tau(\Lambda_1 + \Lambda_2) + \tau^2\Lambda_2\Lambda_1]u_{j,k}, \tag{2.4}$$

то схема (2.4) будет удовлетворять условию сдвига. Устойчивость схемы докажем так же, как доказывается устойчивость схем в [2]. Пусть в пространстве задана некоторая норма сеточной функции $F = \{u_{j,k}\}$. Найдем ограничения на выбор шага τ, при которых разностная схема не увеличивает эту норму при переходе на один шаг по времени. Разностной схеме (2.4) соответствует матрица H перехода с временного слоя t_n на слой $t_n + \tau$. Эту матрицу можно представить в виде

$$H = I + \tau H_1 + \tau H_2 + \tau^2 H_1 H_2 = (1 - \tau/\tau_x)(1 - \tau/\tau_y)I +$$
$$+ (\tau/\tau_x)(1 - \tau/\tau_y)(I + \tau_x H_1) + (\tau/\tau_y)(1 - \tau/\tau_x)(I + \tau_y H_2) + (\tau/\tau_x)(\tau/\tau_y)(I + \tau_x H_x)(I + \tau_y H_y).$$

Здесь I – единичная матрица, матрицы $(I + \tau_x H_1)$, $(I + \tau_y H_2)$ – матрицы перехода “одномерных” разностных схем с временного слоя t_n на слой $t_n + \tau$ вдоль координатных осей x, y соответственно, $\tau_x > 0$, $\tau_y > 0$ – “шаги по времени” этих одномерных схем. Тогда

$$\|H\| \le |(1 - \tau/\tau_x)(1 - \tau/\tau_y)|\|I\| + |(1 - \tau/\tau_y)|(\tau/\tau_x)\|I + \tau_x H_1\| + |(1 - \tau/\tau_x)|(\tau/\tau_y)\|I + \tau_y H_2\| +$$
$$+ (\tau/\tau_x)(\tau/\tau_y)\|I + \tau_x H_1\|\|I + \tau_y H_2\|.$$

Поскольку одномерные схемы являются схемами Годунова и устойчивы при условиях

$$\tau_x \le h_x/a, \quad \tau_y \le h_y/b$$

соответственно, то для этих схем выполняются следующие неравенства:

$$\|I + \tau_x H_1\| \le 1, \quad \|I + \tau_y H_2\| \le 1.$$

Следовательно, для устойчивости схемы (2.4), задаваемой матрицей H, достаточно шаг интегрирования по времени выбрать из условия

$$\tau = \min(\tau_x, \tau_y).$$

Если коэффициенты в уравнении (2.1) переменные, то операторы расщепления в общем случае будут некоммутативными, т.е. $\Lambda_1\Lambda_2 \ne \Lambda_2\Lambda_1$, и точность решений будет зависеть от направления расщепления. Для устранения этой неопределенности и повышения точности строятся симметричные схемы расщепления (см. [20], [21]). Построим такую схему, которая соответствует усредненному оператору $\Lambda = 0.5(\Lambda_1\Lambda_2 + \Lambda_2\Lambda_1)$ (см. [20]):

$$u^{j,k} = 0.5[(E - \tau\Lambda_2)(E - \tau\Lambda_1) + (E - \tau\Lambda_1)(E - \tau\Lambda_2)]u_{j,k}. \tag{2.5}$$

Из (2.5) следует, что время счета по симметричной схеме в сравнении со счетом по схеме (2.4), в которой операторы расщепления коммутативные, увеличивается в два раза. Очевидно, что затраты для решения трехмерных уравнений переноса будут еще бóльшими. Оказывается, что эти затраты можно существенно уменьшить, если от схемы (2.5) перейти к эквивалентной схеме предиктор-корректор. Раскрыв скобки в (2.5) и сгруппировав члены, схему (2.5) запишем в виде

$$u^{j,k} = [E - \tau\Lambda_1^* - \tau\Lambda_2^*]u_{j,k}.$$

Здесь $\Lambda_1^* u_{j,k} = \Lambda_1(E - 0.5\tau\Lambda_2)u_{j,k}$, $\Lambda_2^* u_{j,k} = \Lambda_2(E - 0.5\tau\Lambda_1)u_{j,k}$.

Если ввести обозначения

$$U^*_{j+1/2,k} = [(E-0.5\tau\Lambda_2)u_{j,k}]_{j+1/2,k} = u_{j,k} - 0.5\tau[\Lambda_2 u_{j,k}]_{j+1/2,k} = U_{j+1/2,k} - 0.5c_y(U_{j,k+1/2} - U_{j,k-1/2}),$$
$$U^*_{j,k+1/2} = [(E-0.5\tau\Lambda_1)u_{j,k}]_{j,k+1/2} = u_{j,k} - 0.5\tau[\Lambda_1 u_{j,k}]_{j,k+1/2} = U_{j,k+1/2} - 0.5c_x(U_{j+1/2,k} - U_{j-1/2,k}), \quad (2.6)$$

то схема (2.5) может быть записана в форме схемы предиктор-корректор в виде

$$u^{j,k} = u_{j,k} - \tau a\frac{U^*_{j+1/2,k} - U^*_{j-1/2,k}}{h_j} - \tau b\frac{U^*_{j,k+1/2} - U^*_{j,k-1/2}}{h_k}. \quad (2.7)$$

В схеме (2.7) новые большие величины $U^*_{j+1/2,k}$, $U^*_{j-1/2,k}$ ... вычисляются из реконструкции (2.6) больших величин схемы Годунова. Реконструкция является экономичной, поскольку все величины известны и относятся к граням ячейки, в которой рассчитывается новое состояние. Поэтому численные решения уравнения (2.1) повышенной точности можно получать по новой схеме (2.7) с существенно меньшими затратами, чем по схеме (2.5) с покомпонентным расщеплением.

2.2. Уравнения переноса с тремя пространственными переменными

Рассмотрим задачу Коши для решения линейного уравнения переноса с тремя пространственными переменными

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} + b\frac{\partial u}{\partial y} + d\frac{\partial u}{\partial z} = 0, \quad a > 0, \quad b > 0, \quad d > 0 - \text{const}. \quad (2.8)$$

Поскольку построение разностной схемы Годунова для решения трехмерного уравнения переноса (2.8) не отличается от построения схемы (2.2) для решения двумерного уравнения переноса, то запишем сразу схему в операторной форме:

$$u^{j,k,i} = [E - \tau\Lambda_1 - \tau\Lambda_2 - \tau\Lambda_3]u_{j,k,i}. \quad (2.9)$$

Здесь $\Lambda_3 = d(U_{j,k,i+1/2} - U_{j,k,i+1/2})/h_z$ – это разностный оператор, который аппроксимирует дифференциальный оператор $d\dfrac{\partial}{\partial z}$. В [12] показано, что схема (2.9) при числах Куранта, равных единице, не удовлетворяет условию сдвига $u^{j,k,i} = u_{j-1,k-1,i-1}$. Если решение уравнения (2.8) получать методом покомпонентного расщепления (см. [15])

$$u^{j,k} = (E - \tau\Lambda_3)(E - \tau\Lambda_2)(E - \tau\Lambda_1)u_{j,k} =$$
$$= [E - \tau(\Lambda_1 + \Lambda_2 + \Lambda_3) + \tau^2(\Lambda_1\Lambda_2 + \Lambda_2\Lambda_3 + \Lambda_1\Lambda_3) - \tau^3\Lambda_1\Lambda_2\Lambda_3]u_{j,k}, \quad (2.10)$$

то разностная схема этому условию удовлетворяет. Устойчивость схемы доказывается так же, как была доказана устойчивость схемы (2.4). Шаг интегрирования по времени достаточно выбирать из условия $\tau = \min(\tau_x, \tau_y, \tau_z)$, где τ_z – шаг для "одномерной" схемы вдоль оси z.

В случае переменных коэффициентов операторы расщепления Λ_1, Λ_2, Λ_3 некоммутативные. Поэтому для повышения точности построим симметричную схему, как это было сделано для решения двумерного уравнения переноса (2.1). Для этого рассмотрим комбинации, в которых симметричные схемы уже выписаны для уравнений с двумя переменными. Тогда останется рассмотреть только две группы различных схем расщепления. Первая – это группа, в которой расщепление вначале проведено в одной из координатных плоскостей, а затем вдоль соответствующей координатной оси. Во второй группе расщепление проводится вначале вдоль координатной оси, а затем в соответствующей координатной плоскости. Представителями этих групп будут, соответственно, следующие схемы:

$$u^{j,k,i} = (E - \tau\Lambda_3)[E - \tau\Lambda_1(E - 0.5\tau\Lambda_2) - \tau\Lambda_2(E - 0.5\tau\Lambda_1)]u_{j,k,i},$$
$$u^{j,k,i} = [E - \tau\Lambda_1(E - 0.5\tau\Lambda_2) - \tau\Lambda_2(E - 0.5\tau\Lambda_1)](E - \tau\Lambda_3)u_{j,k,i}.$$

Остальные две пары выписываются аналогичным образом. Если просуммировать все уравнения этих групп (их будет 6), сделать приведение подобных членов и разделить на 6, то получим симметричную схему, которую запишем в виде

$$u^{j,k,i} = \Big\{E - \tau\Lambda_1\Big[E - \tau 0.5\Lambda_2\Big(E - \frac{1}{3}\tau\Lambda_3\Big) - \tau 0.5\Lambda_3\Big(E - \frac{1}{3}\tau\Lambda_2\Big)\Big] -$$

$$- \tau\Lambda_2\Big[E - \tau 0.5\Lambda_1\Big(E - \tau\frac{1}{3}\Lambda_3\Big) - \tau 0.5\Lambda_3\Big(E - \tau\frac{1}{3}\Lambda_1\Big)\Big] - \qquad (2.11)$$

$$- \tau\Lambda_3\Big[E - \tau 0.5\Lambda_2\Big(E - \tau\frac{1}{3}\Lambda_1\Big) - \tau 0.5\Lambda_1\Big(E - \tau\frac{1}{3}\Lambda_2\Big)\Big]\Big\}u_{j,k,i}.$$

Если ввести, как и в двумерном случае, обозначения

$$U^*_{j+1/2,k,i} = \Big\{\Big[E - \tau 0.5\Lambda_2\Big(E - \tau\frac{1}{3}\Lambda_3\Big) - \tau 0.5\Lambda_3\Big(E - \tau\frac{1}{3}\Lambda_2\Big)\Big]u_{j,k,i}\Big\}_{j+1/2,k,i} =$$

$$= U_{j+1/2,k,i} - 0.5\tau\frac{a}{h_y}(U^{**}_{j,k+1/2,i} - U^{**}_{j,k-1/2,i}) - 0.5\tau\frac{d}{h_z}(U^{**}_{j,k,i+1/2} - U^{**}_{j,k,i-1/2}),$$

$$U^{**}_{j,k+1/2,i} = \Big[\Big(E - \tau\frac{1}{3}\Lambda_3\Big)u_{j,k,i}\Big]_{j,k+1/2,i} = U_{j,k+1/2,i} - \frac{1}{3}\tau\frac{d}{h_z}(U_{j,k,i+1/2} - U_{j,k,i-1/2}), \ldots$$

$$\ldots, U^{**}_{j,k,i+1/2} = \Big[\Big(E - \tau\frac{1}{3}\Lambda_2\Big)u_{j,k,i}\Big]_{j,k,i+1/2} = U_{j,k,i+1/2} - \frac{1}{3}\tau\frac{b}{h_y}(U_{j,k+1/2,i} - U_{j,k-1/2,i}), \ldots$$

$$\ldots, U^*_{j,k+1/2,i} = \ldots, \quad U^*_{j,k,i+1/2} = \ldots,$$

то схему (2.11) можно записать в классической форме схемы предиктор-корректор в виде

$$u^{j,k,i} = u_{j,k,i} - c_x(U^*_{j+1/2,k,i} - U^*_{j-1/2,k,i}) - c_y(U^*_{j,k+1/2,i} - U^*_{j,k-1/2,i}) - c_z(U^*_{j,k,i+1/2} - U^*_{j,k,i-1/2}). \qquad (2.12)$$

Величины $U^*_{j+1/2,k,i}$, $U^*_{j-1/2,k,i}$, ... в (2.12) получены, как и в двумерном случае, реконструкцией больших величин схемы Годунова из уравнений

$$U^*_{j+1/2,k,i} = U_{j+1/2,k,i} - 0.5c_y(U_{j,k+1/2,i} - U_{j,k-1/2,i}) - 0.5c_z(U_{j,k,i+1/2} - U_{j,k,i-1/2}) +$$
$$+ \frac{1}{3}0.5c_yc_z[(U_{j,k+1/2,i} - U_{j,k-1/2,i}) - (U_{j,k+1/2,i-1} - U_{j,k-1/2,i-1})] +$$
$$+ \frac{1}{3}0.5c_yc_z[(U_{j,k,i+1/2} - U_{j,k,i-1/2}) - (U_{j,k-1,i+1/2} - U_{j,k-1,i-1/2})],$$

$$U^*_{j,k+1/2,i} = U_{j,k+1/2,i} - 0.5c_x(U_{j+1/2,k,i} - U_{j-1/2,k,i}) - 0.5c_z(U_{j,k,i+1/2} - U_{j,k,i-1/2}) +$$
$$+ \frac{1}{3}0.5c_xc_z[(U_{j,k,i+1/2} - U_{j,k,i-1/2}) - (U_{j-1,k,i+1/2} - U_{j-1,k,i-1/2})] + \qquad (2.13)$$
$$+ \frac{1}{3}0.5c_xc_z[(U_{j+1/2,k,i} - U_{j-1/2,k,i}) - (U_{j+1/2,k,i-1} - U_{j-1/2,k,i-1})],$$

$$U^*_{j,k,i+1/2} = U_{j,k,i+1/2} - 0.5c_x(U_{j+1/2,k,i} - U_{j-1/2,k,i}) - 0.5c_y(V_{j,k+1/2,i} - V_{j,k-1/2,i}) +$$
$$+ \frac{1}{3}c_x0.5c_y[(U_{j,k+1/2,i} - U_{j,k-1/2,i}) - (U_{j-1,k+1/2,i} - U_{j-1,k-1/2,i})] +$$
$$+ \frac{1}{3}c_y0.5c_x[(U_{j+1/2,k,i} - U_{j-1/2,k,i}) - (U_{j+1/2,k-1,i} - U_{j-1/2,k-1,i})].$$

Если операторы $\Lambda_1, \Lambda_2, \Lambda_3$ коммутативные, то схема (2.10) может быть также преобразована к схеме (2.11), (2.12). Таким образом, показано, что симметричные схемы покомпонентного расщепления (2.5), (2.11) могут быть записаны в форме схем предиктор-корректор (2.7), (2.12) соответственно. В этих схемах новые большие величины вычисляются на основе реконструкций, полученных при переходе от схем расщепления к схемам предиктор-корректор. Из метода построения симметричных схем расщепления и преобразования их в схемы предиктор-корректор следует

Теорема. *Пусть дано n-мерное уравнение переноса с переменными коэффициентами, которое решается методом Годунова. Тогда для решения этого уравнения может быть построена симметричная схема покомпонентного расщепления, которая после исключения дробных шагов преобразуется к эквивалентной схеме предиктор-корректор с вычислением больших величин на основе реконструкции, полученной при переходе от схемы расщепления к схеме предиктор-корректор.*

Теорема доказывается методом математической индукции.

3. МЕТОД ГОДУНОВА ДЛЯ РЕШЕНИЯ УРАВНЕНИЙ АКУСТИКИ

3.1. Уравнения акустики с двумя пространственными переменными

Рассмотрим задачу Коши для системы уравнений акустики с постоянными коэффициентами с двумя пространственными переменными:

$$\begin{gathered}\frac{du}{dt}+\frac{\partial p}{\partial x}=0,\\ \frac{dv}{dt}+\frac{\partial p}{\partial y}=0,\\ \frac{dp}{dt}+a_0^2\frac{\partial u}{\partial x}+a_0^2\frac{\partial v}{\partial y}=0,\end{gathered}\tag{3.1}$$

с начальными данными $u(0, x, y) = u(x, y)$, $v(0, x, y) = v(x, y)$, $p(0, x, y) = p(x, y)$. Требуется найти решение для $t > 0$. Здесь t, x, y – независимые переменные по времени и по пространству соответственно, u, v – компоненты вектора скорости вдоль координатных осей x, y соответственно, x, y – это массовые переменные $x = \rho_0\bar{x}$, $y = \rho_0\bar{y}$, $\bar{x}$, $\bar{y}$ – эйлеровы координаты, p – давление, $a_0 = \underline{\rho}_0 c_0$ – массовая скорость звука, ρ_0, c_0 – параметры вещества.

Построим в пространстве (t, x, y) разностную сетку с шагами τ, h_j, h_k, вдоль координатных линий соответственно. Основные величины скорости и давления отнесем к центрам ячеек и обозначим их через $(u^{j,k}, v^{j,k}, p^{j,k})$, $(u_{j,k}, v_{j,k}, p_{j,k})$ в моменты времени $t_n + \tau$, t_n соответственно. Большие величины, которые отнесены к граням двух соседних ячеек, обозначим в виде $U_{j+1/2,k}$, $P_{j+1/2,k}$ на вертикальных и $V_{j,k+1/2}$, $P_{j,k+1/2}$ на горизонтальных гранях. В принятых обозначениях разностную схему, которая построена на основе интегральных законов сохранения (см. [2]) для решения системы уравнений (3.1), запишем в виде

$$\begin{gathered}u^{j,k}=u_{j,k}-\frac{\tau}{h_j}(P_{j+1/2,k}-P_{j-1/2,k}),\\ v^{j,k}=v_{j,k}-\frac{\tau}{h_k}(P_{j,k+1/2}-P_{j,k-1/2}),\\ p^{j,k}=p_{j,k}-\frac{\tau a_0^2}{h_j}(U_{j+1/2,k}-U_{j-1/2,k})-\frac{\tau a_0^2}{h_k}(V_{j,k+1/2}-V_{j,k-1/2}).\end{gathered}\tag{3.2}$$

Большие величины $P_{j+1/2,k}$, $U_{j+1/2,k}$, ... находятся из решения задачи о плоском распаде произвольного разрыва, возникающего на вертикальных гранях двух соседних ячеек, из разностных уравнений (см. [2])

$$\begin{gathered}U_{j+1/2,k}=\frac{a_{j+1,k}u_{j+1,k}+a_{j,k}u_{j,k}}{a_{j+1,k}+a_{j,k}}-\frac{p_{j+1,k}-p_{j,k}}{a_{j+1,k}+a_{j,k}},\\ P_{j+1/2,k}=\frac{a_{j+1,k}p_{j,k}+a_{j,k}p_{j+1,k}}{a_{j+1,k}+a_{j,k}}-a_{j+1,k}a_{j,k}\frac{u_{j+1,k}-u_{j,k}}{a_{j+1,k}+a_{j,k}},\quad a_{j+1,k}=a_{j+1,k}^{(i-1)},\quad a_{j,k}=a_{j,k}^{(i-1)},\end{gathered}\tag{3.3}$$

и относятся к моментам времени $t_n + \tau_{j+1/2,k}$ (см. [12], [24]). Здесь $a_{j,k} = \rho_{j,k}c_{j,k}$ – это массовая скорость звука в ячейке, i – номер итерации. Величины с индексами j, k и $j+1, k$ относятся к ячейкам, которые находятся слева и справа от общей границы. Если $i = 1$, то решается линеаризованная задача о распаде разрыва без итераций. В случае решения уравнений акустики (3.1) $a_{j+1,k} = a_{j,k} = a_0$.

Рассмотрим некоторые интерпретации больших величин, которые понадобятся для построения схем повышенной точности. Введем обозначения

$$U^0_{j+1/2,k} = \frac{a_{j+1,k}u_{j+1,k} + a_{j,k}u_{j,k}}{a_{j+1,k} + a_{j,k}}, \quad P^0_{j+1/2,k} = \frac{a_{j+1,k}p_{j,k} + a_{j,k}p_{j+1,k}}{a_{j+1,k} + a_{j,k}},$$
$$\tau_{j+1/2,k} = 0.5(h_{j+1} + h_j)/(a_{j+1,k} + a_{j,k})$$

и запишем уравнения (3.3) в виде

$$U_{j+1/2,k} = U^0_{j+1/2,k} - \tau_{j+1/2,k}\frac{p_{j+1,k} - p_{j,k}}{0.5(h_{j+1} + h_j)},$$
$$P_{j+1/2,k} = P^0_{j+1/2,k} - \tau_{j+1/2,k}a_{j+1,k}a_{j,k}\frac{u_{j+1,k} - u_{j,k}}{0.5(h_{j+1} + h_j)}. \tag{3.4}$$

Из (3.4) следует, что большие величины на границах ячеек вычисляются из разностных уравнений, которые аппроксимируют дифференциальные уравнения

$$\frac{\partial u}{\partial t} + \frac{\partial p}{\partial x} = 0,$$
$$\frac{\partial p}{\partial t} + a^2_{j+1/2,k}\frac{\partial u}{\partial x} = 0, \quad a^2_{j+1/2,k} = a_{j+1,k}a_{j,k}. \tag{3.5}$$

В случае уравнений акустики (3.1) $\tau_{j+1/2,k} = \tau_{j-1/2,k}$ и формулы (3.3) принимают вид

$$P_{j+1/2,k} = 0.5(p_{j+1,k} + p_{j,k}) - 0.5a_0(u_{j+1,k} - u_{j,k}),$$
$$U_{j+1/2,k} = 0.5(u_{j+1,k} + u_{j,k}) - 0.5(p_{j+1,k} - p_{j,k})/a_0.$$

Большие величины $V_{j,k+1/2}$, $P_{j,k+1/2}$ относятся к моментам времени $t_n + \tau_{j,k+1/2}$, вычисляются с учетом индексов из уравнений, аналогичных уравнениям (3.3), и аппроксимируют вдоль второго направления соответствующую систему одномерных уравнений, аналогичную системе (3.5). Давления $P_{j+1/2,k}$, $P_{j,k+1/2}$ на гранях вычисляются из решения одномерных уравнений и поэтому не удовлетворяют двумерному уравнению в (3.1), которому удовлетворяет течение. Следовательно, это может быть источником ошибок аппроксимации. Шаг интегрирования по времени выбирается из условия устойчивости схемы

$$\tau \le \frac{\tau_x\tau_y}{\tau_x + \tau_y}.$$

Здесь τ_x, τ_y – шаги по времени для решения одномерных уравнений газовой динамики вдоль осей x, y соответственно. ДП схемы (3.2), (3.3) имеет следующий вид:

$$\frac{du}{dt} + \frac{\partial p}{\partial x} = \left(\frac{\tau_{j+1/2,k} + \tau_{j-1/2,k}}{2} - 0.5\tau\right)a_0^2\frac{\partial^2 u}{\partial x^2} - 0.5\tau a_0^2\frac{\partial^2 v}{\partial x\partial y},$$
$$\frac{dv}{dt} + \frac{\partial p}{\partial y} = -0.5\tau a_0^2\frac{\partial^2 u}{\partial y\partial x} + \left(\frac{\tau_{j,k+1/2} + \tau_{j,k-1/2}}{2} - 0.5\tau\right)a_0^2\frac{\partial^2 v}{\partial y^2}, \tag{3.6}$$
$$\frac{dp}{dt} + a_0^2\frac{\partial u}{\partial x} + a_0^2\frac{\partial v}{\partial y} = \left(\frac{\tau_{j+1/2,k} + \tau_{j-1/2,k}}{2} - 0.5\tau\right)a_0^2\frac{\partial^2 p}{\partial x^2} - \left(\frac{\tau_{j,k+1/2} + \tau_{j,k-1/2}}{2} - 0.5\tau\right)a_0^2\frac{\partial^2 p}{\partial y^2}.$$

Из ДП (3.6) следует, что схема (3.2) аппроксимирует исходные уравнения акустики (3.1) с первым порядком по времени и по пространству. Если числа Куранта $c_x = a_0\tau/h_x$, $c_y = a_0\tau/h_y$ равны едини-

це, то правая часть в ДП (3.6) не обращается в нуль, т.е. ДП схемы не удовлетворяет условию сдвига, как это имеет место для одномерных схем Годунова.

3.2. Корректность задачи Коши дифференциального приближения схемы с двумя пространственными переменными

Система уравнений ДП (3.6) в качестве параметров содержит шаг интегрирования по времени и шаги $\tau_{j+1/2,k}$, $\tau_{j,k+1/2}$, от которых зависят свойства решений, в частности, корректность задачи Коши. Поэтому рассмотрим в зависимости от соотношений между этими параметрами корректность задачи Коши для ДП (3.6) в классе финитных функций с ограниченным носителем. Введем в этом классе норму

$$\|\{u, v, p\}\| = \left[\int_S \left(\frac{u^2}{2} + \frac{v^2}{2} + \frac{p^2}{2a_0^2}\right) dxdy\right]^{1/2}.$$

Здесь S ограниченный носитель финитных функций в двумерном пространстве (x, y). Умножив первое уравнение в ДП на u, второе на v, третье на p/a^2, сложив их и проинтегрировав по всему пространству, получим

$$\int\left[\frac{\partial}{\partial t}\left(\frac{u^2}{2} + \frac{v^2}{2} + \frac{p^2}{2a_0^2}\right) + \frac{\partial(pu)}{\partial x} + \frac{\partial(pv)}{\partial y}\right] dxdy =$$

$$= \frac{\partial}{\partial t}\left[\int\left(\frac{u^2}{2} + \frac{v^2}{2} + \frac{p^2}{2a_0^2}\right) dxdy\right] + \int\left[\frac{\partial(pu)}{\partial x} + \frac{\partial(pv)}{\partial y}\right] dxdy =$$

$$= \int\left[(\tau_j^* - 0.5\tau)a_0^2 u\frac{\partial^2 u}{\partial x^2} - 0.5\tau a_0^2 u\frac{\partial^2 v}{\partial x\partial y} - 0.5\tau v\frac{\partial^2 u}{\partial y\partial x} + (\tau_k^* - 0.5\tau)a_0^2 v\frac{\partial^2 v}{\partial y^2}\right] dxdy +$$

$$+ \int\left[(\tau_j^* - 0.5\tau)p\frac{\partial^2 p}{\partial x^2} + (\tau_k^* - 0.5\tau)p\frac{\partial^2 p}{\partial y^2}\right] dxdy.$$

Здесь введены обозначения $\tau_j^* = 0.5(\tau_{j+1/2,k} + \tau_{j-1/2,k})$, $\tau_k^* = 0.5(\tau_{j,k+1/2} + \tau_{j,k-1/2})$.

Проинтегрировав по частям интегралы по пространству с учетом свойств финитных функций, получим

$$\frac{\partial}{\partial t}\left[\int_S\left(\frac{u^2}{2} + \frac{v^2}{2} + \frac{p^2}{2a_0^2}\right) dxdy\right] = \frac{\partial}{\partial t}\|\{u, v, p\}\|^2 = 2\|\{u, v, p\}\|\frac{\partial}{\partial t}\|\{u, v, p\}\| =$$

$$= -\int\left[(\tau_j^* - 0.5\tau)\left(\frac{\partial u}{\partial x}\right)^2 - \tau\frac{\partial u}{\partial x}\frac{\partial v}{\partial y} + (\tau_k^* - 0.5\tau)\left(\frac{\partial v}{\partial y}\right)^2\right] dxdy -$$

$$- \int\left[(\tau_j^* - 0.5\tau)\left(\frac{\partial p}{\partial x}\right)^2 + (\tau_k^* - 0.5\tau)\left(\frac{\partial p}{\partial y}\right)^2\right] dxdy.$$

Следовательно, если подынтегральные выражения в правой части будут больше нуля, то норма функций со временем возрастать не будет. Поскольку подынтегральные выражения являются квадратичными формами, то условия положительности этих форм следует из положительности определителя Сильверста. Определители Сильверста для форм в первом и втором интегралах в правой части имеют вид

$$D_1 = \begin{vmatrix} (\tau_j^* - 0.5\tau) & -0.5\tau \\ -0.5\tau & (\tau_k^* - 0.5\tau) \end{vmatrix} \geq 0 \Rightarrow (\tau_j^* - 0.5\tau)(\tau_k^* - 0.5\tau) - (0.5\tau)^2 > 0,$$

$$D_2 = (\tau_j^* - 0.5\tau)(\tau_k^* - 0.5\tau) \geq 0$$

соответственно. Откуда следует условие

$$0.5\tau \le \frac{\tau_j^* \tau_k^*}{\tau_j^* + \tau_k^*} \Rightarrow \tau \le \frac{\tau_x \tau_y}{\tau_x + \tau_y}$$

ограничения на шаг интегрирования по времени, при котором норма функций со временем возрастать не будет. Это условие совпадает с условием устойчивости разностной схемы (3.2) из [2]. Этим фактом мы в дальнейшем будем пользоваться при исследовании устойчивости разностных схем на основе ДП. На равномерной по пространству разностной сетке (квадратной) шаг интегрирования по времени в два раза меньше, чем шаг интегрирования в схемах для решения одномерных задач.

3.3. Уравнения акустики с тремя пространственными переменными

Рассмотрим задачу Коши для системы уравнений акустики с постоянными коэффициентами с тремя пространственными переменными:

$$\begin{aligned} \frac{du}{dt} + \frac{\partial p}{\partial x} &= 0, \\ \frac{dv}{dt} + \frac{\partial p}{\partial y} &= 0, \\ \frac{dw}{dt} + \frac{\partial p}{\partial z} &= 0, \\ \frac{dp}{dt} + a_0^2 \frac{\partial u}{\partial x} + a_0^2 \frac{\partial v}{\partial y} + a_0^2 \frac{\partial w}{\partial z} &= 0; \end{aligned} \tag{3.8}$$

начальные данные $u(0, x, y, z) = u(x, y, z)$, $v(0, x, y, z) = v(x, y, z)$, $w(0, x, y, z) = w(x, y, z)$, $p(0, x, y, z) = p(x, y, z)$. Требуется найти решение для $t > 0$. Здесь t, x, y, z – независимые переменные по времени и по пространству соответственно, u, v, w – компоненты вектора скорости вдоль координатных осей x, y, z соответственно, x, y, z – это массовые переменные $x = \rho_0 \bar{x}$, $y = \rho_0 \bar{y}$, $z = \rho_0 \bar{z}$, $\bar{x}$, $\bar{y}$, $\bar{z}$ – эйлеровы координаты, p – давление, $a_0 = \rho_0 c_0$ – массовая скорость звука, ρ_0, c_0 – параметры вещества.

Разностная схема предиктор-корректор для решения системы уравнений (3.8) методом Годунова строится так же, как и схема (3.2), и имеет следующий вид:

$$\begin{aligned} u^{j,k,i} &= u_{j,k,i} - \frac{\tau}{h_j}(P_{j+1/2,k,i} - P_{j-1/2,k,i}), \\ v^{j,k,i} &= v_{j,k,i} - \frac{\tau}{h_k}(P_{j,k+1/2,i} - P_{j,k-1/2,i}), \\ w^{j,k,i} &= w_{j,k,i} - \frac{\tau}{h_i}(P_{j,k,i+1/2} - P_{j,k,i-1/2}), \end{aligned} \tag{3.9}$$

$$p^{j,k,i} = p_{j,k,i} - \frac{\tau a_0^2}{h_j}(U_{j+1/2,k,i} - U_{j-1/2,k,i}) - \frac{\tau a_0^2}{h_k}(V_{j,k+1/2,i} - V_{j,k-1/2,i}) - \frac{\tau a_0^2}{h_i}(W_{j,k,i+1/2} - W_{j,k,i-1/2}).$$

Большие величины $W_{j,k,j+1/2}$, $P_{j,k,i+1/2}$ относятся к моменту времени $t_n + \tau_{j,k,j+1/2}$, вычисляются с учетом индексов из уравнений, аналогичных уравнениям (3.3), и аппроксимируют систему одномерных уравнений, аналогичную системе (3.5). Схема (3.9) аналогична схеме (3.2) и поэтому обладает теми же свойствами, что и схема (3.2). Давления на гранях вычисляются из решения одномерных уравнений и поэтому не удовлетворяют трехмерному уравнению в (3.8), которому удовлетворяет течение. Шаг интегрирования по времени выбирается из условия устойчивости схемы

$$\tau \le \frac{\tau_x \tau_y \tau_z}{\tau_x \tau_y + \tau_y \tau_z + \tau_x \tau_z} \tag{3.10}$$

и на равномерных по пространству сетках (кубических) в три раза меньше, чем в одномерной схеме. Следовательно, оба фактора могут быть источниками ошибок аппроксимации. ДП схемы (3.9) запишем в виде

$$\begin{aligned}
\frac{du}{dt} + \frac{\partial p}{\partial x} &= \left(\tau_j^* - \frac{\tau}{2}\right) a_0^2 \frac{\partial^2 u}{\partial x^2} - \frac{\tau}{2} a_0^2 \frac{\partial^2 v}{\partial x \partial y} - \frac{\tau}{2} a_0^2 \frac{\partial^2 w}{\partial x \partial z}, \\
\frac{dv}{dt} + \frac{\partial p}{\partial y} &= -\frac{\tau}{2} a_0^2 \frac{\partial^2 u}{\partial y \partial x} + \left(\tau_k^* - \frac{\tau}{2}\right) a_0^2 \frac{\partial^2 v}{\partial y^2} - \frac{\tau}{2} a_0^2 \frac{\partial^2 w}{\partial z^2}, \\
\frac{dw}{dt} + \frac{\partial p}{\partial z} &= -\frac{\tau}{2} a_0^2 \frac{\partial^2 u}{\partial z \partial x} - \frac{\tau}{2} a_0^2 \frac{\partial^2 v}{\partial y \partial z} + \left(\tau_i^* - \frac{\tau}{2}\right) a_0^2 \frac{\partial^2 w}{\partial z^2}, \\
\frac{dp}{dt} + a_0^2 \frac{\partial u}{\partial x} + a_0^2 \frac{\partial v}{\partial y} + a_0^2 \frac{\partial w}{\partial z} &= (\tau_j^* - 0.5\tau) a_0^2 \frac{\partial^2 p}{\partial x^2} - (\tau_k^* - 0.5\tau) a_0^2 \frac{\partial^2 p}{\partial y^2} + (\tau_i^* - 0.5\tau) a_0^2 \frac{\partial^2 p}{\partial z^2}.
\end{aligned} \tag{3.11}$$

Здесь $\tau_j^* = 0.5(\tau_{j+1/2,k,i} + \tau_{j-1/2,k,i})$, $\tau_k^* = 0.5(\tau_{j,k+1/2,i} + \tau_{j,k-1/2,i})$, $\tau_i^* = 0.5(\tau_{j,k,i+1/2} + \tau_{j,k,i-1/2})$. Из ДП (3.11) следует, что схема аппроксимирует систему уравнений с первым порядком по времени и по пространству. Если числа Куранта равны единице, то правая часть в ДП (3.11) не обращается в нуль, т.е. ДП схемы не удовлетворяет условию сдвига. Корректность задачи Коши для ДП (3.11) в зависимости от шагов интегрирования по времени и шагов τ_j^*, τ_k^*, τ_i^* доказывается так же, как была доказана корректность ДП (3.6) схемы с двумя переменными. Условия корректности ДП (3.11) следуют из условия положительности определителей Сильверста для квадратичных форм и совпадают с условиями устойчивости (3.10) схемы (3.9).

4. СИММЕТРИЧНЫЕ РАЗНОСТНЫЕ СХЕМЫ РАСЩЕПЛЕНИЯ ПО ПРОСТРАНСТВЕННЫМ ПЕРЕМЕННЫМ И ЭКВИВАЛЕНТНЫЕ ИМ СХЕМЫ ПРЕДИКТОР-КОРРЕКТОР

4.1. Разностные схемы для решения уравнений акустики с двумя пространственными переменными

Рассмотрим решение системы уравнений (3.1) методом покомпонентного расщепления (см. [15]) по явным разностным схемам. Согласно этому методу, система (3.1) записывается в виде двух одномерных систем:

$$\Lambda_1 = \begin{cases} \dfrac{du}{dt} + \dfrac{\partial p}{\partial x} = 0, \\ \dfrac{dv}{dt} = 0, \\ \dfrac{dp}{dt} + a_0^2 \dfrac{\partial u}{\partial x} = 0, \end{cases} \qquad \Lambda_2 = \begin{cases} \dfrac{du}{dt} = 0, \\ \dfrac{dv}{dt} + \dfrac{\partial p}{\partial y} = 0, \\ \dfrac{dp}{dt} + a_0^2 \dfrac{\partial v}{\partial y} = 0, \end{cases}$$

которые последовательно решаются. На первом шаге решается первая система уравнений, вдоль оси x, на втором шаге – вторая, вдоль оси y. Начальными данными для второй системы уравнений является решение первой системы уравнений. Последовательность решения может быть и такой: на первом шаге решается вторая система уравнений, на втором – первая. Операторы решений, как и разностные схемы, в первом и втором случаях обозначим через $\Lambda_{12} = \Lambda_2\Lambda_1$, $\Lambda_{21} = \Lambda_1\Lambda_2$ соответственно. Явные разностные схемы для операторов расщепления Λ_1, Λ_2 запишем в следующих видах:

$$\Lambda_1 = \begin{cases} u^{j,k} = u_{j,k} - \dfrac{\bar\tau_j}{h_j}(P_{j+1/2,k} - P_{j-1/2,k}), \\ v^{j,k} = v_{j,k}, \\ \bar p^{j,k} = p_{j,k} - \dfrac{\bar\tau_j a_0^2}{h_j}(U_{j+1/2,k} - U_{j-1/2,k}), \end{cases} \qquad \Lambda_2 = \begin{cases} u^{j,k} = u^{j,k}, \\ v^{j,k} = v_{j,k} - \dfrac{\tau}{h_k}(\bar P_{j,k+1/2} - \bar P_{j,k-1/2}), \\ p^{j,k} = \bar p_{j,k} - \dfrac{\tau a_0^2}{h_k}(V_{j,k+1/2} - V_{j,k-1/2}) \end{cases} \tag{4.1}$$

соответственно. Схемы (4.1) – это схемы Годунова для решения одномерных задач. В этих схемах большие величины компонентов скорости вычисляются, как в схеме Годунова, а давления $\bar{P}_{j+1/2,k}$, $\bar{P}_{j,k+1/2}$ рассчитываются из (3.3) с учетом расщепления по пространственным переменным. Это означает, что давления $\bar{P}_{j+1/2,k}$, $\bar{P}_{j,k+1/2}$ вычисляются из решений одномерных задач о плоском распаде разрыва по величинам, которые были получены на первом шаге по схемам Λ_2, Λ_1 в моменты времени $t_n + \bar{\tau}_k$, $t_n + \bar{\tau}_j$ соответственно. В случае уравнений акустики с постоянными коэффициентами будем иметь

$$\bar{P}_{j+1/2,k} = P_{j+1/2,k} - \frac{\bar{\tau}_k a_0^2}{2h_k}[(V_{j+1,k+1/2} - V_{j+1,k-1/2}) + (V_{j,k+1/2} - V_{j,k-1/2})],$$

$$\bar{P}_{j,k+1/2} = P_{j,k+1/2} - \frac{\bar{\tau}_j a_0^2}{2h_j}[(U_{j+1/2,k+1} - U_{j-1/2,k+1}) + (U_{j+1/2,k} - U_{j-1/2,k})].$$

Разностную схему Λ_{12} после исключения дробного шага запишем в виде

$$u^{j,k} = u_{j,k} - \frac{\tau}{h_j}(P_{j+1/2,k} - P_{j-1/2,k}),$$

$$v^{j,k} = v_{j,k} - \frac{\tau}{h_k}(P_{j,k+1/2} - P_{j,k-1/2}) + \frac{\tau}{h_k}\frac{\bar{\tau}_j a_0^2}{2h_j}[(U_{j+1/2,k+1} - U_{j-1/2,k+1}) + (U_{j+1/2,k-1} - U_{j-1/2,k-1})],$$

$$p^{j,k} = p_{j,k} - \frac{\bar{\tau}_j a_0^2}{h_j}(U_{j+1/2,k} - U_{j-1/2,k}) - \frac{\tau a_0^2}{h_k}(V_{j,k+1/2} - V_{j,k-1/2}).$$

Разностную схему Λ_{21} после исключения дробного шага запишем в виде

$$u^{j,k} = u_{j,k} - \frac{\tau}{h_j}(P_{j+1/2,k} - P_{j-1/2,k}) + \frac{\tau}{h_j}\frac{\bar{\tau}_k a_0^2}{2h_k}[(V_{j+1,k+1/2} - V_{j+1,k-1/2}) + (V_{j-1,k+1/2} - V_{j-1,k-1/2})],$$

$$v^{j,k} = v_{j,k} - \frac{\bar{\tau}_k}{h_k}(P_{j,k+1/2} - P_{j,k-1/2}),$$

$$p^{j,k} = p_{j,k} - \frac{\bar{\tau}_k a_0^2}{h_j}(U_{j+1/2,k} - U_{j-1/2,k}) - \frac{\tau a_0^2}{h_k}(V_{j,k+1/2} - V_{j,k-1/2}).$$

Очевидно, что операторы расщепления некоммутативные, т.е. $\Lambda_2\Lambda_1 \neq \Lambda_1\Lambda_2$. Для повышения точности решений составим усредненный симметричный оператор (см. [20]) в виде

$$\Lambda = 0.5(\Lambda_2\Lambda_1 + \Lambda_1\Lambda_2).$$

Разностная схема, соответствующая этому оператору, имеет вид

$$u^{j,k} = u_{j,k} - \frac{\tau}{h_j}(P_{j+1/2,k} - P_{j-1/2,k}) + \frac{\tau}{2h_j}\frac{\bar{\tau}_k a_0^2}{2h_k}[(V_{j+1,k+1/2} - V_{j+1,k-1/2}) + (V_{j-1,k+1/2} - V_{j-1,k-1/2})],$$

$$v^{j,k} = v_{j,k} - \frac{\tau}{h_k}(P_{j,k+1/2} - P_{j,k-1/2}) + \frac{\tau}{2h_k}\frac{\bar{\tau}_j a_0^2}{2h_j}[(U_{j+1/2,k+1} - U_{j-1/2,k+1}) + (U_{j+1/2,k-1} - U_{j-1/2,k-1})], \quad (4.3)$$

$$p^{j,k} = p_{j,k} - \frac{0.5(\bar{\tau}_k + \tau)a_0^2}{h_j}(U_{j+1/2,k} - U_{j-1/2,k}) - \frac{0.5(\bar{\tau}_j + \tau)a_0^2}{h_k}(V_{j,k+1/2} - V_{j,k-1/2}).$$

Положив $\bar{\tau}_j = \bar{\tau}_k = \tau$, получим симметричную схему, ДП которой запишем в виде

$$\begin{aligned}
\frac{du}{dt} + \frac{\partial p}{\partial x} &= \left(\frac{\tau_{j+1/2,k} + \tau_{j-1/2,k}}{2} - 0.5\tau\right) a_0^2 \frac{\partial^2 u}{\partial x^2}, \\
\frac{dv}{dt} + \frac{\partial p}{\partial y} &= \left(\frac{\tau_{j,k+1/2} + \tau_{j,k-1/2}}{2} - 0.5\tau\right) a_0^2 \frac{\partial^2 v}{\partial y^2}, \\
\frac{dp}{dt} + a_0^2 \frac{\partial u}{\partial x} + a_0^2 \frac{\partial v}{\partial y} &= \left(\frac{\tau_{j+1/2,k} + \tau_{j-1/2,k}}{2} - 0.5\tau\right) a_0^2 \frac{\partial^2 p}{\partial x^2} + \left(\frac{\tau_{j,k+1/2} + \tau_{j,k-1/2}}{2} - 0.5\tau\right) a_0^2 \frac{\partial^2 p}{\partial y^2}.
\end{aligned} \tag{4.4}$$

Условие корректности ДП (4.4) принимает вид $\tau \le \min(\tau_x, \tau_y)$. На равномерной по пространству сетке (квадратной) шаг интегрирования по времени в симметричной схеме (4.3) согласуется с шагом в схемах для решения одномерных задач и в два раза больше, чем шаг по времени в схеме без расщепления. Если числа Куранта c_x, c_y равны единице, то правая часть в ДП (4.4) обращается в ноль и, следовательно, ДП удовлетворяет условию сдвига. Если ввести с учетом уравнений (3.3) обозначения

$$\begin{aligned}
P^*_{j+1/2,k} &= P_{j+1/2,k} - \\
&- \frac{\bar{\tau}_k}{2h_k}\left[\frac{a_{j,k}}{a_{j+1,k} + a_{j,k}} a^2_{j+1,k}(U_{j+1,k+1/2} - U_{j+1/2,k-1/2}) + \frac{a_{j+1,k}}{a_{j+1,k} + a_{j,k}} a^2_{j,k}(U_{j,k+1/2} - U_{j,k-1/2})\right], \\
P^*_{j,k+1/2} &= P_{j,k+1/2} - \\
&- \frac{\bar{\tau}_j}{2h_j}\left[\frac{a_{j,k+1}}{a_{j,k+1} + a_{j,k}} a^2_{j,k}(U_{j+1/2,k} - U_{j-1/2,k}) + \frac{a_{j,k}}{a_{j,k+1} + a_{j,k}} a^2_{j,k+1}(U_{j+1/2,k+1} - U_{j-1/2,k+1})\right]
\end{aligned} \tag{4.5}$$

для новых больших величин давлений и положить $\bar{\tau}_j = \bar{\tau}_k = \tau$, то систему разностных уравнений (4.3) можно записать в виде схемы предиктор-корректор:

$$\begin{aligned}
u^{j,k} &= u_{j,k} - \frac{\tau}{h_j}(P^*_{j+1/2,k} - P^*_{j-1/2,k}), \\
v^{j,k} &= v_{j,k} - \frac{\tau}{h_k}(P^*_{j,k+1/2} - P^*_{j,k-1/2}), \\
p^{j,k} &= p_{j,k} - \frac{\tau a_0^2}{h_j}(U_{j+1/2,k} - U_{j-1/2,k}) - \frac{\tau a_0^2}{h_k}(V_{j,k+1/2} - V_{j,k-1/2}).
\end{aligned} \tag{4.6}$$

Следовательно, решение методом покомпонентного расщепления по симметричной схеме (4.3) будет эквивалентно решению по схеме предиктор-корректор (4.6), в которой новые большие величины давлений получены реконструкцией больших величин схемы Годунова. Реконструкция давлений осуществляется по уравнениям (4.5), которые аппроксимируют двумерное уравнение для давления в (3.1).

Из сравнения ДП (3.11) и (4.4) можно сказать, что источником ошибок аппроксимации в схеме (3.2) являются большие величины, которые вычисляются из решения одномерной задачи о плоском распаде разрыва.

4.2. Разностные схемы для решения уравнений акустики с тремя пространственными переменными

Рассмотрим решение системы уравнений (3.8) методом покомпонентного расщепления. С учетом исследований, проведенных в разд. 4, расщепленную систему уравнений запишем в виде

$$\Lambda_{12} = \begin{cases} \dfrac{dw}{dt} = 0, \\ \dfrac{du}{dt} + \dfrac{\partial p}{\partial x} = 0, \\ \dfrac{dv}{dt} + \dfrac{\partial p}{\partial y} = 0, \\ \dfrac{dp}{dt} + a_0^2 \dfrac{\partial u}{\partial x} + a_0^2 \dfrac{\partial v}{\partial y} = 0, \end{cases} \qquad \Lambda_3 = \begin{cases} \dfrac{du}{dt} = 0, \quad \dfrac{dv}{dt} = 0, \\ \dfrac{dw}{dt} + \dfrac{\partial \bar{p}}{\partial z} = 0, \\ \dfrac{d\bar{p}}{dt} + a_0^2 \dfrac{\partial w}{\partial z} = 0. \end{cases}$$

Оператор решения, как и разностную схему, на этом шаге обозначим через Λ_{312}. Разностный оператор Λ_{12} имеет вид (4.3), а оператор Λ_3 запишем в виде

$$u^{j,k,i} = u^{j,k,i}, \quad v^{j,k,i} = v^{j,k,i},$$

$$w^{j,k,i} = w^{j,k,i} - \tau(\bar{P}_{j,k,i+1/2} - \bar{P}_{j,k,i-1/2})/h_i,$$

$$p^{j,k,i} = p_{j,k,i} - \tau a_0^2(W_{j,k,i+1/2} - W_{j,k,i-1/2})/h_i.$$

Здесь

$$\bar{p}_{j,k,i} = p_{j,k,i} - \frac{0.5(\bar{\tau}_k + \tau)a_0^2}{h_j}(U_{j+1/2,k,i} - U_{j-1/2,k,i}) - \frac{0.5(\bar{\tau}_j + \tau)a_0^2}{h_k}(V_{j,k+1/2,i} - V_{j,k-1/2,i}).$$

Давления $\bar{P}_{j,k,i+1/2}$, ... вычисляются из решения одномерной задачи о плоском распаде разрыва по величинам, которые были получены на первом шаге из решения по схеме Λ_{12}. Исключив дробный шаг, получим следующую схему Λ_{312}:

$$u^{j,k,i} = u_{j,k,i} - \frac{\tau}{h_j}(P_{j+1/2,k,i} - P_{j-1/2,k,i}) + \frac{\tau}{2h_j}\frac{\bar{\tau}_k a_0^2}{2h_k}[(V_{j+1,k+1/2,i} - V_{j+1,k-1/2,i}) - (V_{j-1,k+1/2,i} - V_{j-1,k-1/2,i})],$$

$$v^{j,k,i} = v_{j,k,i} - \frac{\tau}{h_k}(P_{j,k+1/2,i} - P_{j,k-1/2,i}) +$$

$$+ \frac{\tau}{2h_k}\frac{\bar{\tau}_j a_0^2}{2h_j}[(U_{j+1/2,k+1,i} - U_{j-1/2,k+1,i}) - (U_{j+1/2,k-1,i} - U_{j-1/2,k-1,i})],$$

$$w^{j,k,i} = w^{j,k,i} - \frac{\tau}{h_i}(P_{j,k,i+1/2} - P_{j,k,i-1/2}) +$$

$$+ \frac{\tau}{h_i}\frac{0.5(\bar{\tau}_k + \tau)a_0^2}{2h_j}[(U_{j+1/2,k,i+1} - U_{j-1/2,k,i+1}) - (U_{j+1/2,k,i-1} - U_{j-1/2,k,i-1})] -$$

$$- \frac{\tau}{h_i}\frac{0.5(\bar{\tau}_j + \tau)a_0^2}{2h_k}[(V_{j,k+1/2,i+1} - V_{j,k-1/2,i+1}) - (V_{j,k+1/2,i-1} - V_{j,k-1/2,i-1})],$$

$$p^{j,k,i} = p_{j,k,i} - \frac{0.5(\bar{\tau}_k + \tau)a_0^2}{h_j}(U_{j+1/2,k,i} - U_{j-1/2,k,i}) -$$

$$- \frac{0.5(\bar{\tau}_j + \tau)a_0^2}{h_k}(V_{j,k+1/2,i} - V_{j,k-1/2,i}) - \frac{\tau a_0^2}{h_i}(W_{j,k,i+1/2} - W_{j,k,i-1/2}).$$

Если на первом шаге решается вначале вторая система уравнений Λ_3, а затем первая система Λ_{12} по схеме (4.3), то разностная схема Λ_{123} после исключения дробных шагов будет иметь вид

$$u^{j,k,i} = u_{j,k,i} - \frac{\tau}{h_j}(P_{j+1/2,k,i} - P_{j-1/2,k,i}) +$$

$$+ \frac{\tau}{2h_j}\frac{\bar{\tau}_k a_0^2}{2h_k}[(V_{j+1,k+1/2,i} - V_{j+1,k-1/2,i}) - (V_{j-1,k+1/2,i} - V_{j-1,k-1/2,i})] +$$

$$+ \frac{\tau}{h_k}\frac{\bar{\tau}_j a_0^2}{2h_j}[(W_{j+1,k,i+1/2} - W_{j+1,k,i-1/2}) - (W_{j-1,k,i+1/2} - W_{j-1,k,i-1/2})],$$

$$v^{j,k,i} = v_{j,k,i} - \frac{\tau}{h_k}(P_{j,k+1/2,i} - P_{j,k-1/2,i}) +$$

$$+ \frac{\tau}{2h_k}\frac{\bar{\tau}_j a_0^2}{2h_j}[(U_{j+1/2,k+1,i} - U_{j-1/2,k+1,i}) - (U_{j+1/2,k-1,i} - U_{j-1/2,k-1,i})] +$$

$$+ \frac{\tau}{h_k}\frac{\bar{\tau}_j a_0^2}{2h_j}[(W_{j,k+1,i+1/2} - W_{j,k+1,i-1/2}) - (W_{j,k-1,i+1/2} - W_{j,k-1,i-1/2})],$$

$$w^{j,k,i} = w_{j,k,i} - \frac{\tau}{h_i}(P_{j,k,i+1/2} - P_{j,k,i-1/2}),$$

$$p^{j,k,i} = p_{j,k,i} - \frac{0.5(\bar{\tau}_k + \tau)a_0^2}{h_j}(U_{j+1/2,k,i} - U_{j-1/2,k,i}) -$$

$$- \frac{0.5(\bar{\tau}_j + \tau)a_0^2}{h_k}(V_{j,k+1/2,i} - V_{j,k-1/2,i}) - \frac{\tau a_0^2}{h_i}(W_{j,k,i+1/2} - W_{j,k,i-1/2}).$$

Взяв полусумму операторов Λ_{123}, Λ_{312}, получим симметричную разностную схему в виде

$$u^{j,k,i} = u_{j,k,i} - \frac{\tau}{h_j}(P_{j+1/2,k,i} - P_{j-1/2,k,i}) +$$

$$+ \frac{\tau}{2h_j}\frac{\bar{\tau}_k a_0^2}{2h_k}[(V_{j+1,k+1/2,i} - V_{j+1,k-1/2,i}) - (V_{j-1,k+1/2,i} - V_{j-1,k-1/2,i})] +$$

$$+ \frac{\tau}{2h_k}\frac{\bar{\tau}_j a_0^2}{2h_j}[(W_{j+1,k,i+1/2} - W_{j+1,k,i-1/2}) - (W_{j-1,k,i+1/2} - W_{j-1,k,i-1/2})],$$

$$v^{j,k,i} = v_{j,k,i} - \frac{\tau}{h_k}(P_{j,k+1/2,i} - P_{j,k-1/2,i}) +$$

$$+ \frac{\tau}{2h_k}\frac{\bar{\tau}_j a_0^2}{2h_j}[(U_{j+1/2,k+1,i} - U_{j-1/2,k+1,i}) - (U_{j+1/2,k-1,i} - U_{j-1/2,k-1,i})] +$$

$$+ \frac{\tau}{2h_k}\frac{\bar{\tau}_j a_0^2}{2h_j}[(W_{j,k+1,i+1/2} - W_{j,k+1,i-1/2}) - (W_{j,k-1,i+1/2} - W_{j,k-1,i-1/2})],$$

$$w^{j,k,i} = w^{j,k,i} - \frac{\tau}{h_i}(P_{j,k,i+1/2} - P_{j,k,i-1/2}) +$$

$$+ \frac{\tau}{2h_k}\frac{\bar{\tau}_j a_0^2}{2h_j}[(U_{j+1/2,k,i+1} - U_{j-1/2,k,i+1}) - (U_{j+1/2,k,i-1} - U_{j-1/2,k,i-1})] +$$

$$+ \frac{\tau}{2h_k}\frac{\bar{\tau}_j a_0^2}{2h_j}[(V_{j,k+1/2,i+1} - V_{j,k-1/2,i+1}) - (V_{j,k+1/2,i-1} - V_{j,k-1/2,i-1})],$$

$$p^{j,k,i} = p_{j,k,i} - \frac{0.5(\bar{\tau}_j + \tau)a_0^2}{h_j}(U_{j+1/2,k,i} - U_{j-1/2,k,i}) -$$

$$- \frac{0.5(\bar{\tau}_k + \tau)a_0^2}{h_k}(V_{j,k+1/2,i} - V_{j,k-1/2,i}) - \frac{0.5(\bar{\tau}_i + \tau)a_0^2}{h_i}(W_{j,k,i+1/2} - W_{j,k,i-1/2}),$$

которую можно записать в форме схемы предиктор-корректор:

$$\begin{aligned} u^{j,k,i} &= u_{j,k,i} - \frac{\tau}{h_j}(P^*_{j+1/2,k,i} - P^*_{j-1/2,k,i}), \\ v^{j,k,i} &= v_{j,k,i} - \frac{\tau}{h_k}(P^*_{j,k+1/2,i} - P^*_{j,k-1/2,i}), \\ w^{j,k,i} &= w_{j,k,i} - \frac{\tau}{h_i}(P^*_{j,k,i+1/2} - P^*_{j,k,i-1/2}), \end{aligned} \tag{4.7}$$

$$p^{j,k,i} = p_{j,k,i} - \frac{\tau a_0^2}{h_j}(U_{j+1/2,k,i} - U_{j-1/2,k,i}) - \frac{\tau a_0^2}{h_k}(V_{j,k+1/2,i} - V_{j,k-1/2,i}) - \frac{\tau a_0^2}{h_i}(W_{j,k,i+1/2} - W_{j,k,i-1/2}).$$

Здесь $P^*_{j,k+1/2,i}$, ... – это новые большие величины, полученные путем реконструкции больших величин схемы Годунова по формулам

$$\begin{aligned} P^*_{j+1/2,k,i} &= P_{j+1/2,k,i} - \frac{0.5\bar{\tau}a_0^2}{h_j}[(V_{j+1,k+1/2,i} - V_{j+1,k-1/2,i}) + (V_{j,k+1/2,i} - V_{j,k-1/2,i})] - \\ &- \frac{0.5\tilde{\tau}a_0^2}{h_i}[(W_{j+1,k,i+1/2} - W_{j+1,k,i-1/2}) + (W_{j,k,i+1/2} - W_{j,k,i-1/2})], \\ P^*_{j,k+1/2,i} &= P_{j,k+1/2,i} - \frac{0.5\bar{\tau}a_0^2}{h_j}[(U_{j+1/2,k+1,i} - U_{j-1/2,k+1,i}) + (U_{j+1/2,k,i} - U_{j-1/2,k,i})] - \\ &- \frac{0.5\tilde{\tau}a_0^2}{h_i}[(W_{j,k+1,i+1/2} - W_{j,k+1,i-1/2}) + (W_{j,k,i+1/2} - W_{j,k,i-1/2})], \\ P^*_{j,k,i+1/2} &= P_{j,k,i+1/2} - \frac{0.5\bar{\tau}a_0^2}{h_j}[(U_{j+1/2,k,i+1} - U_{j-1/2,k,i+1}) + (U_{j+1/2,k,i} - U_{j-1/2,k,i})] - \\ &- \frac{0.5\tilde{\tau}a_0^2}{h_k}[(V_{j,k+1/2,i+1} - V_{j,k-1/2,i+1}) + (V_{j,k+1/2,i} - V_{j,k-1/2,i})]. \end{aligned} \tag{4.8}$$

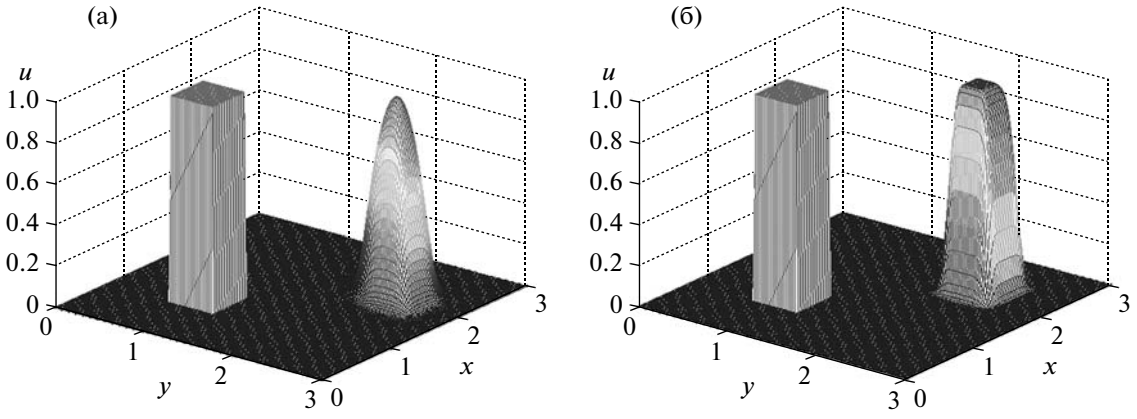
Отсюда следует, что новые большие величины давлений вычисляются из разностных уравнений, которые аппроксимируют уравнение для давления с тремя пространственными переменными в (3.8). ДП схемы (4.7) будет иметь вид

$$\begin{aligned} \frac{du}{dt} + \frac{\partial p}{\partial x} &= \left(\tau_j^* - \frac{\tau}{2}\right)a_0^2\frac{\partial^2 u}{\partial x^2}, \\ \frac{dv}{dt} + \frac{\partial p}{\partial y} &= \left(\tau_k^* - \frac{\tau}{2}\right)a_0^2\frac{\partial^2 v}{\partial y^2}, \\ \frac{dw}{dt} + \frac{\partial p}{\partial z} &= (\tau_i^* - 0.5\tau)a_0^2\frac{\partial^2 w}{\partial z^2}, \end{aligned} \tag{4.9}$$

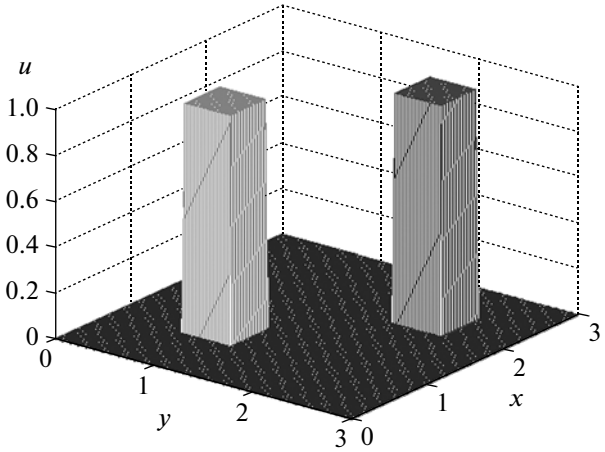
$$\frac{dp}{dt} + a_0^2\frac{\partial u}{\partial x} + a_0^2\frac{\partial v}{\partial y} + a_0^2\frac{\partial w}{\partial z} = (\tau_j^* - 0.5\tau)a_0^2\frac{\partial^2 p}{\partial x^2} + (\tau_k^* - 0.5\tau)a_0^2\frac{\partial^2 p}{\partial y^2} + (\tau_i^* - 0.5\tau)a_0^2\frac{\partial^2 p}{\partial z^2}.$$

Здесь $\tau_j^* = 0.5(\tau_{j+1/2,k,i} + \tau_{j-1/2,k,i})$, $\tau_k^* = 0.5(\tau_{j,k+1/2,i} + \tau_{j,k-1/2,i})$, $\tau_i^* = 0.5(\tau_{j,k,i+1/2} + \tau_{j,k,i-1/2})$. Если выписать выражение для изменения нормы финитных функций, то определитель Сильверста квадратичных форм будет равен

$$D = (\tau_{j+1/2,k,i} - 0.5\tau)(\tau_{j,k+1/2,i} - 0.5\tau)(\tau_{j,k,i+1/2} - 0.5\tau).$$



Фиг. 1.



Фиг. 2.

Отсюда следует условие выбора шага интегрирования по времени

$$\tau \le \min(\tau_x, \tau_y, \tau_z),$$

при котором норма не возрастает. Видно, что шаг интегрирования по времени в схеме расщепления на равномерной по пространству сетке (кубической) в три раза больше, чем шаг по времени в классической схеме без расщепления. Кроме того, если числа Куранта равны единице, то правая часть в (4.9) обращается в ноль. Следовательно, условие сдвига в ДП (4.9) выполняется. С учетом уравнений (3.3), из которых находятся большие величины давлений, уравнения (4.8) в общем случае примут следующий вид:

$$P^*_{j+1/2,k,i} = P_{j+1/2,k,i} - \frac{\tau}{2h_j}[(1-\alpha)a^2_{j+1,k,i}(V_{j+1,k+1/2,i} - V_{j+1,k-1/2,i}) + \alpha a^2_{j,k,i}(V_{j,k+1/2,i} - V_{j,k-1/2,i})] -$$

$$- \frac{\tau}{2h_i}[(1-\alpha)a^2_{j+1,k,i}(W_{j+1,k,i+1/2} - W_{j+1,k,i-1/2}) + \alpha a^2_{j,k,i}(W_{j,k,i+1/2} - V_{j,k,i-1/2})].$$

Здесь $\alpha = a_{j,k,i}/(a_{j+1,k,i} + a_{j,k,i})$. Большие величины давлений по остальным направлениям реконструируются по аналогичным формулам.

5. ЗАДАЧА О ДВИЖЕНИИ ПРЯМОУГОЛЬНОГО ПРОФИЛЯ

Пусть в плоскости (x, y) в области $\Omega = [0 \le x \le 3, 0 \le y \le 3]$ задана функция $u(x, y)$ в виде плоской квадратной “полки”. Значения функции в квадрате со стороной $l = 0.5$ равны единице, а в

остальных точках области значения равны нулю. Левый нижний угол квадрата имеет координаты $x = y = 1$. Полка движется в области со скоростями $V_x = 1$, $V_y = 1$ вдоль координатных осей x, y соответственно. Требуется рассчитать движение полки при переходе из левой нижней части области в правую верхнюю часть.

Задача считалась на сетке 150×150 точек вдоль осей x, y соответственно с числом Куранта, равным 0.9. Результаты расчетов по схемам (2.2) и (2.7) приведены на фиг. 1 и 2.

На фиг. 1 представлены результаты расчетов движения полки по схеме Годунова (фиг. (а)) и модифицированной схеме (2.7) (фиг. (б)).

На фиг. 2 представлены результаты расчетов движения полки по модифицированной схеме (2.7) с числом Куранта, равным единице.

Из представленных результатов расчетов на фиг. 1, 2 видно, что результаты расчетов по модифицированной схеме (2.7) согласуются с результатами расчетов по схеме расщепления по пространственным переменным из [12], точность выше, чем в схеме (2.2), и условие сдвига при числах Куранта, равных единице, выполняется.

ЗАКЛЮЧЕНИЕ

Построены симметричные явные схемы покомпонентного расщепления, которые преобразуются к эквивалентным схемам предиктор-корректор. Предложена реконструкция больших величин, позволяющая устранить отмеченный источник ошибок аппроксимации в схемах Годунова. Реконструкция осуществляется на шаге предиктора после вычисления больших величин методом Годунова или по неявным одномерным схемам. Выбор шага интегрирования по времени в модифицированных явных схемах согласован с выбором в одномерных схемах и на равномерных по пространству разностных сетках (квадратных и кубических) в 2 и 3 раза, соответственно, больше, чем в классических схемах Годунова.

СПИСОК ЛИТЕРАТУРЫ

1. *Годунов С.К.* Разностный метод численного расчета разрывных решений уравнений гидродинамики // Матем. сб. 1959. № 47. Вып. 3. С. 271–306.
2. *Годунов С.К., Забродин А.В., Иванов М.Я. и др.* Численное решение многомерных задач газовой динамики. М.: Наука, 1976.
3. *Куликовский А.Г., Погорелов Н.В., Семенов А.Ю.* Математические вопросы численного решения гиперболических систем уравнений. М.: Физматлит, 2001.
4. *Годунов С.К., Забродин А.В., Прокопов Г.П.* Разностная схема для двумерных нестационарных задач газовой динамики и расчет обтекания с отошедшей ударной волной // Ж. вычисл. матем. и матем. физ. 1961. Т. 1. № 6. С. 1020–1050.
5. *Рождественский Б.Л., Яненко Н.Н.* Системы квазилинейных уравнений и их применение к газовой динамике. М.: Наука, 1978.
6. *Abgrall R.* Approximation du probleme de Riemann vraiment multidimensionneles equations d'Euler par une methode de type Roe, I: La linearization // C.r. Acad. Sci. Ser. I. 1994. V. 319. P. 499–504. II: Solution du pro,leme de Riemann fpproche // P. 625–629.
7. *LeVeque R.J.* Wave propagation algorithms for multidimensional hyperbolic systems // J. Comput. Phys. 1997. V. 131. P. 327.
8. *Gilquin H., Laurens J., Roiser C.* Multi-dimensional Riemann problems for linear hyperbolic systems // Notes Number. Fluid Mech. 1993. V. 43. P. 284.
9. *Brio M., Zakharian A.R., Webb G.M.* Two-dimensional Riemann solver for Euler equations of gas dynamics // J. Comput. Phys. 2001. V. 167. P. 177–195.
10. *Васильев Е.И.* W-модификация метода С.К. Годунова и ее применение для двумерных нестационарных течений запыленного газа // Ж. вычисл. матем. и матем. физ. 1996. Т. 36. № 1. С. 122–135.
11. *Colella P.* Multidimensional upwind methods for hyperbolic conservation laws // J. Comput. Phys. 1990. V. 87. P. 171–200.
12. *Моисеев Н.Я., Силантьева И.Ю.* Разностные схемы произвольного порядка аппроксимации для решения линейных уравнений переноса с постоянными коэффициентами методом Годунова с антидиффузией // Ж. вычисл. матем. и матем. физ. 2008. Т. 48. № 7. С. 1282–1293.
13. *Peaceman D.W., Rachford H.H., Jr.* The numerical solution of parabolic and elliptic differential equations // J. Soc. Industr. Appl. Math. 1955. V. 3. № 1. P. 28–42.
14. *Douglas J., Jr.* On the numerical integration of $u_{xx} + u_{yy} = u_t$ by implicit methods // J. Soc. Industr. Appl. Math. 1955. V. 3. № 1. P. 42–65.

15. *Багриновский К.А., Годунов С.К.* Разностные схемы для многомерных задач // Докл. АН СССР. М.: 1957. Т. 115. С. 431–433.
16. *Яненко Н.Н.* Метод дробных шагов решения многомерных задач математической физики. Новосибирск: Наука, 1967.
17. *Марчук Г.И.* Методы расщепления. М.: Наука, 1988.
18. *Ковеня В.М., Тарнавский Г.А., Черный С.Г.* Применение метода расщепления в задачах аэродинамики. Новосибирск: Наука, СО, 1990.
19. *Годунов С.К.* Воспоминания о разностных схемах. Новосибирск: Научн. книга, 1997.
20. *Годунов С.К., Забродин А.В.* О разностных схемах второго порядка точности для многомерных задач // Ж. вычисл. матем. и матем. физ. 1962. Т. 2. № 4. С. 706–708.
21. *Самарский А.А.* О принципе аддитивности для построения экономичных разностных схем // Докл. АН СССР. 1965. Т. 165. № 6.
22. *Шокин Ю.И.* Метод дифференциального приближения. Новосибирск: Наука, 1979.
23. *Дьяконов В.П.* Maple 8 в математике, физике и образовании. М.: СОЛОН-Пресс, 2003.
24. *Моисеев Н.Я.* Об одной модификации разностной схемы Годунова // ВАНТ. Сер. Методики и программы числ. решения задач матем. физ. 1986. Вып. 3. С. 35–43.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2009, том 49, № 11, с. 1988–2000

УДК 519.634

ЧИСЛЕННЫЙ МЕТОД НАХОЖДЕНИЯ 3D-СОЛИТОНОВ НЕЛИНЕЙНОГО УРАВНЕНИЯ ШРЁДИНГЕРА В АКСИАЛЬНО-СИММЕТРИЧНОМ СЛУЧАЕ[1)]

© 2009 г. О. В. Матусевич, В. А. Трофимов

(119992 Москва, Ленинские горы, МГУ ВМиК)

e-mail: vatro@cs.msu.su

Поступила в редакцию 06.03.2009 г.

Предложен численный метод нахождения солитонов, путем формулировки задачи как задачи на собственные значения (СЗ) и собственные функции (СФ) для системы двух нелинейных уравнений Шрёдингера, описывающих процесс удвоения частоты фемтосекундных импульсов в аксиально-симметричной среде в случае с квадратичной и кубичной нелинейностью. Рассматривается также практически важный частный случай одного уравнения Шрёдингера. Так как трехмерные солитоны для случая кубичной нелинейности неустойчивы к малым возмущениям своей формы, то предложен метод их стабилизации за счет слабой модуляции коэффициента кубичной нелинейности, а также варьирования длины фокусирующих слоев. Подчеркнем, что ранее в литературе для стабилизации предлагалась либо среда с чередующимися по знаку нелинейности слоями, либо среда с сильно изменяющимися по величине (но одного знака) нелинейными слоями. Показано, что применение слабой модуляции в рассмотренном нами случае позволяет увеличить более чем в 4 раза длину среды без коллапса световой волны. Для нахождения СФ и СЗ нелинейной задачи построен итерационный процесс, который позволяет эффективно решать задачи поиска трехмерных солитонов на больших сетках. Библ. 56. Фиг. 4. Табл. 1.

Ключевые слова: нелинейные уравнения Шрёдингера, трехмерные солитоны, численный метод вычисления собственных значений и собственных функций, итерационный процесс.

ВВЕДЕНИЕ

Распространение лазерных фемтосекундных импульсов в различных нелинейных средах широко исследуется в литературе в последние годы. Одним из важных аспектов этой проблемы является нахождение солитонных режимов распространения волн, чему посвящено много работ (см., например, [1]–[21]). Солитонные решения нелинейных уравнений, активно исследуемые в литературе, как известно, представляют интерес для различных разделов физики.

Среди предложенных методов нахождения солитонов наибольшее распространение, на наш взгляд, получили метод обратной задачи, гамильтонов подход, спектральные методы и другие методы, которые изложены, например, в [22]–[38]. В оптике эти методы также нашли широкое применение (см. [39], [40]). При этом лазерное излучение позволяет реализовать так называемые цветные солитоны, когда на нескольких частотах одновременно существуют и распространяются вместе оптические волны вдоль нелинейной среды. Эволюция этих солитонов описывается системами связанных уравнений Шрёдингера. С момента их предсказания в среде с квадратичной нелинейностью (см. [41]) интерес к этим солитонам в литературе постоянно сохраняется в связи с многочисленными потенциальными приложениями их в задачах передачи информации оптическими методами. Выполненные экспериментальные исследования (см. [1]–[7]), по их наблюдению в различных лабораториях, показали практическую реализацию солитонов данного типа.

Важно, однако, подчеркнуть, что наиболее полно изучены одномерные солитоны. Проблема нахождения многомерных солитонов, в особенности для систем нелинейных уравнений Шрёдингера остается актуальной. Именно она рассматривается ниже. Для полноты анализа также обсуждаются и солитоны, формируемые в кубично-нелинейной среде, которые представляют

[1)] Работа выполнена при частичной финансовой поддержке РФФИ (код проекта 08-01-00107-а).

самостоятельный интерес в связи с широким проявлением явления самофокусировки, например при распространении фемтосекундных импульсов в атмосфере, лазерных системах и т.д.

Как хорошо известно из литературы, распространение 3D-солитонов в пространстве при наличии аксиальной симметрии пучка обладает неустойчивостью. Различные способы ее стабилизации предлагались в [8], [9], [39], [40]. Так, в [8], [9] исследовалась возможность стабилизации неустойчивых (2 + 1)-D-солитонов с помощью сильной модуляции коэффициента кубичной нелинейности (либо даже смены его знаков) в направлении распространения волны. В отличие от этих работ, нами рассматривается влияние слабой модуляции кубичной нелинейности и варьирования ширины фокусирующих слоев на эволюцию солитона с целью его стабилизации. Важно подчеркнуть, что слабая модуляция предпочтительна, так как она не приводит к появлению отраженной волны и делает описание распространения оптического излучения в рамках используемого уравнения Шрёдингера корректным. Как показало компьютерное моделирование, за счет специального подбора параметров модуляции коэффициента кубичной нелинейности удается стабилизировать солитон на большом расстоянии.

Заметим, что, как правило, большинство известных солитонов в нелинейной оптике найдено аналитически (см. [39], [40]). Тем не менее, построение солитонных решений нелинейного уравнения Шрёдингера (или систем уравнений) на основе компьютерного моделирования (численного решения соответствующего уравнения) широко обсуждается в этих же источниках. Один из способов состоит, в частности, в применении методов нахождения СФ и СЗ нелинейного уравнения Шрёдингера (или системы таких уравнений). Однако, в отличие от двумерных задач, рассматриваемых в [42]–[44], нахождение 3D-солитонов требует большого объема вычислений и затрат машинного времени. Это связано с необходимостью производить поиск солитонов на сетках, имеющих более миллиона узлов. Как будет показано ниже, использование подхода, описанного в [44], для расчета солитонов делается невозможным из-за кубичной сложности вычислений. Поэтому актуальной является задача нахождения способов ускорения вычислений.

В заключение подчеркнем, что ниже рассматриваются солитоны в физическом смысле.

1. ПОСТАНОВКА ЗАДАЧИ

Как известно, распространение оптического излучения в среде с кубичной нелинейностью в аксиально-симметричном случае описывается следующей системой нелинейных уравнений Шрёдингера:

$$\begin{gathered}\frac{\partial A_1}{\partial z} + i\tilde{D}\Delta_r A_1 + iD_1\frac{\partial^2 A_1}{\partial t^2} + i\gamma A_1^* A_2 e^{-i\Delta kz} + i\alpha_1 A_1(|A_1|^2 + 2|A_2|^2) = 0,\\ \frac{\partial A_2}{\partial z} + i\frac{\tilde{D}}{2}\Delta_r A_2 + \nu\frac{\partial A_2}{\partial t} + iD_2\frac{\partial^2 A_2}{\partial t^2} + i\gamma A_1^2 e^{i\Delta kz} + i\alpha_2 A_2(2|A_1|^2 + |A_2|^2) = 0,\\ 0 < r < L_r, \quad 0 < t < L_t, \quad 0 < z \le L_z, \quad \alpha_2 = 2\alpha_1 = 2\alpha.\end{gathered} \tag{1}$$

Здесь t – безразмерное время в сопровождающей импульс основной волны системе координат, z – нормированная на дифракционную длину пучка основной волны продольная координата; $l_д = 2k_1a^2$; k_1 – волновое число; a – начальный радиус пучка; $\Delta_r = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial}{\partial r}\right)$ – оператор Лапласа по координате r, измеряемой в единицах a; $D_j \sim -0.5\frac{\partial^2 k_j}{\partial \omega_j^2}$ – коэффициенты, характеризующие дисперсию второго порядка; $\tilde{D}$ – коэффициент, характеризующий дифракцию. В выбранной нормировке он равен 1, но оставлен в (1) для удобства моделирования; k_j, ω_j – соответственно, размерное волновое число и частота j-й волны, $j = 1, 2$; γ – коэффициент нелинейной связи взаимодействующих волн; $\Delta k = k_2 - 2k_1$ – безразмерная расстройка их волновых чисел; α_j – коэффициенты самовоздействия волн; A_j – комплексные амплитуды гармоник, нормированные на квадратный корень из максимальной интенсивности первой гармоники на входе среды ($z = 0$). Параметр ν пропорционален разности обратных величин групповых скоростей волн второй гармоники и основной частоты. Далее рассматривается случай равенства нулю параметра ν ($\nu = 0$), L_z –

безразмерная длина нелинейной среды, L_r – ее поперечный размер, L_t – безразмерное время, в течение которого анализируется рассматриваемый процесс.

На входе в нелинейную среду задается начальное распределение импульса:

$$A_j(t, r, z = 0) = A_{j0}(t, r), \quad j = 1, 2, \quad 0 \le t \le L_t, \quad 0 \le r \le L_r. \tag{2}$$

Граничные условия для уравнений системы (1) имеют вид

$$A_j\big|_{t = 0,\, L_t} = 0, \quad r\frac{\partial A_j}{\partial r}\bigg|_{r \to 0} = 0, \quad A_j\big|_{r = L_r} = 0; \tag{3}$$

их можно поставить из-за финитности начального распределения и конечного отрезка по координате z.

Для полноты анализа ниже рассматривается также практически важный частный случай двухчастотного взаимодействия, когда амплитуда второй гармоники тождественно равна нулю: $A_2 \equiv 0$. Распространение фемтосекундного импульса в среде с кубичной нелинейностью описывается в этом случае следующим безразмерным нелинейным уравнением Шрёдингера (НУШ) для комплексной амплитуды $A = A_1$:

$$\frac{\partial A}{\partial z} + i\tilde{D}\Delta_r A + iD\frac{\partial^2 A}{\partial t^2} + i\alpha A|A|^2 = 0. \tag{4}$$

Для этого уравнения задается начальное условие вида (2) и граничные условия (3).

Для нахождения СФ уравнений (1) представим решение в виде

$$A_1 = u(t, r)e^{-i\lambda z}, \quad A_2 = v(t, r)e^{-i\mu z}.$$

Подставляя эти функции в исходные уравнения (1), получаем следующую задачу на СЗ:

$$\begin{gathered}\tilde{D}\Delta_r u + D_1\frac{\partial^2 u}{dt^2} + \gamma u^* v e^{i(2\lambda - \mu - \Delta k)z} + \alpha u(|u|^2 + 2|v|^2) = \lambda u, \quad 0 < r < L_r, \quad 0 < t < L_t, \\ \frac{\tilde{D}}{2}\Delta_r v + D_2\frac{\partial^2 v}{dt^2} + \gamma u^2 e^{-i(2\lambda - \mu - \Delta k)z} + 2\alpha v(2|u|^2 + |v|^2) = \mu v\end{gathered} \tag{5}$$

с граничными условиями вида

$$\begin{gathered}u(0, r) = u(L_t, r) = v(0, r) = v(L_t, r) = 0, \quad 0 < r < L_r, \\ r\frac{\partial u}{\partial r}\bigg|_{r \to 0} = r\frac{\partial v}{\partial r}\bigg|_{r \to 0} = 0, \quad u(t, L_r) = v(t, L_r) = 0, \quad 0 < t < L_t.\end{gathered} \tag{6}$$

Положим $\mu = 2\lambda - \Delta k$, чтобы избавиться в уравнениях (5) от зависимости коэффициента при квадратичной нелинейности от координаты z. Тогда система (5) запишется в виде

$$\begin{gathered}\tilde{D}\Delta_r u + D_1\frac{\partial^2 u}{\partial t^2} + \gamma u^* v + \alpha u(|u|^2 + 2|v|^2) = \lambda u, \\ \frac{\tilde{D}}{2}\Delta_r v + D_2\frac{\partial^2 v}{\partial t^2} + \gamma u^2 + 2\alpha v(2|u|^2 + |v|^2) + \Delta k v = 2\lambda v.\end{gathered} \tag{7}$$

Уравнения (7) имеют только вещественные СЗ. Действительно, умножим первое уравнение системы на u^*, а второе – на v^* и сложим их. Далее, проинтегрируем полученное тождество от 0

до L_t и от 0 до L_r. Воспользовавшись интегрированием по частям и граничными условиями, получим следующее тождество:

$$\int_0^{L_r}\int_0^{L_t}\left(-\tilde{D}\left|\frac{\partial u}{\partial r}\right|^2-\frac{\tilde{D}}{2}\left|\frac{\partial v}{\partial r}\right|^2-D_1\left|\frac{\partial u}{\partial t}\right|^2-D_2\left|\frac{\partial v}{\partial t}\right|^2+2\gamma\mathrm{Re}(u^2v^*)\right)rdrdt+$$

$$+\int_0^{L_r}\int_0^{L_t}(\alpha|u|^2(|u|^2+2|v|^2)+2\alpha|v|^2(2|u|^2+|v|^2)+\Delta k|v|^2)rdrdt=\lambda\int_0^{L_r}\int_0^{L_t}(|u|^2+2|v|^2)rdrdt.$$

Учитывая, что стоящие в правой и левой частях интегралы вещественны, получаем, что λ вещественное. Следовательно, уравнения (7) с вещественными коэффициентами. Поэтому в дальнейшем нас будут интересовать только действительные решения и, значит, знак модуля в уравнении (7) можно опустить.

2. РАЗНОСТНАЯ СХЕМА

Для решения задачи (7) с граничными условиями (6) введем, например, равномерную сетку $\omega=\omega_t\times\omega_r=\{0\le t\le L_t\}\times\{0\le r\le L_r\}$:

$$\omega_t=\{t_j=j\tau, j=\overline{0,N_t}, L_t=\tau N_t\},\quad \omega_r=\{r_k=(k+0.5)h_r, k=\overline{0,N_r}, h_r=L_r/(N_r+0.5)\}.$$

Определим сеточные функции u_h, v_h на ω: $u_{j,k}=u(t_j,r_k)$, $v_{j,k}=v(t_j,r_k)$, и разностный оператор Лапласа во внутренних узлах сетки:

$$\Lambda_\perp\varphi=\frac{1}{r_k(r_{k+1}-r_{k-1})}\left[(r_{k+1}+r_k)\frac{\varphi_{j,k+1}-\varphi_{j,k}}{r_{k+1}-r_k}-(r_k+r_{k-1})\frac{\varphi_{j,k}-\varphi_{j,k-1}}{r_k-r_{k-1}}\right],$$

$$\varphi_{\bar{t}t,j}=\frac{\varphi_{j+1,k}-2\varphi_{j,k}+\varphi_{j-1,k}}{\tau^2},$$

где φ – одна из функций u, v. Тогда разностная схема для уравнений (7) запишется в виде

$$\begin{gathered}\tilde{D}\Delta_\perp u+D_1u_{\bar{t}t,j}+\gamma u_{j,k}v_{j,k}+\alpha u_{j,k}(u_{j,k}^2+2v_{j,k}^2)=\lambda u_{j,k},\quad j=\overline{1,N_t-1},\quad k=\overline{1,N_r-1},\\ \frac{\tilde{D}}{4}\Delta_\perp v+\frac{D_2}{2}v_{\bar{t}t,j}+\frac{\gamma}{2}u_{j,k}^2+\alpha v_{j,k}(2u_{j,k}^2+v_{j,k}^2)+\frac{\Delta k}{2}v_{j,k}=\lambda v_{j,k}.\end{gathered}\tag{8}$$

Разностные уравнения (8) необходимо дополнить следующими условиями в граничных точках:

$$\tilde{D}\frac{u_{j,1}-u_{j,0}}{0.5h_r^2}+D_1u_{\bar{t}t,j}+\gamma u_{j,0}v_{j,0}+\alpha u_{j,0}(u_{j,0}^2+2v_{j,0}^2)=\lambda u_{j,0},$$

$$\frac{\tilde{D}}{4}\frac{v_{j,1}-v_{j,0}}{0.5h_r^2}+\frac{D_2}{2}v_{\bar{t}t,j}+\frac{\gamma}{2}u_{j,0}^2+\alpha v_{j,0}(2u_{j,0}^2+v_{j,0}^2)+\frac{\Delta k}{2}v_{j,0}=\lambda v_{j,0},\tag{9}$$

$$u_{j,N_r}=v_{j,N_r}=0,\quad u_{0,k}=u_{N_r,k}=v_{0,k}=v_{N_r,k}=0,\quad j=\overline{1,N_t-1},\quad k=\overline{1,N_r-1}.$$

Первые два из них аппроксимируют со вторым порядком условия на производную от функций по координате r в нуле (см. [45]). Разностная схема (8) с граничными условиями (9) аппроксимирует уравнения (7) в области $[0,L_t]\times(0,L_r]$ в норме C с порядком $O(\tau^2+h_r^2/r)$.

Так как записанные уравнения (8), (9) нелинейны, то для их разрешения запишем итерационный процесс в виде

$$\begin{gathered}
\tilde{D}\Delta_\perp \overset{s+1}{u} + D_1 \overset{s+1}{u}_{\bar{t}t,j} + \gamma \overset{s}{u}_{j,k} \overset{s+1}{v}_{j,k} + \alpha \overset{s+1}{u}_{j,k}(\overset{s^2}{u}_{j,k} + 2\overset{s^2}{v}_{j,k}) = \overset{s+1\,s+1}{\lambda}\, u_{j,k}, \quad j = \overline{1, N_t - 1}, \quad k = \overline{1, N_r - 1}, \\
\frac{\tilde{D}}{4}\Delta_\perp \overset{s+1}{v} + \frac{D_2}{2}\overset{s+1}{v}_{\bar{t}t,j} + \frac{\gamma}{2}\overset{s}{u}_{j,k}\overset{s+1}{u}_{j,k} + \alpha \overset{s+1}{v}_{j,0}(2\overset{s^2}{u}_{j,k} + \overset{s^2}{v}_{j,k}) + \frac{\Delta k}{2}\overset{s+1}{v}_{j,k} = \overset{s+1\,s+1}{\lambda}\, v_{j,k}, \\
\tilde{D}\frac{\overset{s+1}{u}_{j,1} - \overset{s+1}{u}_{j,0}}{0.5h_r^2} + D_1 \overset{s+1}{u}_{\bar{t}t,j} + \gamma \overset{s}{u}_{j,0}\overset{s+1}{v}_{j,0} + \alpha \overset{s+1}{u}_{j,0}(\overset{s^2}{u}_{j,0} + 2\overset{s^2}{v}_{j,0}) = \overset{s+1\,s+1}{\lambda}\, u_{j,0}, \\
\frac{\tilde{D}}{4}\frac{\overset{s+1}{v}_{j,1} - \overset{s+1}{v}_{j,0}}{0.5h_r^2} + \frac{D_2}{2}\overset{s+1}{v}_{\bar{t}t,j} + \frac{\gamma}{2}\overset{s}{u}_{j,0}\overset{s+1}{u}_{j,0} + \alpha \overset{s+1}{v}_{j,0}(2\overset{s^2}{u}_{j,0} + \overset{s^2}{v}_{j,0}) + \frac{\Delta k}{2}\overset{s+1}{v}_{j,0} = \overset{s+1\,s+1}{\lambda}\, v_{j,0}, \\
\overset{s+1}{u}_{j,N_r} = \overset{s+1}{v}_{j,N_r} = 0, \quad \overset{s+1}{u}_{0,k} = \overset{s+1}{u}_{N_r,k} = \overset{s+1}{v}_{0,k} = \overset{s+1}{v}_{N_r,k} = 0, \quad s = 0, 1, \ldots.
\end{gathered} \tag{10}$$

Итерации в слагаемых, соответствующих квадратичной нелинейности, расставлены аналогично работам [42], [44]. Однако, в отличие от работы [44], несимметричность оператора Лапласа по координате r не позволяет симметризовать матрицу уравнений (10) с помощью замены переменных $w = v\sqrt{2}$.

Для реализации итерационного процесса необходимо задать значения функций на нулевой итерации ($s = 0$). В качестве начального приближения выбирался гауссов пучок:

$$\overset{s=0}{u} = \overset{s=0}{v} = e^{-(r/r_p)^2} e^{-((t - L_t/2)/t_p)^2} \tag{11}$$

либо распределение в виде синуса

$$\overset{s=0}{u}_m = \overset{s=0}{v}_m = \sin\left(\frac{\pi m r}{L_r}\right)\sin\left(\frac{\pi m t}{L_t}\right), \quad m = 1, 2, \ldots. \tag{12}$$

Введем вектор $\psi = (u_{1,0}, v_{1,0}, \ldots, u_{1,N_r-1}, v_{1,N_r-1}, u_{2,0}, v_{2,0}, \ldots, u_{2,N_r-1}, v_{2,N_r-1}, \ldots, u_{N_t-1,0}, v_{N_t-1,0}, \ldots, u_{N_t-1,N_r-1}, v_{N_t-1,N_r-1})$. Тогда систему уравнений (10) можно представить в компактной форме:

$$\overset{s\,s+1}{\Lambda}\,\psi = \overset{s+1\,s+1}{\lambda}\,\psi,$$

где $\overset{s}{\Lambda}$ – вещественная несимметричная ленточная матрица порядка $2N_r(N_t - 1)$ с числом диагоналей, равным $(2N_r + 1)$. Из них ненулевыми являются 1-я, 2-я и $(2N_r + 1)$-я наддиагонали и поддиагонали основной диагонали. Для записи этой матрицы введем следующие обозначения:

$$\begin{aligned}
\overset{s}{a}_{j,k} &= -\frac{2\tilde{D}}{(r_{k+1} - r_k)(r_k - r_{k-1})} - \frac{2D_1}{\tau^2} + \alpha(\overset{s\,2}{u}_{j,k} + 2\overset{s\,2}{v}_{j,k}), \\
\overset{s}{b}_{j,k} &= -\frac{\tilde{D}}{(r_{k+1} - r_k)(r_k - r_{k-1})} - \frac{D_2}{\tau^2} + \alpha(2\overset{s\,2}{u}_{j,k} + \overset{s\,2}{v}_{j,k}) + \frac{\Delta k}{2},
\end{aligned} \tag{13}$$

$$p_k^{(1)} = \frac{r_k + r_{k-1}}{r_k(r_{k+1} - r_{k-1})(r_k - r_{k-1})}, \quad p_k^{(2)} = \frac{1}{(r_{k+1} - r_k)(r_k - r_{k-1})}, \quad p_k^{(3)} = \frac{r_{k+1} + r_k}{r_k(r_{k+1} - r_{k-1})(r_{k+1} - r_k)}.$$

С учетом обозначений (11)–(13) матрицу $\overset{s}{\Lambda}$ можно представить в виде

$$\overset{s}{\Lambda} = \begin{pmatrix}
\overset{s}{a}_{1,0} & \gamma\overset{s}{u}_{1,0} & \tilde{D}p_1^{(3)} & 0 & \cdots & 0 & \frac{D_1}{\tau^2} & 0 & \cdots \\
\frac{\gamma}{2}\overset{s}{u}_{1,0} & \overset{s}{b}_{1,0} & 0 & \frac{\tilde{D}}{2}p_1^{(3)} & 0 & \cdots & 0 & 0 & \frac{D_2}{2\tau^2} & 0 & \cdots \\
\tilde{D}p_2^{(1)} & 0 & \overset{s}{a}_{1,1} & \gamma\overset{s}{u}_{1,1} & \tilde{D}p_2^{(3)} & 0 & 0 & 0 & \frac{D_1}{\tau^2} & 0 \\
0 & \frac{\tilde{D}}{2}p_2^{(1)} & \frac{\gamma}{2}\overset{s}{u}_{1,1} & \overset{s}{b}_{1,1} & 0 & \frac{\tilde{D}}{2}p_2^{(3)} & 0 & 0 & 0 & \frac{D_2}{2\tau^2} \\
 & 0 & \tilde{D}p_3^{(1)} & 0 & \overset{s}{a}_{1,2} & \gamma\overset{s}{u}_{1,2} & \tilde{D}p_3^{(3)} & 0 & & 0 \\
 & & 0 & \frac{\tilde{D}}{2}p_3^{(1)} & \frac{\gamma}{2}\overset{s}{u}_{1,2} & \overset{s}{b}_{1,2} & 0 & \frac{\tilde{D}}{2}p_3^{(3)} & & \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\frac{D_1}{\tau^2} & 0 & \cdots & & & 0 & \overset{s}{a}_{2,0} & \gamma\overset{s}{u}_{2,0} & \tilde{D}p_1^{(3)} & 0 \\
0 & \frac{D_2}{2\tau^2} & 0 & & & 0 & \frac{\gamma}{2}\overset{s}{u}_{2,0} & \overset{s}{b}_{2,0} & 0 & \frac{\tilde{D}}{2}p_1^{(3)} \\
 & 0 & \frac{D_1}{\tau^2} & 0 & & 0 & \tilde{D}p_2^{(1)} & 0 & \overset{s}{a}_{2,1} & \gamma\overset{s}{u}_{2,1} \\
 & & 0 & \frac{D_2}{2\tau^2} & 0 & 0 & 0 & \frac{\tilde{D}}{2}p_2^{(1)} & \frac{\gamma}{2}\overset{s}{u}_{2,1} & \overset{s}{b}_{2,1} \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots
\end{pmatrix}.$$

Заметим, что для случая одного уравнения (4) верны все рассуждения, изложенные выше. Для получения матрицы, соответствующей одному уравнению, необходимо исключить из приведенной выше матрицы все строки и столбцы с четными номерами, а также положить $\overset{s}{v}_{j,k} \equiv 0$ в формулах (10). Полученная матрица будет иметь порядок $N_r(N_t - 1)$ с числом диагоналей, равным $(N_r + 1)$, из которых ненулевыми будут основная диагональ, а также 1-я и $(N_r + 1)$-я наддиагонали и поддиагонали.

Необходимо отметить, что даже при выборе равномерной сетки по координате r невозможно получить симметричную матрицу $\overset{s}{\Lambda}$. Кроме этого, исходя из физической постановки задачи, на каждой итерации происходит нормировка вектора $\overset{s}{\psi}$ в соответствии с условием $\max\limits_j |\overset{s}{\psi}_j| = 1$. Итерационный процесс завершается, если достигается условие

$$|\overset{s+1}{\lambda} - \overset{s}{\lambda}| < \varepsilon|\overset{s}{\lambda}| + \delta, \quad \varepsilon, \delta > 0. \tag{14}$$

Таким образом, задача нахождения функций $u(t, r)$, $v(t, r)$ сводится к нахождению СЗ и СФ матрицы $\overset{s}{\Lambda}$. Для нахождения СЗ несимметричной матрицы теоретически можно использовать QR-алгоритм, предварительно сведя ее к форме Хессенберга (см. [46], [47], [49]). Это позволяет

Таблица

Число узлов сетки	Одинарная точность, Гб	Двойная точность, Гб	Оценка времени на проведение одной итерации, сутки
$N_t = N_r = 100$	0.8	1.6	0.1
$N_t = N_r = 200$	12.8	25.6	7.4
$N_t = N_r = 400$	204.8	409.6	474.1

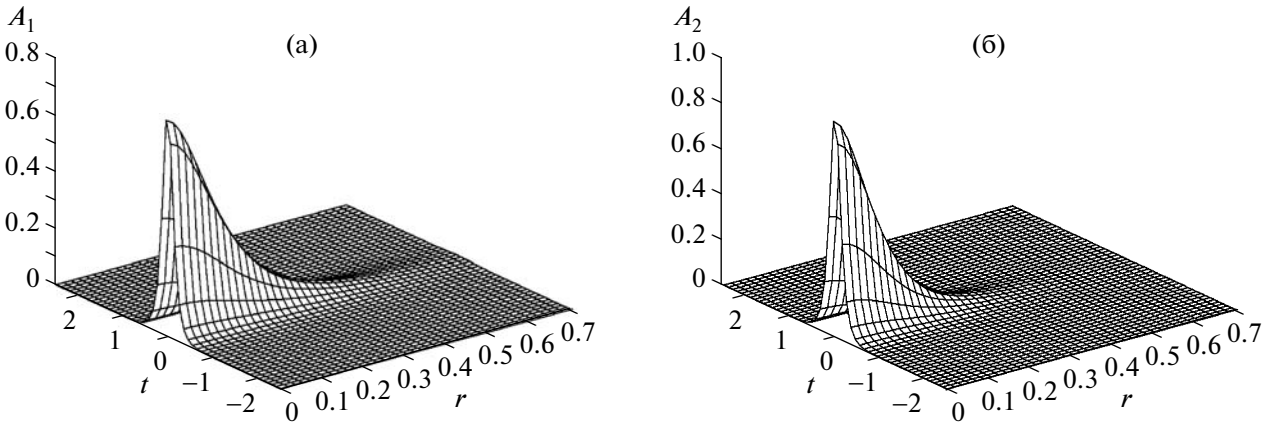
понизить сложность алгоритма с $O(N_t^4 \times N_r^4)$ до $O(N_t^3 \times N_r^3)$. Однако и в этом случае проблема нехватки вычислительных ресурсов (таких как оперативная память и мощность процессора) встает здесь очень остро. Для оценки оперативной памяти рассмотрим верхнюю надтреугольную матрицу. В таблице приводится зависимость размера матрицы $\overset{s}{\Lambda}$ в Гб и числа операций с плавающей точкой от выбранного числа узлов сетки. Подчеркнем, что реальные расчеты и, соответственно, экспериментальная проверка приведенных в таблице данных не проводились. Оценки в ней построены исходя из предположения, что вычисления в этом алгоритме могут быть эффективно распараллелены для исполнения на многопроцессорном кластере (что, вообще говоря, не всегда возможно). В качестве модельного компьютера, нами был выбран многопроцессорный кластер Leo, состоящий из 32 процессоров Xeon 2.6 ГГц (см. [48]). В исследовании, описанном в [48], производительность этого кластера на тестах Linpack достигала 100GFLOP/s. Руководствуясь этим, в таблице примерно оценено время работы одной итерации алгоритма на разных сетках. Как видно, существенные проблемы возникают уже при выборе числа узлов, большего 200. Расчет же на сетке с числом узлов по каждой из размерностей более 400 не представляется возможным.

Поэтому для поиска СЗ несимметричной матрицы $\overset{s}{\Lambda}$ нужен алгоритм, который требует, по крайней мере по памяти, меньше $O(N_t^2 \times N_r^2)$ машинных слов и работы, меньшей чем $O(N_t^3 \times N_r^3)$ операций с плавающей точкой. Одним из таких алгоритмов является алгоритм Арнольди (см. [49]). Наиболее современная его реализация представлена, например, в стандартном пакете ARPACK (см. [50]). Его применение позволяет эффективно хранить матрицу $\overset{s}{\Lambda}$, требуя для этого порядка $O(N_t \times N_r)$ ячеек, а также проводить вычисления за приемлемое время. Так, с использованием этого алгоритма нами найдены СФ уравнений (1) на сетке размером 1024×500 узлов на компьютере Intel Itanium 2 RX6600 1.6 GHz, содержащем 4 двухядерных процессора с общей памятью 96 Гб.

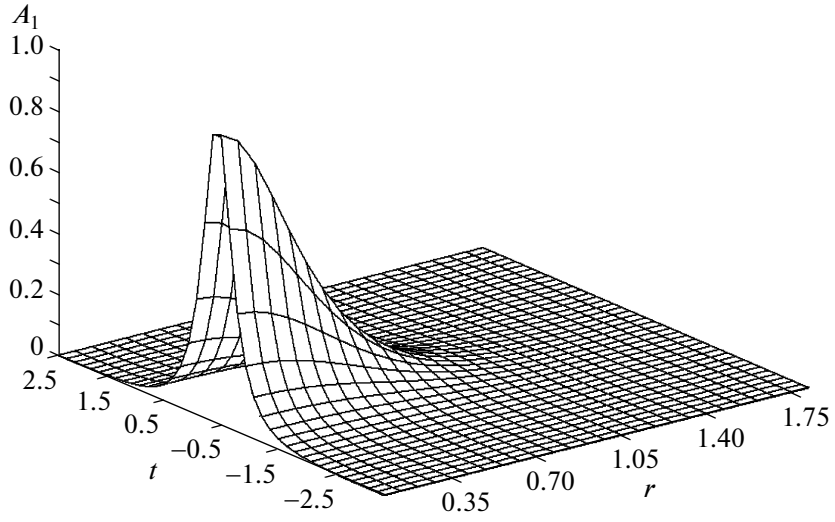
Помимо условия (14), при вычислении СФ контролировалась также невязка Ψ решения в норме C. Эта характеристика важна именно для нахождения солитонного решения уравнения Шрёдингера. В этом случае она не должна превышать 10^{-5}. При нахождении СФ разностной задачи (7) это требование может быть значительно ослаблено.

3. РЕЗУЛЬТАТЫ КОМПЬЮТЕРНОГО МОДЕЛИРОВАНИЯ

В качестве иллюстрации результатов, полученных с помощью описанного выше алгоритма, на фиг. 1 представлены СФ для первого СЗ. С целью подтверждения того, что найденные СФ являются оптическими солитонами, распределения амплитуд, представленные на фиг. 1, использовались в качестве начальных распределений системы (1). Распространение анализировалось на трассе $0 \le z \le 1$. Для этой трассы контролировалось отклонение пиковой интенсивности распространяющейся волны от исходной. Компьютерное моделирование показало, что изменение пиковой интенсивности не превышает 10^{-3}. Важно подчеркнуть, что СФ задачи (1), соответствующие первому СЗ, не изменяются с ростом размера области по времени и поперечной координате, т.е. они являются оптическими солитонами. Отметим также, что форма солитонных реше-



Фиг. 1. Форма двухчастотного солитонного решения уравнения (1) на основной (а) и удвоенной (б) частоте, соответствующая первому СЗ, для $D_1 = 0.08$, $D_2 = 0.14$, $\tilde{D} = 0.1$, $\alpha = 5$, $\gamma = 20$.

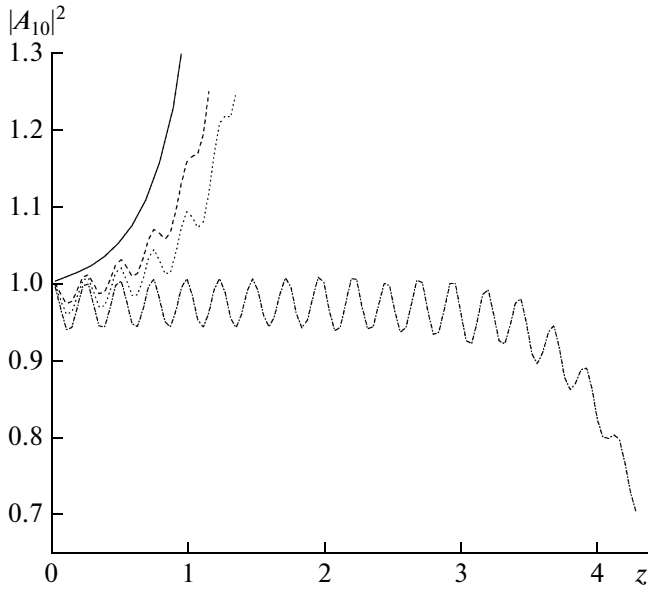


Фиг. 2. Форма одночастотного солитонного решения уравнения (4), соответствующего первому СЗ, для $D_1 \equiv \equiv D = 0.1$, $\tilde{D} = 0.1$, $\alpha = 10$.

ний, отвечающих первому СЗ, не зависит от выбора начального приближения в итерационном процессе (10).

Ввиду своей практической значимости, на фиг. 2 показан одночастотный солитон для случая одного уравнения Шрёдингера (4). Как отмечалось в работах [8], [11], [14], [51]–[56], аксиально-симметричные солитоны неустойчивы к малым начальным возмущениям их формы. Так, возмущения начальной интенсивности, меньшие 0.1%, могут приводить к изменениям пиковой интенсивности волны до 3–5% в двумерном случае (в отсутствие координаты t) на довольно небольшой трассе: $z \le 2$. В 3D-случае имеет место экспоненциальный рост возмущений на трассе порядка 1. Для стабилизации распространения возмущенных солитонов применим периодическую модуляцию керровской нелинейности вдоль направления распространения волны:

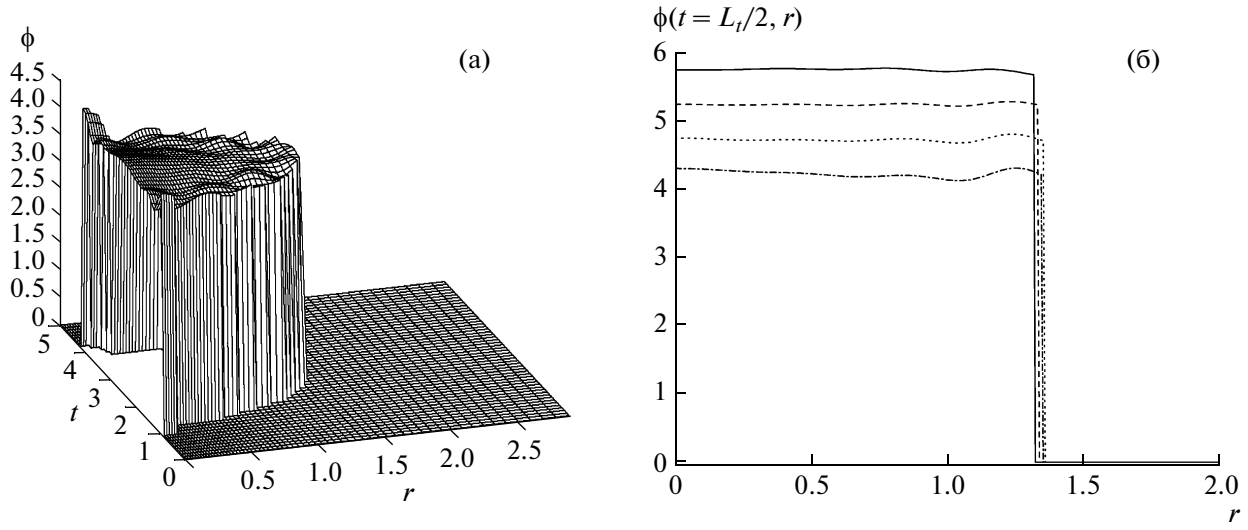
$$\alpha \equiv \alpha(z) = \alpha_0[1 + \delta_\alpha \sin(\omega_\alpha z)]. \tag{15}$$



Фиг. 3. Динамика изменения вдоль продольной координаты оптического излучения, имеющего на входе солитонную форму, изображенную на фиг. 2, с начальным возмущением интенсивности 0.5% при постоянном (сплошная линия) и периодически изменяющемся коэффициенте кубичной нелинейности α с различными амплитудами модуляции δ_α: -0.05 (пунктир); -0.07 (штриховая линия); -0.1 (штрихпунктир) для $D_1 \equiv D = 0.1$, $\tilde{D} = 0.1$, $\alpha = 10$, $\omega_\alpha^{(1)} = 29$, $\omega_\alpha^{(2)} = 23.2$.

Основным отличием рассматриваемого способа стабилизации от ранее известных является то, что ширина отрезков среды с разной нелинейностью различна за счет выбора масштабирующего коэффициента синуса в виде $\omega_\alpha = \begin{cases} \omega_\alpha^{(1)}, & z \in \Omega_z^{(1)}, \\ \omega_\alpha^{(2)}, & z \in \Omega_z^{(2)}, \end{cases}$ что позволяло на несколько порядков уменьшить амплитуду модуляции по сравнению с ее значением, рассматриваемым в других работах. В качестве примера для солитона, изображенного на фиг. 2, на фиг. 3 показана стабилизация возмущенного распространения солитона на трассе длиной более четырех безразмерных единиц за счет специального выбора величин $\omega_\alpha^{(1)}$, $\omega_\alpha^{(2)}$. Из графика видно, что в однородной среде ($\delta_\alpha = 0$) возмущение входной амплитуды 0.5% приводит к экспоненциальному росту пиковой интенсивности солитона (сплошная линия). Модуляция коэффициента кубичной нелинейности амплитудой $\delta_\alpha = -0.05; -0.07$ позволяет незначительно увеличить длину среды, в которой возможно распространение возмущенной волны (пунктир и штриховые кривые, фиг. 3). Дальнейшее увеличение амплитуды модуляции $\delta_\alpha = -0.1$ позволяет добиться стабилизации на большем участке среды (штрихпунктир). В этом случае сначала пиковая интенсивность распространяющейся волны периодически колеблется, а потом даже начинает уменьшаться. Следовательно, полностью стабилизировать распространение солитона не удается. Тем не менее интересно, что пиковая интенсивность уменьшается и нет “схлопывания” пучка. Важно также подчеркнуть, что, несмотря на осцилляции пиковой интенсивности, форма солитона при распространении в нелинейной среде не разрушается. Принципиально, что стабилизация солитона происходит для малой амплитуды модуляции, так как это приводит к отсутствию отражений от неоднородностей коэффициента нелинейности (отраженная волна будет пренебрежимо мала). Поэтому используемое описание в рамках уравнения Шрёдингера без учета отраженной волны справедливо, что, вообще говоря, не очевидно для случая сильной модуляции при $\delta_\alpha \gg 1$.

Отметим еще одну особенность данного режима: на фиг. 3 видно, что после $z = 4$ волна начинает распадаться. Это связано с накоплением дефокусирующей линзы при чередовании слоев с разной величиной нелинейности. Действительно, когда нелинейность уменьшается, пучок и им-



Фиг. 4. Распределение фазы волны, имеющей на входе солитонную форму (фиг. 3), на выходе из нелинейной среды (график (а)) при стабилизации возмущенной волны с $\delta_\alpha = -0.1$; в центре импульса ($t = L_t/2$) в различных сечениях по продольной координате (график (б)) $z = 1$ (сплошная линия); $z = 2$ (пунктир); $z = 3$ (штриховая линия); $z = 4$ (штрихпунктир).

пульс будет расширяться (по времени растягиваться). Это происходит вследствие того, что каждому значению коэффициента нелинейности соответствует солитон определенной амплитуды, пространственного радиуса и временно́й длительности. При уменьшении α должна уменьшиться пиковая интенсивность и увеличатся пространственный и временно́й размеры. Следовательно, лазерное излучение приобретет отрицательную кривизну волнового фронта и чирпирования импульса и оптическое излучение будет дефокусироваться. Затем при возрастании коэффициента нелинейности оптическое излучение будет приобретать волновой фронт с положительной кривизной и положительное чирпирование. Из-за того, что не удается точно подобрать параметры, не получается полностью сбалансировать самофокусировку и дефокусировку волны, что приводит либо к росту пиковой интенсивности $\delta_\alpha = -0.5, -0.07$, либо к дефокусировке пучка и декомпрессии импульса $\delta_\alpha = -0.1$. Заметим, что выбор отрицательных амплитуд модуляции обусловлен корректировкой самофокусировки пучка за счет первого “дефокусирующего” слоя среды.

Сделанное выше рассуждение подтверждается на фиг. 4, где изображено пространственно-временно́е распределение фазы в сечении $z = 4.0$, а также ее распределение на оси импульса в различных продольных сечениях среды. Как видно, она в последнем рассматриваемом сечении среды имеет выпуклую кривизну как по времени, так и по пространству, что говорит о декомпрессии импульса и дефокусировке пучка.

ВЫВОДЫ

В настоящей работе предложен итерационный метод нахождения солитонных решений для системы двух нелинейных уравнений Шрёдингера, описывающий процесс генерации второй гармоники в среде с квадратичным и кубичным откликом в трехмерной постановке в аксиально-симметричном случае. Он основан на нахождении СФ и СЗ соответствующей нелинейной задачи. Проведенный анализ показал, что ограниченность вычислительных ресурсов не позволяет использовать для нахождения СЗ и СФ методы, имеющие сложность порядка $O(N_t^3 \times N_r^3)$, которой обладает, например, QR-алгоритм. Применение алгоритма Арнольди позволяет эффективно использовать оперативную память, а также уменьшить время нахождения СФ до приемлемого. На его основе и с использованием предложенного итерационного метода можно эффективно решать задачи данного класса.

Предложенный метод также применялся для нахождения солитонов трехмерного НУШ, описывающего распространение волн в кубичной среде. Так как в этом случае солитон неустойчив, то для его стабилизации предложена слабая модуляция коэффициента нелинейности. Показано, что обсуждаемый в работе метод позволяет значительно увеличить длину среды, в которой может распространяться возмущенный солитон. Дана интерпретация декомпрессии импульса и дефокусировки пучка при стабилизации солитонов с помощью модуляции коэффициента нелинейности при отсутствии смены его знака. Необходимо подчеркнуть, что в работе исследовалась устойчивость солитона и методы стабилизации развивающейся неустойчивости для одного уравнения Шрёдингера. Вопрос об устойчивости солитонов для системы двух уравнений требует дополнительного исследования.

СПИСОК ЛИТЕРАТУРЫ

1. *Buryak A.V., Kivshar Yu.S.* Spatial optical solitons governed by quadratic nonlinearity // Opt. Letts. 1994. V. 19. № 20. P. 1612–1615.
2. *Buryak A.V., Trapani P.D., Skryabin D.V. et al.* Optical solitons due to quadratic nonlinearities: from basic physics to futuristic applications // Phys. Rept. 2002. V. 370. № 2. P. 63–235.
3. *Etrich C., Lederer F., Malomed B.A. et al.* Optical solitons in media with a quadratic nonlinearity // Progress in Optics. 2000. V. 41. P. 483–568.
4. *Brull L., Lange H.* Stationary, oscillatory and solitary wave type solution of singular nonlinear Schrödinger equations // Math. Meth. in Appl. Sci. 1986. V. 8. № 4. P. 559–575.
5. *Liu X., Beckwitt K., Wise F.W.* Two-dimensional optical spatiotemporal solitons in quadratic media // Phys. Rev. E. 2000. V. 62. № 1. P. 1328–1340.
6. *Stegeman G., Hagan D.J., Torner L.* $\chi^{(2)}$ cascading phenomena and their applications to all-optical signal processing, mode-locking, pulse compression and solitons // Opt. Quantum Electron. 1996. V. 28. № 12. P. 1691–1740.
7. *Ashihara S., Nishina J., Shimura T. et al.* Soliton compression of femtosecond pulses in quadratic media // JOSA B. 2002. V. 19. № 10. P. 2505–2510.
8. *Towers I., Malomed B.A.* Stable (2 + 1)-dimentional solitons in a layered medium with sign-alternating Kerr nonlinearity // JOSA B. 2002. V. 19. № 3. P. 537–543.
9. *Steblina V., Kivshar Yu.S., Lisak M. et al.* Self-guided beams in a diffractive $\chi^{(2)}$ medium: variational approach // Opt. Communs. 1995. V. 118. P. 345–352.
10. *Yang J., Malomed B.A., Kaup D.J.* Embedded solitons in second-harmonic-generating systems // Phys. Rev. Lett. 1999. V. 83. № 10. P. 1958–1961.
11. *Mihalache D., Mazilu D., Malomed B.A. et al.* Stable three-dimensional optical solitons supported by competing quadratic and self-focusing cubic nonlinearities // Phys. Rev. E. 2006. V. 74. № 4. P. 047601–1:4.
12. *Malomed B.A.* Soliton management in periodic systems. New York: Springer, 2006.
13. *Sakaguchi H., Malomed B.A.* Resonant nonlinearity management for nonlinear Schrödinger solitons // Phys. Rev. E. 2004. V. 70. P. 066613.
14. *Berge L., Mezentsev V.K., Rasmussen J.J. et al.* Self-guiding light in layered nonlinear media // Opt. Letts. 2000. V. 25. № 14. P. 1037–1039.
15. *Rozanov N.N., Fedorov S.V., Shatsev A.N.* Incoherent weak coupling of laser solitons // Optics & Spectroscopy. 2007. V. 102. № 1. P. 83–85.
16. *Бабушкин И.В., Лойко Н.А., Розанов Н.Н.* Пространственные солитоноподобные структуры в тонкопленочной системе с поперечным фотонным кристаллом в цепи обратной связи // Оптика и спектроскопия. 2007. Т. 102. № 2. С. 285–291.
17. *Schiek R., Baek Y., Stegeman G. et al.* Interactions between one-dimensional quadratic soliton-like beams // Optical and Quantum Electronics. 1998. V. 30. № 7–10. P. 861–879.
18. *Driben R., Oz Y., Malomed B.A. et al.* Mismatch management for optical and matter-wave quadratic solitons // Phys. Rev. E. 2007. V. 75. P. 026612.
19. *Saito H., Ueda M.* Dynamically stabilized bright solitons in a two-dimensional Bose-Einstein condensate // Phys. Rev. Letts. 2003. V. 90. P. 040403.
20. *Abdullaev F.K., Caputo J.G. , Kraenkel R.A., Malomed B.A.* Controlling collapse in Bose-Einstein condensates by temporal modulation of the scattering length // Phys. Rev. A. 2003. V. 67. P. 013605.

21. *Towers I., Buryak A.V., Sammut R. A., Malomed B.A.* Stable localized vortex solitons // Phys. Rev. E. 2001. V. 63. № 5. P. 055601(R).

22. *Захаров В.Е., Манаков С.В., Новиков С.П., Питаевский Л.П.* Теория солитонов: Метод обратной задачи. М.: Наука,1980.

23. *Кернер Б.С., Осипов В.В.* Автосолитоны. М.: Наука, 1991.

24. *Abdullaev F., Darmanyan S., Khabibullaev P.* Optical solitons. Berlin, Heidelberg, 1993.

25. *Абловиц М., Сигур Х.* Солитоны и метод обратной задачи. М.: Мир, 1987.

26. *Ablowitz M.J., Clarkson P.A.* Solitons. Nonlinear Evolution Equations and Inverse Scattering. Cambridge: Cambridge Univ. Press. London Math. Soc. Lect. Notes Ser. 149. 1991.

27. *Тахтаджян Л.А., Фаддеев Л.Д.* Гамильтонов подход в теории солитонов. М.: Наука, 1986.

28. *Лэм Дж.Л.* Введение в теорию солитонов. М.: Мир, 1983.

29. *Богоявленский О.И.* Опрокидывающие солитоны: Нелинейные интегрируемые уравнения. М.: Наука, 1991.

30. *Наянов В.И.* Многополевые солитоны. М.: Физматлит, 2006.

31. *Додд Р., Эилбек Дж., Гиббон Дж. и др.* Солитоны и нелинейные волновые уравнения. М.: Мир, 1988.

32. *Мива Т., Джимбо М., Датэ Э.* Солитоны: дифференциальные уравнения, симметрии и бесконечномерные алгебры. М.: МЦНМО, 2005.

33. *Инфельд Э., Роуландс Дж.* Нелинейные волны, солитоны и хаос. М.: Физматлит, 2005.

34. *Ньюэлл А.* Солитоны в математике и физике. М.: Мир, 1989.

35. *Калоджеро Ф., Дегасперис А.* Спектральные преобразования и солитоны. Методы решения и исследования нелинейных эволюционных уравнений. М.: Мир, 1985.

36. *Давыдов А.С.* Солитоны в молекулярных системах. Киев: Наук. думка, 1984.

37. *Новокшенов В.Ю.* Введение в теорию солитонов. Москва–Ижевск: РХД, 2002.

38. *Раджараман Р.* Солитоны и инстантоны в квантовой теории поля. М.: Мир, 1985.

39. *Кившарь Ю.С., Агравал Г.П.* Оптические солитоны. От световодов к фотонным кристаллам. М.: Физматлит, 2005.

40. *Ахмедиев Н.Н., Анкевич А.* Солитоны. Нелинейные импульсы и пучки. М.: Физматлит, 2003.

41. *Карамзин Ю.Н., Сухоруков А.П.* Нелинейное взаимодействие дифрагирующих световых пучков в среде с квадратичной нелинейностью, взаимофокусировка пучков и ограничение эффективности оптических преобразователей частоты // Письма в ЖЭТФ. 1974. Т. 20. № 11. С. 734–739.

42. *Варенцова С.А., Трофимов В.А.* О разностном методе нахождения собственных мод нелинейного уравнения Шрёдингера // Вестн. МГУ. Сер. 15. 2005. № 3. С. 16–22.

43. *Trofimov V.A., Varentsova S.A.* Computational method for finding of soliton solutions of a nonlinear Schrödinger equation // Lect. Notes Math. Berlin, Heidelberg: Springer, 2005. V. 3401. P. 550–557.

44. *Матусевич О.В., Трофимов В.А.* Итерационный метод нахождения собственных функций системы двух уравнений Шрёдингера с комбинированной нелинейностью // Ж. вычисл. матем. и матем. физ. 2008. Т. 48. № 4. С. 713–724.

45. *Самарский А.А., Андреев В.Б.* Разностные методы для эллиптических уравнений. М.: Наука, 1976.

46. *Голуб Дж., Ван Лоун Ч.* Матричные вычисления. М.: Мир, 1999.

47. *Самарский А.А., Гулин А.В.* Численные методы. М.: Наука, 1989.

48. http://parallel.ru/cluster/leo_linpack.html

49. *Деммель Дж.* Вычислительная линейная алгебра. Теория и приложения. М.: Мир, 2001.

50. http://www.caam.rice.edu/software/ARPACK/

51. *Malomed B.A., Mihalache D., Wise F., Torner L.* Spatiotemporal optical solitons // J. Opt. B. 2005. V. 7. № 5. P. 53–72.

52. *Nijhof J.H.B., Doran N.J., Forysiak W., Knox F.M.* Stable soliton-like propagation in dispersion managed systems with net anomalous, zero and normal dispersion // Electron. Letts. 1997. V. 33. № 20. P. 1726–1727.

53. *Lakoba T., Yang J., Kaup D.J., Malomed B.A.* Conditions for stationary pulse propagation in the strong dispersion management regime // Opt. Communs. 1998. V. 149. № 4–6. P. 366–375.

54. *Moll K.D., Gaeta A.L., Fibich G.* Self-similar optical wave collapse: observation of the Townes soliton // Phys. Rev. Letts. 2003. V. 90. № 20. P. 203902–1:4.

55. *Montesinos G.D., Perez-Garcia V.M.* Numerical studies of stabilized Townes solitons // Math. Comput. in Simulation. 2005. V. 69. № 5. P. 447–456.

56. *Montesinos G.D., Perez-Garcia V.M., Torres P.J.* Stabilization of solitons of the multidimensional nonlinear Schrödinger equation: matter-wave breathers // Physica D. 2004. V. 191. P. 193–210.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2009, том 49, № 11, с. 2001–2009

УДК 519.634

УПРАВЛЕНИЕ МАГНИТОГИДРОДИНАМИЧЕСКИМ ТЕЧЕНИЕМ ПРИ СОЗДАНИИ МАГНИТНОГО ПОЛЯ ЗАДАННОЙ КОНФИГУРАЦИИ[1)]

© 2009 г. А. Ю. Чеботарёв

(690041 Владивосток, ул. Радио, 7, ИПМ ДВО РАН)

e-mail: cheb@iam.dvo.ru

Поступила в редакцию 04.03.2009 г.
Переработанный вариант 28.05.2009 г.

Рассматриваются задачи управления для уравнений магнитной гидродинамики вязкой несжимаемой жидкости, состоящие в создании в заданный момент времени магнитного поля требуемой структуры за счет действия сторонних электродвижущих сил при условии минимизации работы над токами проводимости или минимизации выделения джоулева тепла. На основе оценок решения субдифференциальной задачи Коши для системы типа Навье–Стокса доказана возможность получения требуемого магнитного поля и даны условия разрешимости задач управления. Библ. 14.

Ключевые состава: уравнения магнитной гидродинамики, оптимальное управление, вариационные неравенства, задачи управляемости.

1. ВВЕДЕНИЕ. ПОСТАНОВКИ ЗАДАЧ

Рассмотрим течение вязкой несжимаемой и проводящей жидкости в ограниченной односвязной области $\Omega \subset \mathbb{R}^3$ с границей Γ. В безразмерных переменных течение описывается уравнениями магнитной гидродинамики (МГД):

$$\partial u/\partial t - \nu\Delta u + (u\nabla)u = -\nabla p + S\cdot \operatorname{rot} B \times B, \quad x \in \Omega, \quad t \in (0, T), \tag{1}$$

$$\partial B/\partial t + \operatorname{rot} E = 0, \quad j = \operatorname{rot} B = 1/\nu_m(E + u \times B + E_c), \tag{2}$$

$$\operatorname{div} u = 0, \quad \operatorname{div} B = 0. \tag{3}$$

Здесь u, B, E и j – векторные поля скорости, магнитной индукции, электрической напряженности и плотности тока соответственно, p – давление, $\nu = 1/\mathrm{Re}$, $\nu_m = 1/\mathrm{R}_m$, $S = \mathrm{M}^2/\mathrm{Re}\,\mathrm{R}_m$, где Re – число Рейнольдса, R_m – магнитное число Рейнольдса, M – число Гартмана. Через E_c обозначено векторное поле сторонних электродвижущих сил (ЭДС), играющих роль управления.

К уравнениям (1)–(3) добавляют начальные условия

$$u|_{t=0} = u_0(x), \quad B|_{t=0} = B_0(x), \quad x \in \Omega, \tag{4}$$

и условия на границе Γ области течения

$$u = 0, \quad B\cdot n = 0, \quad n \times E = 0, \quad (x, t) \in \Gamma \times (0, T), \tag{5}$$

где n – единичный вектор внешней нормали к границе Γ.

В дальнейшем, поскольку давление p и напряженность электрического поля E можно исключить из системы (1) – (3), под решением системы будем понимать пару $y = \{u, B\}$, определяя поле E из второго уравнения (2).

Рассмотрим задачу управляемости, состоящую в создании в данный момент времени T магнитного поля заданной конфигурации $B_s(x)$, $x \in \Omega$, за счет действия сторонних ЭДС.

[1)] Работа выполнена при финасовой поддержке гранта ДВО РАН (проект 09-II-CO-01-002) и гранта программы поддержки ведущих научных школ (проект НШ-2810.2008.1).

Задача 1. Найти векторное поле E_c и соответствующее ему решение $y = \{u, B\}$ системы (1)–(3), удовлетворяющее условиям (4), (5) такое, что

$$B|_{t=T} = B_s. \tag{6}$$

Заметим, что величина $\int_\Omega \operatorname{rot} B \cdot E_c dx$ соответствует работе, совершаемой сторонними ЭДС E_c в единицу времени, над токами проводимости $j = \operatorname{rot} B$, тогда как интеграл $\int_\Omega E_c^2 dx$ пропорционален джоулеву теплу, выделяемому сторонними токами в единицу времени (см. [1, с. 424]). Из (1)–(6) после умножения (1) на $(1/S)u$ и первого уравнения в (2) на B, интегрирования по частям по Ω, по времени от 0 до T и сложения полученных равенств вытекает, что работа сторонних ЭДС определяется по формуле

$$\frac{1}{2S}\int_\Omega (u^2|_{t=T} - u_0^2)dx + \frac{1}{2}\int_\Omega (B_s^2 - B_0^2)dx + \int_0^T\int_\Omega \left(\nu_m|\operatorname{rot} B|^2 + \frac{\nu}{S}|\operatorname{rot} u|^2\right)dxdt.$$

Здесь и далее, если w – вектор-функция, то через w^2 обозначаем ее скалярный квадрат $w \cdot w$.

Таким образом, приходим к следующим постановкам задач о минимизации выделения джоулева тепла или работы сторонних ЭДС.

Задача 2. Найти векторные поля E_c, u, B, удовлетворяющие (1)–(6) и минимизирующие функционал

$$J_h = \int_0^T\int_\Omega E_c^2 dxdt.$$

Задача 3. Найти векторные поля E_c, u, B, удовлетворяющие (1)–(6) и минимизирующие функционал

$$J_a = \frac{1}{2S}\int_\Omega u^2|_{t=T}dx + \int_0^T\int_\Omega \left(\nu_m|\operatorname{rot} B|^2 + \frac{\nu}{S}|\operatorname{rot} u|^2\right)dxdt$$

при условии

$$\int_0^T\int_\Omega E_c^2 dxdt \le M.$$

Здесь $M > 0$ – заданное число.

Математические вопросы для классических краевых задач в модели (1)–(3) изучены в [2]. Задача о минимизации работы для уравнений Навье–Стокса и задачи управляемости изучены в [3]. В [4]–[6] рассмотрены неравенства Навье–Стокса и эволюционные МГД-неравенства.

Основной результат работы – теорема о разрешимости задачи 1 – получен на основе оценок решения задачи Коши для субдифференциального включения, описывающего управление с обратной связью для уравнений (1)–(3). Отметим, что задачи создания магнитного поля заданной конфигурации возникают во многих приложениях магнитной гидродинамики, например при проектировании гидромагнитных систем и МГД-генераторов. Теоретический анализ и прикладные вопросы управления эволюционными МГД-системами рассмотрены в [7]–[12].

2. ФОРМАЛИЗАЦИЯ ЗАДАЧ УПРАВЛЕНИЯ

Далее, не нарушая общности, будем считать, что параметр S в модели (1)–(3) равен 1, поскольку всегда можно сделать переобозначения $B := \sqrt{S}B$, $E := \sqrt{S}E$, $E_c := \sqrt{S}E_c$.

2.1. Пространства и операторы для модели МГД

Пусть Ω – ограниченная односвязная область в $\mathbb{R}^3$ с границей $\Gamma \in C^2$. Рассмотрим линейные многообразия гладких вектор-функций

$$\mathcal{U}_1 = \{ v \in C^\infty(\overline{\Omega}) : \operatorname{div} v = 0, x \in \Omega,\ v = 0, x \in \Gamma \},$$

$$\mathcal{U}_2 = \{ v \in C^\infty(\overline{\Omega}) : \operatorname{div} v = 0, x \in \Omega, n \cdot v = 0, x \in \Gamma \}.$$

Обозначим через V_1, V_2 замыкания $\mathcal{U}_1$, $\mathcal{U}_2$ по норме $W_2^1(\Omega)$, через H_1, H_2 – замыкания $\mathcal{U}_1$, $\mathcal{U}_2$ по норме $L^2(\Omega)$, при этом фактически $H_1 = H_2$. Здесь и далее через $W_p^l(\Omega)$ обозначаем пространства Соболева. Скалярное произведение и норма в пространствах H_1, H_2 определяются обычным образом:

$$(u, v)_0 = \int_\Omega (u \cdot v)dx, \quad |u| = \sqrt{(u, u)_0}.$$

Этим же символом $(\cdot, \cdot)_0$ будем обозначать отношения двойственности между V_1 (соответственно, V_2) и сопряженным пространством V_1' (соответственно, V_2'). Поскольку область Ω односвязна, то билинейная форма

$$((u, v)) = (\operatorname{rot} u, \operatorname{rot} v)_0 = \int_\Omega (\operatorname{rot} u \cdot \operatorname{rot} v)dx \quad \forall u, v \in V_1, V_2$$

определяет скалярное произведение в V_1 и V_2, при этом определяемая им норма $\|u\| = \sqrt{((u, u))}$ эквивалентна норме пространства $W_2^1(\Omega)$. Рассмотрим также пространства

$$V = V_1 \times V_2, \quad H = H_1 \times H_2, \quad V \subset H = H' \subset V'.$$

Указанные вложения являются плотными и непрерывными. Нормы в пространствах V и H, соответственно, также обозначаем через $\|\cdot\|$, $|\cdot|$; $(\cdot, \cdot)$ – отношение двойственности между V' и V и скалярное произведение в H;

$$(y, z) = (u, v)_0 + (B, w)_0, \quad (y, z)_V = ((u, v)) + ((B, w)) \quad \forall y = \{u, B\}, \quad z = \{v, w\}.$$

В дальнейшем если X – банахово пространство, то через $L^p(0, T; X)$ обозначаем пространство L^p функций, определенных на $(0, T)$, со значениями в X.

Начально-краевую задачу (1)–(5) сведем к задаче Коши для дифференциального уравнения с операторными коэффициентами. С этой целью определим отображения $A_1 : V_1 \longrightarrow V_1'$, $A_2 : V_2 \longrightarrow V_2'$, $A : V \longrightarrow V'$, $\mathcal{B} : V \times V \longrightarrow V'$, $F : L^2(\Omega) \longrightarrow V'$, используя равенства

$$(Ay, z) = \nu((u, v)) + \nu_m((B, w)) = \nu(A_1 u, v)_0 + \nu_m(A_2 B, w)_0,$$

$$(\mathcal{B}(y_1, y_2), z) = (\operatorname{rot} u_1 \times u_2 - \operatorname{rot} B_2 \times B_1, v)_0 - (u_2 \times B_1, \operatorname{rot} w)_0,$$

$$(F(E_c), z) = (E_c, \operatorname{rot} w)_0,$$

которые выполняются для всех $y = \{u, B\}$, $y_1 = \{u_1, B_1\}$, $y_2 = \{u_2, B_2\}$, $z = \{v, w\}$ из пространства V, $E_c \in L^2(\Omega)$.

Оператор A удовлетворяет условиям

$$(Ay, y) \geq \alpha \|y\|^2, \quad \alpha = \min\{\nu, \nu_m\}, \quad (Ay, z) = (Az, y) \quad \forall y, z \in V, \tag{7}$$

а отображения $\mathcal{B}(y, z)$ и $\mathcal{B}[y] = \mathcal{B}(y, y)$ таковы, что

$$(\mathcal{B}(y, z), z) = 0,$$

$$(\mathcal{B}[y], z) = (\operatorname{rot} u \times u, v)_0 - (\operatorname{rot} B \times B, v)_0 - (u \times B, \operatorname{rot} w)_0.$$

Пусть $D(A) = \{y \in V : Ay \in H\}$. Для оператора $\mathcal{B}(y, z)$ справедливы оценки (см. [2])

$$(\mathcal{B}(y_1, y_2), y_3) \le C|y_1|^{1/4} \cdot \|y_1\|^{3/4} \cdot \|y_2\| \cdot |y_3|^{1/4} \cdot \|y_3\|^{3/4}, \quad y_1, y_2, y_3 \in V. \tag{8}$$

$$(\mathcal{B}(y_1, y_2), y_3) \le C\|y_1\| \cdot \|y_2\|^{1/2} \cdot |Ay_2|^{1/2} \cdot |y_3|, \quad y_1 \in V, \quad y_2 \in D(A), \quad y_3 \in H. \tag{9}$$

Здесь постоянная $C > 0$ зависит только от Ω, Re, R_m.

Из разложения Вейля пространства вектор-функций $L^2(\Omega)$ следует, что на величину $\int_\Omega E_c \operatorname{rot} w dx$ влияет только вихревая часть векторного поля E_c. Поэтому в дальнейшем будем считать, что $\operatorname{div} E_c = 0$, а оператор F определен на пространстве $H_3 = \operatorname{rot} V_2 = \{\operatorname{rot} w, w \in V_2\}$.

Постановка задачи (1)–(5) сводится теперь к следующей задаче Коши (см. [2], [5]):

$$y' + Ay + \mathcal{B}[y] = F(E_c), \quad y(0) = y_0. \tag{10}$$

Здесь $y \in L^2(0, T; V)$, $y' \in L^1(0, T; V')$, $y_0 = \{u_0, B_0\} \in H$, $E_c \in L^2(0, T; H_3)$.

2.2. Задачи управления

Математическая формализация сформулированных выше задач выглядит следующим образом. Пусть $U = L^2(0, T; H_3)$ – пространство управлений и, соответственно, $Y = \{z \in L^2(0, T; V) : z' \in L^1(0, T; V')\}$ – пространство состояний.

Задача 1. Пусть $y_0 \in H$, $B_s \in H_2$. Требуется найти $E_c \in U$ и соответствующее решение $y = \{u, B\} \in Y$ задачи (10), удовлетворяющее условию $B|_{t=T} = B_s$.

Пусть $\{E_c, y\}$, являющуюся решением задачи 1, будем называть допустимой.

Задача 2. Найти допустимую пару $\{E_c, y\} \in U \times Y$ такую, что

$$\|E_c\|_U = \int_0^T \int_\Omega E_c^2 dx dt \longrightarrow \inf.$$

Данная задача представляет собой фактически проблему отыскания нормального решения (т.е. решения с минимальной нормой) задачи 1.

Задача 3. Найти допустимую пару $\{E_c, y\} \in U \times Y$ такую, что

$$J_a(y) = \frac{1}{2}\int_\Omega u^2|_{t=T} dx + \int_0^T \int_\Omega (\nu_m |\operatorname{rot} B|^2 + \nu |\operatorname{rot} u|^2) dx dt \longrightarrow \inf, \tag{11}$$

при условии, что

$$\|E_c\|_U^2 \le M^2. \tag{12}$$

Здесь $M > 0$ – заданное число.

Основную трудность при исследовании разрешимости сформулированных задач представляет доказательство существования решения задачи 1, которое основано на рассмотрении управления с обратной связью (см. [13], [14]). Отметим, что в данной работе разрешимость задач управления доказана при дополнительных условиях гладкости исходных данных u_0, B_0, B_s.

3. ЗАДАЧА КОШИ ДЛЯ СУБДИФФЕРЕНЦИАЛЬНОГО ВКЛЮЧЕНИЯ

Пусть вектор-функция B_s принадлежит пространству V_2. Определим функционал

$$\Phi(y) = \rho\|B - B_s\|, \quad y = \{u, B\} \in V, \quad \rho > 0.$$

Функционал Ф является выпуклым и полунепрерывным снизу, при этом его субдифференциал имеет вид (см. [9])

$$\partial\Phi(y) = \{\chi \in V' : (\chi, z) = \rho(\operatorname{sign} \operatorname{rot}(B - B_s), \operatorname{rot} w)_0 \ \forall z = \{v, w\} \in V\},$$

где

$$\operatorname{sign}\operatorname{rot} w = \begin{cases} \operatorname{rot} w/\|w\|, & w \neq 0, \\ \operatorname{rot}\xi, & \|\xi\| \leq 1, \quad w = 0. \end{cases}$$

Рассмотрим эволюционное вариационное неравенство типа Навье–Стокса

$$(y' + Ay + \mathcal{B}[y], z - y) + \Phi(z) - \Phi(y) \geq 0 \quad \forall z \in V, \quad y(0) = y_0. \tag{13}$$

Если $\partial\Phi(y)$ – субдиффренциал функции Φ, то для каждого элемента $\chi \in \partial\Phi(y)$ справедлива оценка $\|\chi(t)\|_{V'} \leq \rho$. В [5] доказано существование слабого решения неравенства (13), которое с учетом структуры функционала Φ обладает свойствами

$$y \in L^{\infty}(0, T; H) \cap L^{2}(0, T; V), \quad y' \in L^{1}(0, T; V'), \quad y(0) = y_0.$$

При этом для произвольного $z \in V$ справедливо равенство

$$(y' + Ay + \mathcal{B}[y] + \chi, z) = 0,$$

выполняющееся в смысле распределений на $(0, T)$.

Получим априорные оценки решения неравенства (13), которые будут гарантировать существование и единственность сильного решения на некотором интервале $(0, T_*)$, где T_* не зависит от параметра ρ. Под сильным решением задачи (13) на интервале $(0, T)$ понимается функция $y \in L^{\infty}(0, T; V)$, $y' \in L^{2}(0, T; V)$ такая, что

$$-(y' + Ay + \mathcal{B}[y]) \in \partial\Phi(y) \text{ п.в. на } (0, T). \tag{14}$$

Пусть $y_0 \in V$, $B_s \in V_2$, $A_2 B_s \in H_2$. Введем обозначения

$$\psi = B - B_s, \quad \xi = \{u, \psi\}, \quad y_s = \{0, B_s\}, \quad y = \xi + y_s.$$

Тогда если y – сильное решение задачи (13), то

$$\xi' + A\xi + \mathcal{B}(\xi + y_s, \xi + y_s) = -\chi - Ay_s \text{ п.в. на } (0, T), \tag{15}$$

где $\chi \in \partial\Phi(y)$. Умножая последнее равенство скалярно на ξ и учитывая, что $(\chi, \xi) = \rho\|\psi\|$, получаем неравенство

$$\frac{1}{2}\frac{d}{dt}|\xi|^2 + \alpha\|\xi\|^2 + (\mathcal{B}(\xi + y_s, y_s), \xi) \leq -\rho\|\psi\| - (Ay_s, \xi), \quad \alpha = \min\{\nu, \nu_m\}. \tag{16}$$

Следствием неравенства (16) и неравенства (9) для квадратичного оператора $\mathcal{B}$ является оценка

$$\frac{d}{dt}|\xi|^2 + \|\xi\|^2 + \rho\|\psi\| \leq C(1 + \|\xi\|)|\xi|. \tag{17}$$

Здесь и далее через C обозначаем постоянные, зависящие только от исходных данных задачи, в частности от $\|y_s\|$, $|Ay_s|$. Из дифференциального неравенства (17) вытекают оценки

$$|\xi(t)| \leq C, \quad \int_0^T \|\xi(\tau)\|^2 d\tau \leq C, \quad \rho\int_0^T \|\psi(\tau)\| d\tau \leq C. \tag{18}$$

Полученные оценки справедливы для произвольного временно́го интервала. Покажем далее, что существует интервал $(0, T_*)$, на котором выполняются более сильные оценки. В [5], [6] существование сильного решения доказано путем получения априорных оценок для галеркинских приближений $y_k = y_k^{\varepsilon}$, определяемых из системы

$$(y_k' + Ay_k + \mathcal{B}[y_k] + \nabla\Phi_{\varepsilon}(y_k), z_j) = 0, \quad j = 1, 2, \ldots, k, \quad y_k(0) = P_k y_0. \tag{19}$$

Здесь $\{z_j\}$ – ортонормированный в H базис пространства V, состоящий из собственных элементов оператора A, $Az_j = \lambda_j z_j$, P_k – оператор проектирования в H на подпространство, натянутое на $z_1, \ldots, z_k$. Через Φ_ε обозначена регуляризация функционала Φ (см. [9])

$$\Phi_\varepsilon(y) = \inf\left\{\frac{\|y-z\|^2}{2\varepsilon} + \Phi(z);\ z \in V\right\}, \quad y \in V, \quad \varepsilon > 0.$$

Для данного функционала Φ при $y = \{u, B\}$, $z = \{v, w\}$ получаем

$$\Phi_\varepsilon(y) = \rho\begin{cases}\dfrac{1}{2\varepsilon}\|B - B_s\|^2 & \text{при} \quad \|B - B_s\| \le \varepsilon,\\[2ex] \|B - B_s\| - \dfrac{\varepsilon}{2} & \text{в противном случае.}\end{cases}$$

Соответственно, имеем

$$(\nabla\Phi_\varepsilon(y), z) = \rho\begin{cases}\dfrac{1}{\varepsilon}(A_2(B - B_s), w) & \text{при} \quad \|B - B_s\| \le \varepsilon,\\[2ex] \dfrac{1}{\|B - B_s\|}(A_2(B - B_s), w) & \text{в противном случае.}\end{cases} \tag{20}$$

Структура градиента (20) позволяет получить неравенство

$$(y_k' + Ay_k + \mathcal{B}[y_k], Ay_k - Ay_{sk}) \le 0, \tag{21}$$

где $y_{sk} = P_k y_s \in V$, $Ay_{sk} \in V$. Отметим, что $\|y_{sk}\| \le \|y_s\|$, $|Ay_{sk}| \le |Ay_s|$. Воспользовавшись неравенством (9) и неравенством Юнга, оценим в (21) слагаемые с квадратичным оператором $\mathcal{B}[y_k]$:

$$-(\mathcal{B}[y_k], Ay_k) \le C\|y_k\|^{3/2}|Ay_k|^{3/2} \le C\left(\frac{3\alpha_1^{4/3}}{4}|Ay_k|^2 + \frac{1}{\alpha_1^4}\|y_k\|^6\right),$$

$$(\mathcal{B}[y_k], Ay_{sk}) \le C\|y_k\|^{3/2}|Ay_k|^{1/2}|Ay_s| \le C|Ay_s|\left(\frac{\alpha_2^4}{4}|Ay_k|^2 + \frac{3}{4\alpha_2^{4/3}}\|y_k\|^2\right).$$

Выбирая достаточно малые значения α_1, α_2, получаем из (21) неравенство

$$\frac{d}{dt}\|y_k - y_{sk}\|^2 + |Ay_k|^2 \le C_1(1 + \|y_k\|^2 + \|y_k\|^6),$$

где постоянная C_1 зависит только от $|Ay_s|$ и от постоянной из неравенства (9). Следовательно, имеем

$$\frac{d}{dt}\|y_k - y_{sk}\|^2 + |Ay_k|^2 \le C_2(1 + \|y_k - y_{sk}\|^6), \tag{22}$$

причем C_2 зависит только от C_1, $\|y_s\|$, $|Ay_s|$. Следствием дифференциального неравенства (22) является ограниченность последовательности $\{y_k\}$ в $L^\infty(0, T_*; V)$ и Ay_k в $L^2(0, T_*; H)$ при условии, что

$$T_* < \frac{1}{2C_2(1 + \sigma)^2}, \quad \sigma = \|y_0 - y_s\|^2. \tag{23}$$

Подчеркнем, что величина T_* не зависит от параметра ρ. Полученных оценок достаточно для предельного перехода в системе (19) при $k \longrightarrow \infty$ и при $\varepsilon \longrightarrow 0$. В пределе получаем существование на достаточно малом интервале времени, не зависящем от ρ, сильного решения задачи (13), единственность которого устанавливается стандартным образом (см. [5]).

Теорема 1. *Пусть* $y_0 \in V$, $B_s \in V_2$, $A_2B_s \in H_2$. *Тогда найдется* $T_* > 0$, *не зависящее от* ρ *и такое, что на* $(0, T_*)$ *существует единственное сильное решение у вариационного неравенства* (13), *при этом справедливы оценки* (18) *и неравенство*

$$\|y(t)\| \le K,$$

где K также не зависит от ρ.

4. РАЗРЕШИМОСТЬ ЗАДАЧИ 1

Доказательство разрешимости задачи 1 основано на следующем результате.

Теорема 2. *Пусть* $y_0 \in V$, $B_s \in V_2$, $A_2B_s \in H_2$. *Существует такое значение параметра* $\rho > 0$, *зависящее только от* Ω, ν, ν_m, $|y_0|$, $\|y_0\|$, $\|B_s\|$, $|A_2B_s|$, *для которого найдется слабое решение* (13), *удовлетворяющее условию* $B|_{t=T} = B_s$.

Доказательство. Из условий на исходные данные вытекает существование (см. [5]) слабого решения $y = \{u, B\}$ неравенства (13), при этом $y \in L^\infty(0, T; H) \cap L^2(0, T; V)$, $y' \in L^1(0, T; V')$. В силу теоремы 1, на некотором интервале $(0, T_*)$, не зависящем от параметра ρ, данное решение является сильным. Тогда, умножая скалярное (15) на $z = \{0, \psi\}$, $\psi = B - B_s$, получаем равенство

$$\frac{1}{2}\frac{d}{dt}|\psi|^2 + (A_2(\psi + B_s), \psi)_0 - (u \times (\psi + B_s), \operatorname{rot}\psi)_0 = -\rho\|\psi\|, \tag{24}$$

которое выполняется п.в. на $(0, T_*)$. Из (24) следует, что на интервале $(0, T_*)$ выполняется неравенство

$$|\psi|\frac{d|\psi|}{dt} + \nu_m\|\psi\|^2 + \rho\|\psi\| \le \nu_m\|B_s\| \cdot \|\psi\| + C\|u\| \cdot \|B\| \cdot \|\psi\|. \tag{25}$$

Учтем ограниченность сильного решения, $\|u\| \le K$, $\|B\| \le K$, и выберем $\rho > \|B_s\| + CK^2$. Тогда имеем

$$|\psi|\frac{d|\psi|}{dt} + (\rho - \nu_m\|B_s\| - CK^2)\|\psi\| \le 0.$$

В силу неравенства Пуанкаре–Стеклова $\|\psi\| \ge C_\Omega|\psi|$, где постоянная C_Ω зависит только от области течения Ω, если $|\psi(t)| \ne 0$, получаем

$$\frac{d|\psi|}{dt} + \rho_1 \le 0, \quad \rho_1 = C_\Omega(\rho - \nu_m\|B_s\| - CK^2).$$

Следовательно,

$$0 \le |\psi(t)| \le |B_0 - B_s| - \rho_1 t.$$

Таким образом, $\psi|_{t=T_*} = 0$, если $\rho_1 > |B_0 - B_s|/T_*$. Покажем далее, что $\psi = 0$ на интервале (T_*, T). Отметим, что в целом на интервале $(0, T)$ справедливо неравенство

$$\frac{1}{2}\frac{d}{dt}|\psi|^2 + (A_2(\psi + B_s), \psi)_0 - (u \times (\psi + B_s), \operatorname{rot}\psi)_0 \le -\rho\|\psi\|,$$

которое выполняется в смысле распределений и может быть получено из системы галеркинских приближений (19). Оценим выражение $(u \times (\psi + B_s), \operatorname{rot}\psi)_0$ в левой части последнего неравенства. Заметим, что из первого неравенства (18) следует равномерная по $t \in (0, T)$ оценка $|u(t)|$. Следовательно,

$$|(u \times (\psi + B_s), \operatorname{rot}\psi)_0| \le |u| \cdot \|B_s\|_{L^\infty(\Omega)}\|\psi\| + C\|u\| \cdot \|\psi\|^2 \le C(1 + \|u\| \cdot \|\psi\|)\|\psi\|.$$

Поэтому имеем

$$\frac{1}{2}\frac{d}{dt}|\psi(t)|^2 \le -q(t)\|\psi\|,$$

где функция $q(t)$ оценивается снизу выражением

$$\mu(t) = \rho + \nu_m\|\psi\| - \nu_m\|B_s\| - C(1 + \|u\| \cdot \|\psi\|).$$

Заметим, что $\mu(T_*) = \rho - \nu_m\|B_s\| - C > 0$ при соответствующем выборе параметра ρ. Пусть $t_0 > T_*$ – наименьший из таких моментов времени, где $\mu(t_0) = 0$ и, соответственно, $\mu(t) > 0$ в интервале (T_*, t_0). Тогда функция $|\psi(t)|^2$ не возрастает на (T_*, t_0) и, значит, $\psi(t_0) = 0$. Следовательно, $\mu(t_0) = \mu(T_*) > 0$ и равенство $\mu(t_0) = 0$ невозможно. Таким образом, $\psi = 0$ на (T_*, T) и $B|_{t=T} = B_s$. Теорема доказана.

Теорема 2. *Пусть $u_0 \in V_1$, $B_0 \in V_2$, $B_s \in V_2$, $A_2B_s \in H_2$. Тогда задача* 1 *имеет по крайней мере одно решение.*

Доказательство теоремы 3 вытекает из теоремы 2, если в качестве управления E_c выбрать управление с обратной связью

$$E_c = -\rho\,\mathrm{sign}\,\mathrm{rot}(B - B_s),$$

где B – магнитное поле, соответствующее решению $y = \{u, B\}$ вариационного неравенства (13). Отметим при этом, что $|E_c(t)| \le \rho$.

5. СУЩЕСТВОВАНИЕ ОПТИМАЛЬНЫХ УПРАВЛЕНИЙ

Вопрос о разрешимости задач 2, 3 связан со слабой замкнутостью множества допустимых пар, т.е. множества решений задачи 1. Из (10) сразу следует энергетическое неравенство (см. [11])

$$\frac{1}{2}|y(t)|^2 + \alpha\int_0^t\|y(\tau)\|^2 d\tau \le \frac{1}{2}|y_0|^2 + \int_0^t|E_c(\tau)| \cdot \|y(\tau)\|\,d\tau. \tag{26}$$

Лемма. *Множество допустимых пар слабо замкнуто в пространстве $U \times L^2(0, T; V)$.*

Доказательство леммы основано на оценке (26) и проводится аналогично работе [11]. В силу слабой полунепрерывности снизу нормы $\|\cdot\|_U$, разрешимость задачи 2 следует теперь из непустоты множества решений задачи 1.

Теорема 4. *Пусть $u_0 \in V_1$, $B_0 \in V_2$, $B_s \in V_2$, $A_2B_s \in H_2$. Тогда существует решение задачи* 2.

Рассмотрим далее задачу 3, в которой имеется ограничение на множество управлений $\|E_c\|_U \le M$. Здесь существование решения зависит от соотношения между исходными данными и параметром M (см. [3, с. 191]). Ясно, что при достаточно больших M решение задачи 1, т.е. допустимая пара, построенная в теореме 3, удовлетворяет условию (12), поскольку

$$\|E_c\|_U \le \rho\sqrt{T}.$$

Из леммы о слабой замкнутости множества допустимых пар и слабой полунепрерывности снизу функционала J_a, определяющего работу сторонних ЭДС, следует разрешимость задачи 3 при больших M.

Пусть M_0 – нижняя грань значений M, для которых множество допустимых пар, удовлетворяющих неравенству (12), непусто. Разрешимость задачи 3 при $M = M_0$ проверяется так же, как в [3, с. 192]. Возникает вопрос о положительности значения M_0. Рассмотрим случай, когда

$$|y_0|^2 = |u_0|^2 + |B_0|^2 < |B_s|^2. \tag{27}$$

Условие (27) выполняется, например, если $y_0 = 0$, $B_s \ne 0$. Из энергетического неравенства (26) вытекает, что

$$|B_s|^2 + \alpha\int_0^T\|y(\tau)\|^2 d\tau \le |y_0|^2 + \frac{1}{\alpha}\|E_c\|_U^2 \le |y_0|^2 + \frac{1}{\alpha}M^2.$$

Следовательно, при достаточно малых M множество допустимых пар, удовлетворяющих неравенству (12), пусто и поэтому $M_0 > 0$.

Теорема 5. *Пусть $u_0 \in V_1$, $B_0 \in V_2$, $B_s \in V_2$, $A_2B_s \in H_2$ и справедливо неравенство* (27). *Найдется число $M_0 > 0$ такое, что задача* 3 *разрешима при $M \ge M_0$, а при $M < M_0$ решение задачи* 3 *не существует.*

СПИСОК ЛИТЕРАТУРЫ

1. *Тамм И.Е.* Основы теории электричества. М.: Наука, 1974.
2. *Sermange M., Temam R.* Some mathematical questions related to the MHD equations // Communs Pure and Appl. Math. 1983. V. 36. P. 635–664.
3. *Фурсиков А.В.* Оптимальное управление распределенными системами. Теория и приложения. Новосибирск: Изд-во научная книга, 1999.
4. *Chebotarev A.Yu.* Sudbifferential inverse problems for evolution Navier–Stokes systems // J. Inverse and Ill Posed Problems. 2000. V. 8. № 3. P. 275–287.
5. *Чеботарёв А.Ю., Савенкова А.С.* Вариационные неравенства в магнитной гидродинамике // Матем. заметки. 2007. Т. 82. Вып. 1. С. 135–149.
6. *Чеботарёв А.Ю.* Субдифференциальные краевые задачи магнитной гидродинамики // Дифференц. ур-ния. 2007. Т. 43. № 12. С. 1700–1709.
7. *Vazquev R., Krstic M.* Control of turbulent and magnetohydrodynamic channel flows. Boundary stabilization and state estimation. Boston: Birkhäuser, 2008.
8. *Ravindran S.S.* On the dynamics Magnetohydrodynamic systems // Nonlinear Analysis: Modelling and Cintrol. 2008. V. 13. № 3. P. 351–377.
9. *Wang L.* Optimal control of magnetohydrodynamic equations with state constrant // J. Optimizat. Theory and Appl. 2004. V. 122. № 3. P. 599–626.
10. *Barbu V., Havarneanu T., Popa C., Sritharan S.S.* Exact controllabity for the magnetohydrodynamic equations // Commus Pure and Appl. Math. V. 56. Issue 6. P. 732–783.
11. *Чеботарёв А.Ю.* Оптимальное управление в нестационарных задачах магнитной гидродинамики // Сибирский журнал индустр. матем. 2007. Т. 10. № С. 138–148.
12. *Амосова Е.В.* Оптимальное управление магнитогидродинамическим течением вязкого теплопроводного газа // Ж. вычисл. матем. и матем. физ. 2008. Т. 48. № 4. С. 623–632.
13. *Barbu V.* Analysis and control of nonlinear infinite dimensional systems New York: Acad. Press. 1993.
14. *Barbu V.* The time optimal control of Navier–Stokes equations // Systems and Control Letts. 1997. V. 30. P. 93–100.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2009, том 49, № 11, с. 2010–2019

УДК 519.676

МОДИФИКАЦИЯ ДВУХЭТАПНЫХ АЛГОРИТМОВ МЕТОДА МОНТЕ-КАРЛО НА ОСНОВЕ СВОЙСТВ СИММЕТРИИ ПЕРВОГО ЭТАПА[1)]

© 2009 г. Г. А. Михайлов*, С. А. Роженко**

*(*630090 Новосибирск, пр-т Акад. Лаврентьева, 6, ИВМ и МГ СО РАН;*
***Университетский пр-т, 2, Новосибирский гос. ун-т)*
e-mail: gam@osmf.sscc.ru; sergoj@mail.ru

Поступила в редакцию 04.05.2009 г.

Дана модификация двухэтапных алгоритмов метода Монте-Карло с учетом свойства симметрии, т.е. инвариантности, первого этапа относительно некоторого начального векторного параметра моделируемой траектории. Предлагаемая модификация состоит в формальном переносе моделирования указанного параметра на второй этап алгоритма. В "методе расщепления" это означает рандомизацию начальных точек вспомогательных траекторий. Показано, что такую рандомизацию можно улучшить, фактически применяя принцип Беллмана. Библ. 3. Фиг. 3. Табл. 4.

Ключевые слова: метод Монте-Карло, двухэтапный алгоритм, метод расщепления, оценка трудоемкости алгоритма.

1. ДВУХЭТАПНЫЕ АЛГОРИТМЫ МЕТОДА МОНТЕ-КАРЛО

Двухэтапные алгоритмы метода Монте-Карло строятся для решения задач, которые формально сводятся к вычислению интегралов вида

$$J = \int_U \left(\int_V q(u, v)\mathbf{P}_2(dv|u) \right) \mathbf{P}_1(du),$$

где $\mathbf{P}_1(\cdot)$ – вероятностная мера в U, $\mathbf{P}_2(\cdot|u)$ – вероятностная мера в V при $u \in U$.

Базовым для метода Монте-Карло является представление

$$J = \mathbf{E}\zeta, \quad \zeta = q(\xi, \eta), \quad \xi \in U, \quad \eta \in V. \tag{1}$$

Случайный вектор (ξ, η) численно моделируется в два этапа. На первом из них реализуется ξ соответственно мере $\mathbf{P}_1(\cdot)$, а на втором реализуется η соответственно условной мере $\mathbf{P}_2(\cdot|\xi)$. Искомое приближенное значение величины J из (1) строится путем осреднения выборочных значений ζ, получаемых в результате численного моделирования (см. [1]). Известно, что трудоемкость алгоритма метода Монте-Карло оценивается величиной $S = t\mathbf{D}\zeta$, где $\mathbf{D}\zeta$ – дисперсия, а t – средние арифметические затраты на одну реализацию алгоритма (см. [1]).

В литературе (см., например, [1]) рассматриваются две модификации стандартного двухэтапного алгоритма: метод математических ожиданий (ММО) и метод расщепления.

Концепция ММО тривиальна: если функция $q_1(u) = \int_V q(u, v)\mathbf{P}_2(dv|u)$ определена аналитически, то на втором этапе алгоритма просто вычисляется значение случайной величины $\zeta^{(1)} = q_1(\xi)$. Известно (см. [1]), что $\mathbf{D}\zeta^{(1)} \le \mathbf{D}\zeta$, но целесообразность ММО определяется путем сравнительного анализа соответствующей трудоемкости.

На практике чаще используется другая модификация двухэтапного моделирования – метод расщепления. Этот метод состоит в том, что для каждого значения ξ на втором этапе реализуется

[1)] Работа выполнена при частичной финансовой поддержке РФФИ (код проекта 09-01-00035-а).

K условно-независимых значений $\eta_1, \ldots, \eta_K$ соответственно условной мере $\mathbf{P}_2(\cdot|\xi)$ и вычисляется значение случайной величины

$$\zeta^{(K)} = \frac{1}{K}\sum_{k=1}^{K} q(\xi, \eta_k).$$

При этом $\mathbf{E}\zeta^{(K)} = \mathbf{E}\zeta$. Вследствие "формулы полной дисперсии" (см. [1]) имеем

$$\mathbf{D}\zeta^{(K)} = \mathbf{D}_\xi\mathbf{E}_\eta(\zeta|\xi) + \frac{1}{K}\mathbf{E}_\xi\mathbf{D}_\eta(\zeta|\xi). \tag{2}$$

Таким образом,

$$\mathbf{D}\zeta^{(K)} = A_0 + A_1/K, \quad t^{(K)} = t_0 + Kt,$$

где $A_0 = \mathbf{D}_\xi\mathbf{E}_\eta(\zeta|\xi)$, $A_1 = \mathbf{E}_\xi\mathbf{D}_\eta(\zeta|\xi)$, t_0 – средние затраты на одну реализацию первого этапа, а t_1 – второго этапа.

Оптимальное значение K равно (см., например, [1])

$$K^* = \sqrt{\frac{A_0 t_1}{A_1 t_0}}, \tag{3}$$

с точностью до перехода к "целой части". В этом же смысле оптимальная трудоемкость определяется формулой

$$S^* = \left(\sqrt{t_0 A_0} + \sqrt{t_1 A_1}\right)^2. \tag{4}$$

2. МОДИФИКАЦИИ НА ОСНОВЕ СВОЙСТВ СИММЕТРИИ ПЕРВОГО ЭТАПА

Пусть ξ однозначно определяется вектором (ξ_1, ξ_2), $\xi_1 \in U_1$, $\xi_2 \in U_2$, причем ξ_1 и ξ_2 независимы и реализация ξ_2 сравнительно малотрудоемка. Тогда эффективным может оказаться преобразование этапов моделирования, которое определяется заменой ξ на ξ_1, η на вектор (ξ_2, η) и, соответственно, U на U_1, V на $U_2 \times V$. Характерной в этом плане является задача переноса частиц с рассеянием от источника, расположенного в центре сферически-однородного шара. Пусть ξ – внутренняя часть траектории частицы (до вылета из шара), η – остальная часть траектории. Через ξ_2 обозначим случайные начальное направление частицы и угол поворота траектории вокруг соответствующего луча, а через ξ_1 – внутреннюю часть траектории, "освобожденную" от этих начальных параметров (т.е. для некоторого фиксированного начального направления, например вдоль которой-либо из координатных осей, с заданным углом поворота). Траектория ξ определяется вектором (ξ_1, ξ_2) путем вращения траектории ξ_1 от фиксированных начальных параметров к случайным параметрам ξ_2.

Особенно эффективна такая модификация в случае изотропного источника для вычисления функционала $J = \mathbf{E}q(\eta)$. Известно, что при этом распределение точки ρ вылета частицы из сферы (т.е. последней точки внутренней траектории) является равномерным на поверхности сферы. Равномерно распределенным в этом случае является также азимутальный угол φ в точке вылета, т.е. угол вращения направления вылета вокруг радиус-вектора ρ. Следовательно, здесь в качестве ξ_2 можно рассмотреть вектор (ρ, φ), а в качестве ξ_1 – угол между направлением вылета и радиус-вектором ρ. Значение ξ_1 определяется путем моделирования внутренней части траектории, а ξ_2 моделируется независимо со сравнительно малой трудоемкостью. В методе математических ожиданий переход к новым этапам, очевидно, уменьшает дисперсию, так как при этом осуществляется дополнительное осреднение по распределению ξ_2.

Следующее утверждение показывает целесообразность соответствующей модификации алгоритма расщепления.

Теорема. *Пусть требуется вычислить величину* $\mathbf{E}q(\xi_1, \xi_2, \eta)$, *где* ξ_1, ξ_2 *и* η – *некоторые последовательно моделируемые случайные величины, а* q – *вычисляемая по ним статистика.*

Пусть $\zeta_1^{(K)}$ *и* $\zeta_2^{(K)}$ *– случайные величины, моделируемые с использованием метода расщепления следующим образом:*

$$\zeta_1^{(K)} = \frac{1}{K}\sum_{k=1}^{K} q(\xi_1, \xi_2, \eta^{(k)}), \quad \zeta_2^{(K)} = \frac{1}{K}\sum_{k=1}^{K} q(\xi_1, \xi_2^{(k)}, \eta^{(k)}).$$

Тогда

$$\mathbf{D}\zeta_1^{(K)} - \mathbf{D}\zeta_2^{(K)} = \frac{K-1}{K}\mathbf{E}_{\xi_1}\mathbf{D}_{\xi_2}\{\mathbf{E}_\eta(q(\xi_1, \xi_2, \eta)\big|\xi_1, \xi_2)\big|\xi_1\} \geq 0. \tag{5}$$

Доказательство. Введем обозначение $\zeta = q(\xi_1, \xi_2, \eta)$. Заметим, что величина $\zeta_1^{(K)}$ получается из $\zeta^{(K)}$ заменой ξ на (ξ_1, ξ_2), а величина $\zeta_2^{(K)}$ – заменой ξ на ξ_1 и η на (ξ_2, η). Учтя это, из (2) получим

$$\mathbf{D}\zeta_1^{(K)} = \mathbf{D}\zeta - \frac{K-1}{K}\mathbf{E}_{(\xi_1, \xi_2)}\mathbf{D}_\eta(\zeta|\xi_1, \xi_2) =$$

$$= \mathbf{D}\zeta - \frac{K-1}{K}\mathbf{E}_{\xi_1}\mathbf{E}_{\xi_2}[\mathbf{D}_\eta(\zeta|\xi_1, \xi_2)\big|\xi_1],$$

$$\mathbf{D}\zeta_2^{(K)} = \mathbf{D}\zeta - \frac{K-1}{K}\mathbf{E}_{\xi_1}\mathbf{D}_{(\xi_2, \eta)}(\zeta|\xi_1).$$

Вычислим разницу дисперсий:

$$\mathbf{D}\zeta_1^{(K)} - \mathbf{D}\zeta_2^{(K)} = \frac{K-1}{K}\mathbf{E}_{\xi_1}\{\mathbf{D}_{(\xi_2, \eta)}(\zeta|\xi_1) - \mathbf{E}_{\xi_2}[\mathbf{D}_\eta(\zeta|\xi_1, \xi_2)\big|\xi_1]\} = \frac{K-1}{K}\mathbf{E}_{\xi_1}A,$$

где

$$A = \mathbf{E}_{(\xi_2, \eta)}(\zeta^2\big|\xi_1) - (\mathbf{E}_{(\xi_2, \eta)}(\zeta|\xi_1))^2 -$$

$$- \mathbf{E}_{\xi_2}\{\mathbf{E}_\eta(\zeta^2\big|\xi_1, \xi_2)\big|\xi_1\} + \mathbf{E}_{\xi_2}\{[\mathbf{E}_\eta(\zeta|\xi_1, \xi_2)]^2\big|\xi_1\} =$$

$$= \mathbf{E}_{\xi_2}\{[\mathbf{E}_\eta(\zeta|\xi_1, \xi_2)]^2\big|\xi_1\} - [\mathbf{E}_{(\xi_2, \eta)}(\zeta|\xi_1)]^2 =$$

$$= \mathbf{E}_{\xi_2}\{[\mathbf{E}_\eta(\zeta|\xi_1, \xi_2)]^2\big|\xi_1\} - \{\mathbf{E}_{\xi_2}[\mathbf{E}_\eta(\zeta|\xi_1, \xi_2)\big|\xi_1]\}^2 =$$

$$= \mathbf{D}_{\xi_2}\{\mathbf{E}_\eta(\zeta|\xi_1, \xi_2)\big|\xi_1\} \geq 0.$$

Отсюда получаем формулу (5). Теорема доказана.

Если затраты на моделирование ξ_2 пренебрежимо малы по сравнению с затратами на моделирование ξ_1 и η, то трудоемкость модифицированного алгоритма не превосходит трудоемкость традиционного алгоритма при одинаковом значении параметра расщепления K. Это соотношение, очевидно, усиливается при оптимальных параметрах расщепления.

В заключение заметим, что можно дополнительно использовать весовую модификацию моделирования ξ_2. Известно, что оптимальная плотность вероятности для такой модификации, согласно стохастическому аналогу принципа Беллмана (см. [1], [3]), определяется выражением

$$f_{\xi_2}(u_2) = c^* f_{\xi_2}(u_2)\sqrt{\mathbf{E}(\zeta^2\big|u)},$$

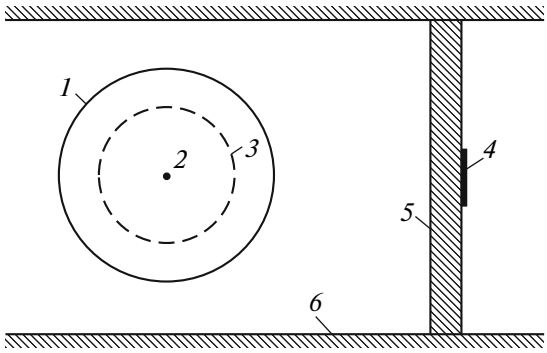
где $f_{\xi_2}(u_2)$ – плотность распределения $\mathbf{P}_2(du_2)$, а c^* – нормирующая константа. Соответствующая весовая модификация оценки такова:

$$\zeta^* = \zeta\frac{1}{c^*\sqrt{\mathbf{E}(\zeta^2\big|u)}}.$$

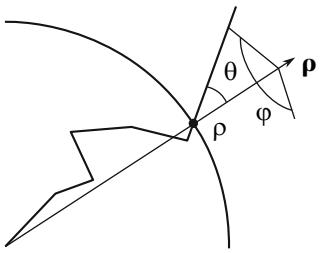
Ясно, что

$$A_0^* = \mathbf{D}\mathbf{E}(\zeta^*|\xi_1) = \mathbf{D}\mathbf{E}(\zeta|\xi_1) = A_0,$$

$$A_1^* = \mathbf{E}\mathbf{D}(\zeta^*|\xi_1) \leq \mathbf{E}\mathbf{D}(\zeta|\xi_1) = A_1.$$



Фиг. 1. Схема эксперимента: *1* – однородный шар; *2* – изотропный источник частиц; *3* – сфера расщепления; *4* – детектор частиц; *5* – защита; *6* – граница полости.



Фиг. 2.

Следовательно, согласно (4), при использовании дополнительной модификации трудоемкость оптимального расщепления, как правило, уменьшается.

Ввиду сложности оценки функции $\mathbf{E}(\zeta^2|u)$ целесообразно связывать такую модификацию с некоторым разбиением пространства U_2, введя вспомогательную дискретную координату (см. п. 3.3).

3. МОДИФИКАЦИЯ РАСЩЕПЛЕНИЯ ТРАЕКТОРИЙ В СЛУЧАЕ ИЗОТРОПНОГО ИСТОЧНИКА

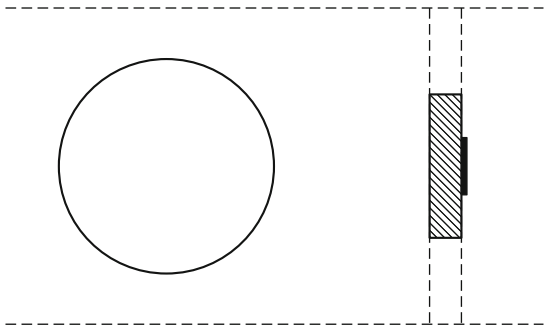
3.1. Постановка задачи

Далее рассматривается алгоритм расщепления с “вращением” траекторий в случае изотропного “точечного” источника. С целью простоты изложения этот алгоритм строится и оптимизируется для конкретной модельной задачи радиационного контроля.

Пусть в центре однородного шара, расположенного в бесконечной цилиндрической полости, находится изотропный источник частиц. На некотором расстоянии от шара находится перегородка (защита), за которой расположен детектор частиц (фиг. 1). Фиксируется среднее число частиц, попавших в детектор.

Внутри шара, защиты и вне полости поток частиц определяется математической моделью процесса переноса, а внутри полости частицы летят свободно. Традиционное “расщепление” траекторий частиц (см. [1]) реализуется на некоторой промежуточной сфере. Если частица пересекает эту сферу, то из точки пересечения в направлении пробега испускается K новых частиц, которые далее учитываются с весом $1/K$; при этом используется тот факт, что “остаток” экспоненциального распределения также распределен экспоненциально (см. [2]).

В предлагаемом модифицированном алгоритме расщепления, в отличие от стандартного алгоритма, новые частицы стартуют из точек, случайно выбранных на сфере согласно равномерному распределению. При этом сохраняется угол между траекторией и нормалью к сфере, но луч вылета частицы дополнительно вращается вокруг нормали на случайный угол.



Фиг. 3.

Для траектории частицы, пересекающей сферу расщепления (фиг. 2), введем следующие обозначения:

ξ_1 – угол θ между направлением пробега и нормалью к сфере в точке расщепления ρ либо искусственно введенное значение (ИВЗ), если частица поглотилась до пересечения сферы;

ξ_2 – вектор (ρ, φ), состоящий из координат точки расщепления и угла поворота направления частицы вокруг радиус-вектора ρ либо ИВЗ, если частица не долетела до сферы расщепления;

η – часть траектории после пересечения сферы расщепления;

$\zeta = q(\xi_1, \xi_2, \eta)$ – случайная величина, фиксируемая детектором, т.е. $\mathbf{E}\zeta$ – вероятность попадания частицы в детектор.

Стандартное расщепление можно определить как однократную реализацию пары (ξ_1, ξ_2) и K-кратную реализацию η. Модифицированное расщепление определяется как однократная реализация ξ_1 и K-кратная реализация пары (ξ_2, η). В разд. 2 было показано, что дисперсия случайной величины $\zeta^{(k)}$ для модифицированного расщепления меньше, чем для стандартного. Поскольку затраты на моделирование ξ_2 заметно меньше затрат на моделирование ξ_1 и η, затраты на моделирование при использовании "вращения" возрастают несущественно и рассматриваемая модификация уменьшает трудоемкость оценки.

3.2. Оценка трудоемкости на основе упрощенной модели

Далее оценим сравнительную трудоемкость модифицированного алгоритма расщепления в данной задаче. Для этого рассмотрим упрощенную постановку задачи. Оставим только цилиндрическую часть защиты непосредственно перед детектором (фиг. 3). Частицы, не попавшие в нее, будем игнорировать. Расщепление будем проводить на поверхности шара и заменим введенные в разд. 2 величины ξ_1, ξ_2, η на зависимые бернуллиевы случайные величины ξ_1, ξ_2, η:

$\xi_1 = 1$, если частица вылетела из шара;

$\xi_2 = 1$, если частица, вылетев из шара, долетела до защиты;

$\eta = 1$, если частица попала в детектор.

Заметим, что при расщеплении на поверхности сферы величина ξ_1 будет иметь одинаковое значение для всех новых частиц. При этом обычный метод соответствует однократной реализации пары (ξ_1, ξ_2) и K-кратной реализации η, а в модифицированном методе ξ_1 реализуется один раз и K раз реализуется пара (ξ_2, η).

Найдем аналитические формулы для трудоемкостей обычного и модифицированного методов при выбранных параметрах моделирования. Есть

$$\mathbf{P}(\xi_1 = 1) = p_0, \quad \mathbf{P}(\xi_2 = 1 | \xi_1 = 1) = p_1, \quad \mathbf{P}(\eta = 1 | \xi_2 = 1) = p_2, \quad p = p_0 p_1 p_2.$$

Через t_0, t_1 и t_2 обозначим затраты на моделирование на соответствующих этапах. Для случая бернуллиевых случайных величин легко находятся параметры $A_0, A_1, \bar{t}_0, \bar{t}_1$: в традиционном методе

$$A_0 = pp_2(1 - p_0 p_1), \quad A_1 = p(1 - p_2), \quad \bar{t}_0 = t_0 + p_0 t_1, \quad \bar{t}_1 = p_0 t_2;$$

в модифицированном методе

$$A_0 = pp_1p_2(1-p_0), \quad A_1 = p(1-p_1p_2), \quad \bar{t}_0 = t_0, \quad \bar{t}_1 = p_0(t_1+t_2).$$

Далее можно находить оптимальные значения параметра расщепления и трудоемкости по формулам (3) и (4).

3.3. Выборка "вращения" по важности

Для дальнейшего уменьшения трудоемкости воспользуемся выборкой по важности. Разобьем сферу расщепления на две области плоскостью, перпендикулярной к оси симметрии. Из ближайшей к детектору области I_1 частицы попадают в детектор с большей вероятностью, чем из области I_2. Обозначим через s_1, s_2 – площади областей I_1 и I_2, $r_1 = s_1/(s_1+s_2)$, γ – соответствующая бернуллиева величина, т.е.

$$\gamma = \begin{cases} 1, & \mathbf{P} = r_1, \\ 0, & \mathbf{P} = 1-r_1 = r_2. \end{cases}$$

Базовую случайную оценку можно определить формулой

$$\zeta = \gamma\zeta^{(1)} + (1-\gamma)\zeta^{(2)},$$

где $\zeta^{(i)}$ – случайное значение $q(\xi_1, \xi_2, \eta)$ при условии, что частица стартует из области I_1, $i = 1, 2$. Таким образом, величина γ является индикатором "старта".

Введем для областей I_1, I_2 модифицирующие "ценностные множители" q_1 и q_2, которые меняют распределение индикатора старта по формуле

$$\gamma = \begin{cases} 1, & p = q_1r_1/a, \\ 0, & p = q_2r_2/a, \end{cases} \qquad a = q_1r_1 + q_2r_2.$$

Будем моделировать случайную величину

$$\zeta'' = \frac{1}{K}\sum_{i=1}^{K}\zeta_i', \quad \text{где} \quad \zeta' = \gamma'\zeta^{(1)}\frac{a}{q_1} + (1-\gamma')\zeta^{(2)}\frac{a}{q_2}.$$

При $q_1 = q_2$ она совпадает с ζ.

Поскольку $\mathbf{E}(\zeta''|\xi_1) = \mathbf{E}(\zeta'|\xi_1) = r_1\mathbf{E}(\zeta^{(1)}|\xi_1) + r_2\mathbf{E}(\zeta^{(2)}|\xi_1) = \mathbf{E}(\zeta|\xi_1)$, то

$$\mathbf{E}\zeta'' = \mathbf{E}_{\xi_1}\mathbf{E}(\zeta''|\xi_1) = \mathbf{E}\zeta.$$

Пусть $A_0 = \mathbf{D}_{\xi_1}\mathbf{E}(\zeta|\xi_1)$, $A_1 = \mathbf{E}_{\xi_1}\mathbf{D}(\zeta|\xi_1)$. Тогда

$$\mathbf{D}\zeta'' = \mathbf{D}_{\xi_1}\mathbf{E}(\zeta''|\xi_1) + \mathbf{E}_{\xi_1}\mathbf{D}(\zeta''|\xi_1) = A_0 + \mathbf{E}_{\xi_1}\mathbf{D}\left(\frac{1}{K}\sum_{i=1}^{K}\zeta_i'\Big|\xi_1\right) = A_0 + \frac{1}{K}\mathbf{E}_{\xi_1}\mathbf{D}(\zeta'|\xi_1).$$

Учитывая, что $\gamma'(1-\gamma') = 0$, получаем

$$\mathbf{E}_{\xi_1}\mathbf{E}((\zeta')^2|\xi_1) = \mathbf{E}\left(\gamma'\zeta^{(1)}\frac{a}{q_1}\right)^2 + \mathbf{E}\left((1-\gamma')\zeta^{(2)}\frac{a}{q_2}\right)^2 =$$

$$= \frac{q_1r_1}{a}\frac{a^2}{q_1^2}\mathbf{E}(\zeta^{(1)})^2 + \frac{q_2r_2}{a}\frac{a^2}{q_2^2}\mathbf{E}(\zeta^{(2)})^2 =$$

$$= \frac{ar_1}{q_1}\mathbf{E}(\zeta^{(1)})^2 + \frac{ar_2}{q_2}\mathbf{E}(\zeta^{(2)})^2,$$

$$\mathbf{E}_{\xi_1}(\mathbf{E}(\zeta'|\xi_1))^2 = \mathbf{E}_{\xi_1}(\mathbf{E}(\zeta|\xi_1))^2 = \mathbf{E}\zeta^2 - \mathbf{E}_{\xi_1}\mathbf{D}(\zeta|\xi_1) = \mathbf{E}\zeta^2 - A_1.$$

Следовательно,

$$\mathbf{D}\zeta'' = A_0 + \frac{1}{K}\left[A_1 - \mathbf{E}\zeta^2 + \frac{ar_1}{q_1}\mathbf{E}(\zeta^{(1)})^2 + \frac{ar_2}{q_2}\mathbf{E}(\zeta^{(2)})^2\right]. \quad (6)$$

Предположим, что затраты на моделирование ζ'' не зависят от q_1, q_2. Тогда, минимизируя (6) по $x = q_1r_1/a$, получаем

$$\frac{x^2}{(1-x)^2} = \frac{r_1^2\mathbf{E}(\zeta^{(1)})^2}{r_2^2\mathbf{E}(\zeta^{(2)})^2}.$$

Отсюда следует, что

$$\frac{q_1r_1}{q_2r_2} = \frac{r_1}{r_2}\sqrt{\frac{\mathbf{E}(\zeta^{(1)})^2}{\mathbf{E}(\zeta^{(2)})^2}} = \frac{r_1}{r_2}\sqrt{\frac{\mathbf{E}(\zeta^2|\gamma=1)}{\mathbf{E}(\zeta^2|\gamma=2)}}.$$

Таким образом, оптимальные значения q_1, q_2 определяются по формулам

$$q_1 = \sqrt{\mathbf{E}(\zeta^2|\gamma=1)}, \quad q_2 = \sqrt{\mathbf{E}(\zeta^2|\gamma=2)}, \quad (7)$$

т.е. и в дискретном варианте оптимизации “вращения” реализуется стохастический вариант принципа Беллмана (см. [3] и заключение разд. 2).

Учитывая, что $a = r_1q_1 + r_2q_2$, получаем выражение дисперсии для оптимальных весов:

$$\mathbf{D}\zeta'' = A_0 + \frac{1}{K}[A_1 - \mathbf{E}\zeta^2 + (r_1q_1 + r_2q_2)^2]. \quad (8)$$

Имея оценки оптимальных весов q_1 и q_2, можно использовать формулы оценки значения оптимального параметра расщепления на основе упрощенной модели, подставив в них $A_1 - \mathbf{E}\zeta^2 + (r_1q_1 + r_2q_2)^2$ вместо A_1.

Заметим, что вероятности r_1, r_2 могут варьироваться. Их значения, как и значения модельных величин p_0, p_1, p_2, t_0, t_1, t_2, влияют лишь на параметр K и тем самым на трудоемкость вычислительного алгоритма с сохранением состоятельности результативных оценок. Отметим также, что в довольно большой окрестности оптимального варианта трудоемкость может мало отличаться от минимальной; это показывают тестовые результаты (см. п. 4.3).

4. ЧИСЛЕННЫЕ ЭКСПЕРИМЕНТЫ

4.1. Описание алгоритма численного моделирования

Рассмотрим стационарный односкоростной процесс переноса излучения без деления с постоянными сечениями рассеяния и поглощения. Он представляет собой марковскую цепь столкновений частицы с элементами вещества. В результате столкновения может произойти поглощение или рассеяние, которое меняет направление движения частицы. Рассеяние предполагается изотропным.

Длина l свободного пробега частицы между двумя последовательными столкновениями распределена с плотностью $f(t) = \sigma e^{-\sigma t}$, $t > 0$. Вероятность того, что при столкновении частица поглотится, равна $\mathbf{P}_c = \sigma_c/\sigma$.

Моделирование очередного пробега частицы выполняется следующим образом:

1) выбирается новое направление ω движения частицы;

2) определяется длина l свободного пробега соответственно плотности вероятности $f(t)$ по формуле $l = -\ln(\alpha/\sigma)$, где $\alpha \in (0, 1)$ – стандартное случайное число (см. [1]); если частица на пробеге пересекает границу сред, то она фиксируется на этой границе, а дополнительная длина свободного пробега выбирается в соответствии с полным сечением σ новой среды, при этом используется тот факт, что “остаток” экспоненциального распределения также распределен экспоненциально (см. [2]);

3) реализуется поглощение с вероятностью $\mathbf{P}_c$;

4) в случае выживания определяется новая точка столкновения по формуле $\mathbf{r} = \mathbf{r}' + l\omega$.

Вследствие изотропности рассеяния пересчет координат вектора ω направления пробега осуществляется по формулам (см. [1])

$$\omega_1 = \sqrt{1-\mu^2}\cos\varphi, \quad \omega_2 = \sqrt{1-\mu^2}\sin\varphi, \quad \omega_3 = \mu, \tag{9}$$

где $\mu = 1 - 2\alpha_1$, $\varphi = 2\pi\alpha_2$, причем α_1, α_2 равнометно распределены в (0, 1) и независимы.

4.2. Детали реализации алгоритма

Моделирование выполняется соответственно схеме эксперимента, изображенной на фиг. 1, со следующими параметрами: радиус шара 10, радиус полости 20, толщина защиты 1, диаметр детектора 2, расстояние от центра шара до защиты 20.

Были выбраны следующие значения сечений: для шара $\sigma = 1$, $\sigma_s = 0.9$, для стенок полости и защиты $\sigma = 1$, $\sigma_s = 0.5$.

Был использован "прыгающий" мультипликативный датчик псевдослучайных чисел (little frog, см. [1]) с модулем 2^{40} и множителем 5^{17}. При переходе к моделированию новой случайной траектории делается фиксированный "прыжок" вдоль последовательности псевдослучайных чисел из расчета, что на одну траекторию достаточно 2^{15} случайных чисел. Таким образом коррелируются случайные оценки для разных вариантов задачи.

Расщепление траектории осуществляется при пересечении частицей границы шара. Радиус-векторы точек старта новых частиц при расщеплении выбираются изотропно, с помощью формулы (9).

После выбора точки вылета новой частицы выбирается угол $\varphi = 2\pi\alpha$ поворота траектории в касательной плоскости.

Координаты направления вылета вычисляются по формулам (см. [1, с. 261])

$$\omega_1 = \omega_1'\mu - (\omega_2'\sin\varphi + \omega_1'\omega_3'\cos\varphi)\sqrt{\frac{1-\mu^2}{1-\omega_3'^2}},$$

$$\omega_2 = \omega_2'\mu + (\omega_1'\sin\varphi - \omega_2'\omega_3'\cos\varphi)\sqrt{\frac{1-\mu^2}{1-\omega_3'^2}},$$

$$\omega_3 = \omega_3'\mu + (1-\omega_3'^2)\cos\varphi\sqrt{\frac{1-\mu^2}{1-\omega_3'^2}},$$

где $(\omega_1', \omega_2', \omega_3') = \rho/|\rho|$, $\mu = \cos\theta = (\omega, \omega')$ (см. фиг. 2), φ – угол между плоскостями $(\omega'; Oz)$ и (ω', ω). Угол φ является изотропным и моделируется по формуле $\varphi = 2\pi\alpha$.

4.3. Анализ результатов

Для упрощенной модели были численно получены значения параметров

$$t_0 = 0.00117, \quad t_1 = 0.00101, \quad t_2 = 0.00297, \quad p_0 = 0.039, \quad p = 1.044 \times 10^{-5}.$$

Параметр p_2 зависит от радиуса R защиты в упрощенной модели. Оптимальные "ценности" $q_1 = 0.0397$ и $q_1 = 0.00552$ были получены при $r_1 = 0.15$ и $r_2 = 0.85$.

В табл. 1 приведено сравнение традиционного и модифицированного методов расщепления на основе упрощенной модели. Через K_1, K_2, K_3 обозначены оптимальные значения параметра расщепления для традиционного, модифицированного методов и для комбинации модифицированного метода расщепления с выборкой по важности соответственно. Через S_i обозначены соответствующие трудоемкости методов: S_0 – алгоритма без расщепления, S_1 – традиционного метода расщепления, S_2 – модифицированного метода расщепления, S_3 – комбинации модифицированного метода расщепления с выборкой по важности.

Таблица 1

R	p_1	p_2	K_1	K_2	K_3	S_0/S_1	S_1/S_2	S_2/S_3
1	0.00063	0.43	12	541	352	1.90	31.0	2.08
1.05	0.00069	0.39	13	541	352	2.06	28.6	2.08
1.1	0.00076	0.35	14	541	352	2.22	26.5	2.08
1.2	0.00090	0.30	15	541	352	2.56	23.0	2.08

Таблица 2. Результаты расчетов по традиционному методу при $N = 5000000$

K	$\mathbf{E}\zeta$	δ	t, мс	S
1	0.940×10^{-5}	1.37×10^{-6}	0.0101	9.54×10^{-8}
5	1.136×10^{-5}	1.04×10^{-6}	0.0107	5.85×10^{-8}
8	1.118×10^{-5}	9.74×10^{-7}	0.0111	5.26×10^{-8}
10	1.124×10^{-5}	9.64×10^{-7}	0.0113	5.27×10^{-8}
15	$\mathbf{1.080 \times 10^{-5}}$	$\mathbf{9.18 \times 10^{-7}}$	**0.0120**	$\mathbf{5.04 \times 10^{-8}}$
20	1.077×10^{-5}	9.08×10^{-7}	0.0126	5.19×10^{-8}
25	1.067×10^{-5}	8.90×10^{-7}	0.0132	5.23×10^{-8}
30	1.051×10^{-5}	8.72×10^{-7}	0.0138	5.26×10^{-8}
50	1.048×10^{-5}	8.69×10^{-7}	0.0163	6.16×10^{-8}
100	1.032×10^{-5}	8.41×10^{-7}	0.0226	7.98×10^{-8}
300	1.045×10^{-5}	8.47×10^{-7}	0.0475	1.70×10^{-7}

Таблица 3. Результаты расчетов по модифицированному методу при $N = 1000000$

K	$\mathbf{E}\zeta$	δ	t, мс	S
100	1.047×10^{-5}	3.27×10^{-7}	0.0257	2.75×10^{-9}
200	1.061×10^{-5}	2.37×10^{-7}	0.0413	2.33×10^{-9}
300	1.033×10^{-5}	1.93×10^{-7}	0.0569	2.12×10^{-9}
400	1.045×10^{-5}	1.69×10^{-7}	0.0725	2.08×10^{-9}
450	$\mathbf{1.047 \times 10^{-5}}$	$\mathbf{1.61 \times 10^{-7}}$	**0.0803**	$\mathbf{2.07 \times 10^{-9}}$
500	1.053×10^{-5}	1.54×10^{-7}	0.0880	2.09×10^{-9}
600	1.052×10^{-5}	1.42×10^{-7}	0.1036	2.09×10^{-9}
800	1.050×10^{-5}	1.26×10^{-7}	0.1349	2.12×10^{-9}
2000	1.044×10^{-5}	8.87×10^{-8}	0.3220	2.53×10^{-9}

Таблица 4. Результаты расчетов по комбинации модифицированного метода с выборкой по важности при $N = 1000000$

K	$\mathbf{E}\zeta$	δ	t, мс	S
100	1.073×10^{-5}	2.16×10^{-7}	0.0267	1.25×10^{-9}
200	1.055×10^{-5}	1.61×10^{-7}	0.0431	1.12×10^{-9}
250	1.060×10^{-5}	1.45×10^{-7}	0.0513	1.08×10^{-9}
300	1.057×10^{-5}	1.34×10^{-7}	0.0596	1.07×10^{-9}
350	$\mathbf{1.048 \times 10^{-5}}$	$\mathbf{1.24 \times 10^{-7}}$	**0.0678**	$\mathbf{1.05 \times 10^{-9}}$
400	1.051×10^{-5}	1.18×10^{-7}	0.0761	1.06×10^{-9}
500	1.048×10^{-5}	1.07×10^{-7}	0.0925	1.06×10^{-9}
600	1.047×10^{-5}	1.01×10^{-7}	0.1089	1.10×10^{-9}

Расчеты методом Монте-Карло дали следующие оптимальные значения:

$$K_1 = 15, \quad K_2 = 450, \quad K_3 = 350,$$

при которых

$$S_0/S_1 = 1.89, \quad S_1/S_2 = 24.3, \quad S_2/S_3 = 1.98.$$

Таким образом, упрощенная модель дала удовлетворительную оценку для $R \in (1, 1.2)$. Полный же выигрыш трудоемкости при использовании модифицированного расщепления по сравнению с расчетом без расщепления равен $S_0/S_3 = 90.9$.

В табл. 2–4 приведены результаты расчетов методом Монте-Карло. Использованы обозначения: N – количество моделируемых траекторий, K – количество частиц расщепления, $\mathbf{E}\zeta$ и δ – оценки среднего реализуемой случайной величины (вероятности попадания частиц в детектор) и соответствующей среднеквадратической погрешности, t – время моделирования одной случайной траектории, S – трудоемкость. Жирным шрифтом в таблицах выделены оптимальные варианты.

СПИСОК ЛИТЕРАТУРЫ

1. *Михайлов Г.А., Войтишек А.В.* Численное статистическое моделирование. Методы Монте-Карло / Учебное пособие. М.: Издат. центр “Академия”, 2006.
2. *Боровков А.А.* Теория вероятностей. М.: Наука, 1986.
3. *Mikhailov G.A.* Recurrent formulae and the Bellman principle in the Monte Carlo method // Rus. J. Numer. Analys. and Math. Modelling. 1994. № 3. P. 281–300.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2009, том 49, № 11, с. 2020–2040

УДК 519.71

ОБ ОДНОМ МЕТОДЕ ПОИСКА ПЛАВНО МЕНЯЮЩИХСЯ ЗАКОНОМЕРНОСТЕЙ В ПУЧКАХ ВРЕМЕННЫХ РЯДОВ

© 2009 г. Н. В. Филипенков

(119991 Москва, ул. Вавилова, 40, ВЦ РАН)

e-mail: filipenkov@mail.ru

Поступила в редакцию 18.11.2008 г.
Переработанный вариант 03.06.2009 г.

Предлагается новый подход к поиску закономерностей в пучках нестационарных k-значных временных рядов. Этот подход позволяет выявлять закономерности, которые подвергаются "плавным" структурным изменениям с течением времени. Для определения подобного рода изменений предлагается мера сходства закономерностей и описывается ее применение как веса на графе закономерностей. Найденные закономерности могут быть использованы как для прогнозирования следующих элементов пучка временных рядов, так и для анализа явления, описанного пучком, и для моделирования явления. Это делает возможным применение предложенного алгоритма в широком пласте задач прогнозирования временных рядов, а также в задачах изучения и описания процессов, которые могут быть представлены пучком временных рядов. Описаны способы непосредственного практического использования разработанных методов анализа и прогноза временных рядов, и рассматривается применение методов для краткосрочного прогнозирования модельных рядов и реального пучка временных рядов, составленного из курсов акций компаний, имеющих сходную область деятельности. Библ. 23. Фиг. 5. Табл. 8.

Ключевые слова: временные ряды, интеллектуальный анализ данных, мера сходства закономерностей, вычислительный алгоритм.

1. ВВЕДЕНИЕ

В настоящее время анализ временных рядов является крайне актуальной задачей в различных сферах деятельности человека: медицине, экономике, физике, кибернетике. При этом часто возникает необходимость в исследовании сразу нескольких процессов или показателей одного процесса в их взаимосвязи и взаимовлиянии – изучения пучков временных рядов. Пучки временных рядов могут, например, описывать процессы жизнедеятельности человека, стоимость акций на бирже, курсы валют и т.д. Пучок временных рядов, учитывая множество характеристик явления, позволяет описать процесс или систему процессов наиболее полно, что, в свою очередь, позволяет сделать более точный прогноз. Возможность системного анализа процессов, их более точного описания определила высокий интерес исследователей к изучению пучков временных рядов (см. [1]–[6]).

Изучение пучков временных рядов подразумевает не только прогнозирование значений рядов, но предполагает решение задачи в рамках области интеллектуального анализа данных. Это означает, что необходимо выявить и описать закономерности, определяющие поведение временных рядов. Найденные закономерности могут быть представлены в виде уравнений (см. [2], [3], [5]), ассоциативных (см. [7], [8]), "эпизодических" (см. [9]) или прочих правил (см. [6], [10], [11]).

В большинстве реальных задач измерения проводятся в дискретные моменты времени, поэтому во многих работах рассматриваются дискретные временные ряды. При этом в работах, посвященных поиску правил (см. [6], [7], [10], [11]), рассматриваются пучки, где значениями временных рядов являются элементы некоторого конечного алфавита. Для поиска правил в "непрерывных" временных рядах используются методы дискретизации или символьного представления (см. [10], [12]).

Пучок временных рядов отражает характеристики явления во времени, но само явление может меняться с течением времени. Нестационарными в общем смысле называются временные ряды, свойства которых непостоянны во времени. Такие ряды составляют большинство, так как почти все явления под воздействием различных факторов претерпевают изменения. Для анализа

нестационарных временных рядов был предложен целый ряд адаптивных методов: экспоненциальное сглаживание и его модификации (см. [13], [14]), модели семейства ARIMA (см. [2]), модели семейства ARCH (см. [3], [15]), множественная регрессия (см. [5]), модели, основанные на использовании спектральных характеристик рядов (см. [16]–[18]).

В настоящей работе рассматривается задача поиска закономерностей в пучках дискретных нестационарных временных рядов с конечным алфавитом значений. В работе предложен подход, который позволяет учитывать плавное изменение закономерностей с течением времени. Для определения плавного изменения вводится понятие меры сходства закономерностей, указывающей на близкие закономерности. При этом задача поиска закономерностей рассматривается как задача интеллектуального анализа данных, что позволяет в явном виде описывать найденные закономерности.

1.1. Основные определения

Пучком временных рядов $\mathfrak{S}$ называется совокупность взаимосвязанных временных рядов S_i, $i \in \{1, 2, ..., N\}$. Каждый ряд S_i представляет собой последовательность чисел конечнозначной логики E_{k_i}. Каждому элементу ряда соответствует некоторый момент времени, и эти моменты времени для всех рядов одинаковы. Поэтому одинакова и длина всех рядов, которая обозначается через T. Таким образом, пучок временных рядов $\mathfrak{S}$ есть матрица размера $N \times T$, где элемент i-й строки принадлежит множеству E_{k_i}. Значения ряда S_i, $i \in \{1, 2, ..., N\}$, в момент времени $t \in \{1, 2, ..., T\}$ обозначим через $a(i, t)$ или $a_{i,t}$.

Маской ω на прямоугольнике $N \times \Delta$ назовем булеву матрицу размера $N \times \Delta$ (здесь параметр Δ определяет максимальный отступ по времени). Число единиц в маске ω будем называть *мощностью* маски и обозначать через $\|\omega\|$. Элемент маски, находящийся в i-й строке и j-м столбце, будем обозначать через $\omega(i, j)$ или $\omega_{i,j}$. *Закономерностью* R назовем набор (p, ω, f) с такими особенностями:

1) число $p \in \{1, 2, ..., N\}$ указывает на целевой ряд (ряд, значения которого определяются закономерностью R);

2) маска ω указывает на значения рядов, являющиеся аргументами функции f;

3) частично определенная функция f задает зависимость значений целевого ряда от переменных, на которые указывает маска ω:

$$f\colon E_{k_{i_1}} \times \ldots \times E_{k_{i_{\|\omega\|}}} \longrightarrow E_{k_p} \cup \{\lambda\},$$

где $\omega(i_1, j_1)$, ..., $\omega(i_{\|\omega\|}, j_{\|\omega\|})$ – единичные элементы матрицы ω, p – номер целевого ряда, символ λ обозначает, что f не определена на соответствующем наборе значений переменных.

Если значения всех рядов представляют собой числа k-значной логики ($E_{k_1} = ... = E_{k_N} = E_k$), то функция f принадлежит множеству P_k^* всех частично определенных функций k-значной логики.

Задача состоит в поиске закономерностей. Найденные закономерности позволяют прогнозировать значения целевого ряда, делать выводы о характере зависимостей между рядами, моделировать целевой ряд или весь пучок временных рядов.

1.2. Алгоритм поиска постоянных закономерностей

Рассмотрим один из подходов к поиску закономерностей в пучках временных рядов, который предполагает отсутствие изменений в закономерностях с течением времени. Для простоты выкладок предположим, что $k_i = k$, $i = 1, 2, ..., N$, т.е. будем рассматривать N k-значных временных рядов длины T.

1.2.1. Построение закономерности по маске. Пусть заданы пучок временных рядов $\mathfrak{S} \in E_k^{N \times T}$, целевой ряд $p \in \{1, 2, ..., N\}$ и маска $\omega \in E_2^{N \times \Delta}$ с единичными элементами $\omega(i_1, j_1)$, ..., $\omega(i_{\|\omega\|}, j_{\|\omega\|})$, упорядоченными лексикографически. Тогда при фиксированном целевом ряде p на основании $\mathfrak{S}$ и ω строится множество пар (α_t, v_t), $t \in \{1, 2, ..., T - \Delta\}$, где

$$\alpha_t = \{a(i_1, t + j_1 - 1), \ldots, a(t_{\|\omega\|}, t + j_{\|\omega\|} - 1)\} \in E_k^{\|\omega\|}, \quad v_t = a(p, t + \Delta) \in E_k.$$

Пусть α – произвольный набор из $E_k^{\|\omega\|}$ и $v \in E_k$. *Частотой* $\nu(\alpha, v, p, \omega, \mathfrak{S})$ называется число раз, которое пара (α, v) встречается среди пар (α_t, v_t), $t \in \{1, 2, \ldots, T-\Delta\}$. *Достоверностью* $\mathrm{Conf}_{\mathrm{set}}(\alpha, v, p, \omega, \mathfrak{S})$ *значения* v *на наборе* α *при маске* ω *на пучке* $\mathfrak{S}$ называется частота появления значения v на наборе α:

$$\mathrm{Conf}_{\mathrm{set}}(\alpha, v, p, \omega, \mathfrak{S}) = \frac{\nu(\alpha, v, p, \omega, \mathfrak{S})}{\sum_{i=0}^{k-1}\nu(\alpha, i, p, \omega, \mathfrak{S})}.$$

Поддержкой $\mathrm{Supp}_{\mathrm{set}}(\alpha, p, \omega, \mathfrak{S})$ набора α при маске ω на пучке $\mathfrak{S}$ называется частота появления набора α в исследуемой выборке:

$$\mathrm{Supp}_{\mathrm{set}}(\alpha, p, \omega, \mathfrak{S}) = \frac{\sum_{i=0}^{k-1}\nu(\alpha, i, p, \omega, \mathfrak{S})}{T-\Delta}.$$

Термины “поддержка” (support) и “достоверность” (confidence) были введены как меры эффективности ассоциативных правил (см. [7], [8], [19]).

Достоверностью $\mathrm{Conf}(R, \mathfrak{S})$ *закономерности* $R = (p, \omega, f)$ *на пучке временных рядов* $\mathfrak{S}$ называется доля правильных прогнозов закономерности R на пучке временных рядов $\mathfrak{S}$:

$$\mathrm{Conf}(R, \mathfrak{S}) = \mathrm{Conf}(p, \omega, f, \mathfrak{S}) = \frac{\sum_{\alpha \in E_k^{\|\omega\|}}\nu(\alpha, f(\alpha), p, \omega, \mathfrak{S})}{T-\Delta} =$$

$$= \sum_{\alpha \in E_k^{\|\omega\|}} \mathrm{Conf}(\alpha, f(\alpha), p, \omega, \mathfrak{S})\mathrm{Supp}(\alpha, p, \omega, \mathfrak{S}).$$

Поиск постоянных закономерностей на пучке временных рядов $\mathfrak{S}$, определяющих поведение целевого ряда $p = 1, 2, \ldots, N$, состоит в последовательном переборе масок. Для каждой маски ω определяется закономерность $R = (p, \omega, f)$, которая наиболее точно описывает исследуемый пучок временных рядов $\mathfrak{S}$, т.е. которая максимизирует достоверность закономерности R. Легко видеть, что при фиксированном целевом ряде p и маске ω оптимальная закономерность на пучке временных рядов $\mathfrak{S}$ строится путем выбора функции f.

Теорема 1. *При фиксированном целевом ряде* p *и маске* ω *максимальную достоверность* $\mathrm{Conf}(R, \mathfrak{S})$ *на пучке временных рядов* $\mathfrak{S}$ *имеет закономерность* $R_0 = (p, \omega, f_0)$, *где значения* $f_0(\alpha)$ *выбираются следующим образом:*

$$f_0(\alpha) = \arg\max_{v \in E_k}\nu(\alpha, v, \omega, \mathfrak{S}).$$

Доказательство. Рассмотрим цепочку соотношений

$$\max_f \mathrm{Conf}(R, \mathfrak{S}) = \max_f \mathrm{Conf}(p, \omega, f, \mathfrak{S}) = \max_f \frac{\sum_{\alpha \in E_k^{\|\omega\|}}\nu(\alpha, f(\alpha), p, \omega, \mathfrak{S})}{T-\Delta} \le$$

$$\le \frac{\sum_{\alpha \in E_k^{\|\omega\|}}\max_f \nu(\alpha, f(\alpha), p, \omega, \mathfrak{S})}{T-\Delta} \le \frac{\sum_{\alpha \in E_k^{\|\omega\|}}\max_{v \in E_k} \nu(\alpha, v, p, \omega, \mathfrak{S})}{T-\Delta} =$$

$$= \frac{\sum_{\alpha \in E_k^{\|\omega\|}}\nu(\alpha, f(\alpha), p, \omega, \mathfrak{S})}{T-\Delta} = \mathrm{Conf}(R_0, \mathfrak{S}).$$

Из утверждения следует, что при обучении для каждого набора α должны выбираться значения v, максимизирующие достоверность $\mathrm{Conf}_{\mathrm{set}}(\alpha, v, p, \omega, \mathfrak{S})$ значения v на наборе α. При этом отметим, что при некотором фиксированном наборе α значения $\mathrm{Conf}_{\mathrm{set}}(\alpha, v, p, \omega, \mathfrak{S})$ для различных $v \in E_k$ представляют собой вектор оценок. Алгоритм выбора оптимального значения $f_0(\alpha) = \arg\max_{v \in E_k}\nu(\alpha, v, p, \omega, \mathfrak{S})$ представляет собой решающее правило. Таким образом, данное представление позволяет применять к построенным таким образом закономерностям конструкции

алгебраического подхода, предложенные в работах Ю.И. Журавлёва (см. [20]) и К.В. Рудакова (см. [21]).

Если $\mathrm{Supp}_{\mathrm{set}}(\alpha, p, \omega, \mathfrak{S}) = 0$, то функция f является неопределенной на наборе α. В таких случаях будем обозначать $f(\alpha) = \lambda$.

Заметим, что при выборе наилучшей закономерности может рассматриваться достоверность $\mathrm{Conf}(R, \mathfrak{S}_{\mathrm{valid}})$ не на пучке временных рядов, на котором происходило обучение, а на независимом пучке временных рядов $\mathfrak{S}_{\mathrm{valid}}$, который используется для валидации закономерностей. Такой подход обычно позволяет избежать переобучения.

С целью выбора только "полезных" закономерностей иногда задаются пороговые значения достоверности $\mathrm{Conf}_{\min}$ и поддержки $\mathrm{Supp}_{\min}$, при которых считается, что функция определена на данном наборе:

$$f(\alpha) = \begin{cases} v_m,\ v_m = \underset{v \in E_k}{\arg\max} \nu(\alpha, v, p, \omega, \mathfrak{S}), & \text{если} \quad \mathrm{Conf}_{\mathrm{set}}(\alpha, v_m, p, \omega, \mathfrak{S}) \geq \mathrm{Conf}_{\min} \\ & \text{и} \quad \mathrm{Supp}_{\mathrm{set}}(\alpha, p, \omega, \mathfrak{S}) \geq \mathrm{Supp}_{\min}; \\ \lambda & \text{иначе}. \end{cases}$$

При этом $\mathrm{Conf}_{\mathrm{set}}(\alpha, \lambda, p, \omega, \mathfrak{S})$ полагается равным 0.

Будем называть закономерность $R = (p, \omega, f)$ *применимой* на наборе α при маске ω на пучке $\mathfrak{S}$, если $f(\alpha) \neq \lambda$. Если значения функции f определялись на основе частот, рассчитанных на пучке $\mathfrak{S}$, будем говорить, что закономерность R *построена* на пучке временных рядов $\mathfrak{S}$.

Понятие поддержки вводится не только для набора, но и для закономерности. Оно характеризует ту долю наборов, на которой закономерность оказывается применима. *Поддержкой* $\mathrm{Supp}(R, \mathfrak{S})$ *закономерности* R *на пучке* $\mathfrak{S}$ называется следующая величина:

$$\mathrm{Supp}(R, \mathfrak{S}) = \mathrm{Supp}(p, \omega, f, \mathfrak{S}) = \sum_{\alpha \in E_k^{\|\omega\|} : \mathrm{Supp}_{\mathrm{set}}(\alpha, p, \omega, \mathfrak{S}) \geq \mathrm{Supp}_{\min}} \mathrm{Supp}_{\mathrm{set}}(\alpha, p, \omega, \mathfrak{S}).$$

Из определений показателей качества закономерностей следует справедливость следующих неравенств для произвольной закономерности R и произвольного пучка временных рядов $\mathfrak{S}$:

$$0 \leq \mathrm{Conf}(R, \mathfrak{S}) \leq 1, \tag{1}$$

$$0 \leq \mathrm{Supp}(R, \mathfrak{S}) \leq 1. \tag{2}$$

1.2.2. Оценка необходимой длины пучка временных рядов. Выше был описан алгоритм построения функции f по маске ω для некоторого фиксированного целевого ряда p на основании пучка временных рядов $\mathfrak{S}$. Малая длина T пучка временных рядов $\mathfrak{S}$ приводит к появлению неопределенных значений λ на большом числе наборов функции f. Это способно сделать найденную закономерность $R = (p, \omega, f)$ непригодной для применения на практике. Ниже приводится оценка необходимой длины пучка временных рядов T при фиксированных k и $\|\omega\|$, которая основана на следующих заключениях. Если набор $\alpha \in E_k^{\|\omega\|}$ не появился в множестве $\{\alpha_t\}$, $t \in \{1, 2, \ldots, T - \Delta\}$, то значение функции f на нем равно λ. Поэтому информативной оценкой достаточности длины пучка временных рядов является вероятность того, что в множестве $\{\alpha_t\}$ присутствуют все наборы из $E_k^{\|\omega\|}$.

Теорема 2. *Пусть все наборы из* $E_k^{\|\omega\|}$ *появляются в множестве* $\{\alpha_t\}$ *с равной вероятностью. Обозначим через* $M = k^{\|\omega\|}$ *число всех наборов из* $E_k^{\|\omega\|}$, *через* $L = T - \Delta$ *– число элементов в множестве* $\{\alpha_t\}$ $(M < L)$. *Вероятность того, что в множестве* $\{\alpha_t\}$ *присутствуют все наборы из* $E_k^{\|\omega\|}$, *обозначим через* $\mathbf{P}$. *Тогда* $\mathbf{P} = M!\mathrm{S}(L, M)/M^L$, *где* $\mathrm{S}(L, M)$ *– число Стирлинга* II *рода.*

Доказательство. Заметим, что $\mathbf{P}$ можно интерпретировать как вероятность появления в выборке длины L всех элементов некоторого множества мощности M, причем элементы этого множества появляются в выборке с равной вероятностью; $\mathrm{S}(L, M)$ – тут число всех разбиений множества $\{1, 2, \ldots, L\}$ на M непустых подмножеств; $M!\mathrm{S}(L, M)$ – число всех сюръективных отображений из $\{1, 2, \ldots, L\}$ в $\{1, 2, \ldots, M\}$. В то же время $M!\mathrm{S}(L, M)$ – это число всех выборок длины L,

Таблица 1

k \ $\|\omega\|$	1	2	3	4	5	6
2	16	26	48	100	213	463
3	21	54	177	604	2063	6977
4	26	100	463	2186	10145	46241
5	31	162	982	5886	34435	197299

Таблица 2

k \ $\|\omega\|$	1	2	3	4	5	6
2	18	31	61	125	265	567
3	25	68	220	735	2458	8164
4	31	125	567	2603	11813	52916
5	38	202	1184	6904	39527	222766

состоящих из элементов множества мощности M, в которых присутствуют все элементы этого множества; M^L – число всех возможных выборок длины L, состоящих из элементов множества мощности M.

В табл. 1, 2 приведено минимальное значение T, при котором достигается соотношение $\mathbf{P}(k, \|\omega\|, T) \geqslant \mathbf{P}_0$, $\mathbf{P}_0 = 0.95$ и 0.99 соответственно. Для определенности Δ полагается равным 10. Значение вероятности $\mathbf{P}(k, \|\omega\|, T)$ определяется по приведенной выше формуле, где $M = k^{\|\omega\|}$ и $L = T - \Delta$.

1.2.3. Полнота системы закономерностей. Понятие поддержки $\mathrm{Supp}(R, \mathfrak{S})$ закономерности R на пучке $\mathfrak{S}$ характеризует, как часто закономерность оказывается применима на пучке $\mathfrak{S}$. Однако если закономерность была построена на пучке $\mathfrak{S}$, то оценка "применимости" может оказаться существенно завышенной. Чтобы избежать завышенной оценки, поддержку закономерности следует рассчитывать на пучке временных рядов, который не использовался для построения этой закономерности. Но не всегда имеющихся данных достаточно для того, чтобы корректно оценить применимость. Поэтому вводится понятие полноты системы закономерностей.

Системой закономерностей называется произвольная непустая совокупность закономерностей $R_1, ..., R_n : R_1 = (p_1, \omega_1, f_1), ..., R_n = (p_n, \omega_n, f_n)$, у которых совпадает целевой ряд и размерность масок:

$$p_1 = ... = p_n,$$

$$\omega_1, ..., \omega_n \in E_2^{N \times \Delta}.$$

Пусть фиксирован целевой ряд p. Любая закономерность может быть представлена как отображение $\hat{R}$ из $E_k^{N \times \Delta}$ в $E_{k_p} \cup \{\lambda\}$, что изображено на следующей коммутативной диаграмме:

$$\begin{array}{ccc} E_k^{N\times\Delta} & \xrightarrow{\hat{R}} & E_{k_p} \cup \{\lambda\} \\ & {\scriptstyle\hat{\omega}}\searrow \quad \nearrow{\scriptstyle f} & \\ & E_k^{\|\omega\|} & \end{array}$$

Здесь отображение $\hat{\omega}$ каждому набору $\beta \in E_k^{N \times \Delta}$ ставит в соответствие набор $\alpha \in E_k^{\|\omega\|}$, составленный из элементов набора β, соответствующих единичным элементам маски ω. Если закономерность применима на наборе α, будем говорить, что она применима на наборе β.

Таким образом, для каждой закономерности $R = (p, \omega, f)$, $\omega \in E_2^{N\times\Delta}$ и для каждого набора $\beta \in E_k^{N\times\Delta}$ существует отношение $r \subseteq E_k^{N\times\Delta}$, характеризующее применимость закономерности R на наборе β :

$$\beta \in r, \quad \text{если} \quad f(\alpha) \neq \lambda,$$
$$\beta \notin r, \quad \text{если} \quad f(\alpha) = \lambda.$$

Обозначим через $r_1 \cup r_2$ стандартную операцию объединения отношений r_1 и r_2, определенных на одном и том же конечном множестве $E_k^{N\times\Delta}$. Объединение отношений $r_\cup = r_1 \cup \ldots \cup r_n$ называется *полным*, если $r_\cup(\beta) = 1\ \forall\beta \in E_k^{N\times\Delta}$.

Система закономерностей $R_1, \ldots, R_n$: $R_1 = (p, \omega_1, f_1), \ldots, R_n = (p, \omega_n, f_n)$ с функциями $f_1, \ldots, f_n$ называется *полной*, если полно объединение отображений $r_\cup = r_1 \cup \ldots \cup r_n$, где каждое отношение r_i соответствует функции f_i, $i = 1, 2, \ldots, n$. В этом случае будем говорить, что система закономерностей $R_1, \ldots, R_n$ *покрывает* весь куб $E_k^{N\times\Delta}$. Из определения следует, что полная система закономерностей всегда способна сделать прогноз, так как всюду определена.

К сожалению, не всегда построенная система закономерностей является полной. Поэтому вводится понятие *полноты* $C(R_1, \ldots, R_n)$ системы закономерностей как доли наборов из $E_k^{N\times\Delta}$, которые покрываются системой закономерностей:

$$C = \frac{|r_\cup|}{k^{N\times\Delta}},$$

где $|r_\cup|$ – мощность отношения как подмножества декартового произведения $E_k^{N\times\Delta}$.

Легко видеть, что для полной системы закономерностей $C = 1$.

1.2.4. Алгоритм поиска постоянных закономерностей. Алгоритм поиска постоянных закономерностей на пучке временных рядов $\mathfrak{S}$ состоит в последовательном переборе масок, максимизирующем достоверность $\mathrm{Conf}(R, \mathfrak{S}_{\mathrm{valid}})$ построенных закономерностей. Здесь $\mathfrak{S}_{\mathrm{valid}}$ – отрезок пучка временных рядов $\mathfrak{S}$, используемый для валидации закономерностей.

В качестве методов оптимизации предлагается использовать методы селекции признаков, применяемые в распознавании образов. Это обусловлено сходством задачи выбора оптимального подмножества признаков и задачи выбора единичных элементов маски. В настоящей работе применяется алгоритм плавающего поиска, рассматриваемый в [22].

Алгоритм поиска постоянных закономерностей с целевым рядом p на пучке временных рядов $\mathfrak{S}$ может быть представлен в виде следующей схемы действий.

Шаг 1. Инициализация. $C = 0$ (C – значение полноты системы закономерностей, найденных на данный момент). Выбор первой маски ω.

Шаг 2. Построение по маске ω функции f. Получение новой закономерности $R = (p, \omega, f)$.

Шаг 3. Вычисление текущего значения полноты C системы закономерностей.

Шаг 4. Если $C = 1$, то конец выполнения алгоритма.

Шаг 5. Выбор очередной маски в соответствии с алгоритмом плавающего поиска, максимизирующим достоверность $\mathrm{Conf}(R, \mathfrak{S})$.

Шаг 6. Переход к шагу 2.

2. АНАЛИЗ ВРЕМЕННЫХ РЯДОВ С ИЗМЕНЯЮЩИМИСЯ ЗАКОНОМЕРНОСТЯМИ

Закономерность, определяющая поведение целевого ряда, может меняться с течением времени. Описанный выше алгоритм не способен адекватно реагировать на подобного рода изменения. Наиболее часто встречающейся в литературе идеей при поиске нестационарных закономерностей является использование “старых” наборов с меньшим весом по отношению к “более новым” наборам. Но данная методика не позволяет выделить закономерности, подвергающиеся структурным изменениям, т.е. изменениям маски или функции. Предлагаемый ниже подход призван исправить этот недостаток. Он учитывает возможные структурные изменения и ориен-

тирован не только на прогнозирование значений ряда, но и на поиск закономерностей, которые затем могут быть использованы при анализе и моделировании процесса, описываемого пучком временных рядов.

Идея подхода состоит в разбиении исходного пучка временных рядов на отрезки, на каждом из которых применяется алгоритм поиска постоянных закономерностей. Наиболее близкие в смысле некоторой меры сходства закономерности, полученные на различных отрезках, считаются этапами эволюции одной закономерности. Ключевую роль в алгоритме играет мера сходства на закономерностях.

2.1. Меры сходства

Ниже будет рассмотрена мера сходства на закономерностях, полученных в результате анализа k-значных временных рядов. Предполагается, что закономерности в процессе структурной деформации не могут менять номер целевого ряда. Поэтому мера сходства вводится на закономерностях с одинаковым параметром p.

2.1.1. Метрика на масках одинаковой мощности. Пусть σ_0 – симметрическая группа перестановок, действующих на множестве $\Omega = \{(1, 1), ..., (N, \Delta)\}$. Пусть также $\mathfrak{U}$ – произвольное множество, ω – произвольная матрица размера $N \times \Delta$ над множеством $\mathfrak{U}$ и s – перестановка из группы σ_0. Определим действие перестановки s на матрице ω равенством

$$s(\omega) = s(\|\omega_{ij}\|_{N\times\Delta}) = \|\omega'_{ij}\|_{N\times\Delta},$$

где $\omega'_{ij} = \omega'_{s(i,j)}$ при $i \in \{1, 2, ..., N\}$ и $j \in \{1, 2, ..., \Delta\}$.

Введем расстояние между масками одинаковой мощности следующим образом:

$$\rho_m(\omega_1, \omega_2) = \min_{s \in \sigma_0 : s(\omega_1) = \omega_2} W(\omega_1, s).$$

Здесь функция $W(\omega_1, s)$ с параметрами h и v ($h, v \in \mathbb{R}$, $h > 0$, $v > 0$) задает стоимость преобразования одной маски в другую:

$$W(\omega_1, s) = \sum_{(i,j) \in \mathbb{S} : \omega_1(i,j) = 1} (h|j - j_s| + v|i - i_s|), \quad \text{где} \quad (i_s, j_s) = s(i,j),$$

или, что то же самое,

$$W(\omega_1, s) = \sum_{(i,j) \in \mathbb{S}} \omega_1(i,j)(h|j - j_s| + v|i - i_s|), \quad \text{где} \quad (i_s, j_s) = s(i,j).$$

Теорема 3. *Описанное отображение ρ_m является метрикой.*

Доказательство. Так как параметры h и v положительны, то ρ_m отображает любую пару масок одинаковой мощности в неотрицательное число. Если $\rho_m(\omega_1, \omega_2) = 0$, то перестановка s в определении отображения такова, что $i = i_s$ и $j = j_s$, где $(i_s, j_s) = s(i, j)$, т.е. маски совпадают. Справедливо и обратное утверждение: если маски совпали, то $\rho_m(\omega_1, \omega_2) = 0$. Симметричность следует из существования в группе σ_0 обратной перестановки для любой из перестановок. Докажем выполнение неравенства треугольника: $\rho_m(x, z) \leqslant \rho_m(x, y) + \rho_m(y, z)$ для всех масок x, y и z. Пусть $\rho_m(x, z) > \rho_m(x, y) + \rho_m(y, z)$. Покажем, что тогда $\rho_m(x, z) > \min_{s \in \sigma_0 : s(x) = z} W(x, s)$, т.е. $\exists s_0 \in \sigma_0 : s_0(x) = z$ и $W(x, s_0) < \rho_m(x, z)$. Возьмем $s_0 = s_2 \circ s_1$. Здесь $\circ$ – операция суперпозиции в группе σ_0, $s_1 = \arg\min_{s \in \sigma_0 : s(x) = y} W(x, s)$ и $s_2 = \arg\min_{s \in \sigma_0 : s(y) = z} W(y, s)$. Отсюда $\rho_m(x, y) = W(x, s_1)$ и $\rho_m(y, z) = W(y, s_2)$.

Далее, $(i_{s_0}, j_{s_0}) = s_0(i,j)$, $(i_{s_1}, j_{s_1}) = s_1(i,j)$ и $(i_{s_2}, j_{s_2}) = s_2(i,j)$,

$$W(x, s_0) = \sum_{(i,j) \in \mathbb{S} : x(i,j) = 1} (h|j - j_{s_0}| + v|i - i_{s_0}|) =$$

$$= \sum_{(i,j) \in \mathbb{S} : x(i,j) = 1} (h\left|(j - j_{s_1}) + (j_{s_1} - j_{s_0})\right| + v\left|(i - i_{s_0}) + (i_{s_1} - i_{s_0})\right|) \leqslant$$

$$\leqslant \sum_{(i,j) \in \mathbb{S} : x(i,j) = 1} (h(\left|j - j_{s_1}\right| + \left|j_{s_1} - j_{s_0}\right|) + v(\left|i - i_{s_1}\right| + \left|i_{s_1} - i_{s_0}\right|)) =$$

$$= \sum_{(i,j) \in \mathbb{S} : x(i,j) = 1} (h\left|j - j_{s_1}\right| + v\left|i - i_{s_1}\right|) +$$

$$+ \sum_{(i,j) \in \mathbb{S} : x(i,j) = 1} (h\left|j_{s_1} - j_{s_0}\right| + v\left|i_{s_1} - i_{s_0}\right|) =$$

$$= W(x, s_1) + \sum_{(i,j) \in \mathbb{S} : x(i,j) = 1} (h\left|j_{s_1} - j_{(s_2 \circ s_1)}\right| + v\left|i_{s_1} - i_{(s_2 \circ s_1)}\right|) =$$

$$= W(x, s_1) + \sum_{(k,l) \in \mathbb{S} : y(k,l) = 1} (h\left|l - l_{s_2}\right| + v\left|k - k_{s_2}\right|) =$$

$$= W(x, s_1) + = W(y, s_2) = \rho_m(x, y) + \rho_m(y, z) < \rho_m(x, z).$$

Мы пришли к противоречию с определением расстояния как минимума функции W. Следовательно, наше предположение неверно и аксиома треугольника выполнена для всех масок x, y и z. Таким образом, доказано, что отображение ρ_m является метрикой.

2.1.2. Мера сходства на масках произвольной мощности. Пусть ω_1, ω_2 – маски, вообще говоря, различной мощности. Введем отображение ρ'_m следующим образом:

$$\rho'_m(\omega_1, \omega_2) = \begin{cases} \min\limits_{s \in \sigma_0 : s(\omega_1) \wedge \omega_2 = s(\omega_1)} W(\omega_1, s) + \dfrac{\min\{h, v\}}{N \times \Delta}(\|\omega_2\| - \|\omega_1\|), & \text{если} \quad \|\omega_1\| \leqslant \|\omega_2\|, \\ \min\limits_{s \in \sigma_0 : s(\omega_2) \wedge \omega_1 = s(\omega_2)} W(\omega_2, s) + \dfrac{\min\{h, v\}}{N \times \Delta}(\|\omega_1\| - \|\omega_2\|), & \text{если} \quad \|\omega_1\| > \|\omega_2\|. \end{cases}$$

Здесь $\wedge$ – умножение матриц по Адамару. Будем использовать ту же функцию W с параметрами $h > 0$ и $v > 0$:

$$W(\omega_1, s) = \sum_{(i,j) \in \mathbb{S} : \omega_1(i,j) = 1} (h\left|j - j_s\right| + v\left|i - i_s\right|), \quad \text{где} \quad (i_s, j_s) = s(i, j). \tag{3}$$

Таким образом, мы пытаемся переставить элементы меньшей маски, чтобы получить подмножество элементов большей маски.

Теорема 4. *Выполнены все аксиомы метрики, за исключением неравенства треугольника.*

Доказательство. Выполнение аксиом очевидно. Заметим лишь, что второе слагаемое добавлено в определение ρ'_m для справедливости следующего высказывания: $\rho'_m(x, y) = 0 \Leftrightarrow x = y$. Покажем, что неравенство треугольника, вообще говоря, не выполнено. Для этого рассмотрим маски

$$x = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}, \quad y = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \quad \text{и} \quad z = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}.$$

Тогда $\rho'_m(x, y) = \rho'_m(y, z) = h/4$, а $\rho'_m(x, z) = h + v$. Поэтому при соответствующем выборе h и v будет $\rho'_m(x, z) > \rho'_m(x, y) + \rho'_m(y, z)$.

Описанную меру сходства можно сделать метрикой, изменив слагаемое, отвечающее за различие в мощности масок:

$$\rho''_m(\omega_1, \omega_2) = \begin{cases} \min\limits_{s \in \sigma_0 : s(\omega_1) \wedge \omega_2 = s(\omega_1)} W(\omega_1, s) + (\vee N + h\Delta)(\|\omega_2\| - \|\omega_1\|), & \text{если} \quad \|\omega_1\| \leqslant \|\omega_2\|, \\ \min\limits_{s \in \sigma_0 : s(\omega_2) \wedge \omega_1 = s(\omega_2)} W(\omega_2, s) + (\vee N + h\Delta)(\|\omega_1\| - \|\omega_2\|), & \text{если} \quad \|\omega_1\| > \|\omega_2\|. \end{cases}$$

Здесь используется функция W, которая задается формулой (3).

Выполнение неравенства треугольника здесь доказывается аналогично доказательству для метрики на масках равной мощности.

Метрику ρ''_m целесообразно нормировать. В качестве коэффициента нормировки используется $\rho''_m(\emptyset, \Theta)$, где $\emptyset$ – маска, у которой нет переменных, а Θ – маска, состоящая только из единиц. Полученная метрика задается следующей формулой:

$$\rho_m^{\text{norm}}(\omega_1, \omega_2) = \frac{\rho''_m(\omega_1, \omega_2)}{\rho''_m(\emptyset, \Theta)} = \frac{\min\limits_{s \in \sigma_0 : s(\omega_1) \wedge \omega_2 = s(\omega_1)} W(\omega_1, s) + (\vee N + h\Delta)(\|\omega_2\| - \|\omega_1\|)}{N\Delta(\vee N + h\Delta)},$$

если $\|\omega_1\| \leqslant \|\omega_2\|$. Аналогично – если $\|\omega_2\| \leqslant \|\omega_1\|$.

Предложенная нормализация в метрике $\rho_m^{\text{norm}}(\omega_1, \omega_2)$ делает справедливым неравенство $0 \leq \rho_m^{\text{norm}}(\omega_1, \omega_2) \leq 1$, которое в дальнейшем применяется в мере сходства закономерностей. Однако оно способствует тому, что при малом числе аргументов в маске значения меры сходства масок $\rho_m^{\text{norm}}(\omega_1, \omega_2)$ концентрируются вблизи нуля. Для предотвращения этого эффекта на практике применяется следующая схема.

Пусть в рамках рассматриваемой задачи поиска закономерностей фиксировано максимальное количество аргументов функции. Это означает, что существует некоторый параметр μ, который определяет максимальное количество единиц в маске. Тогда удобно воспользоваться следующей метрикой:

$$\rho_m^{\mu}(\omega_1, \omega_2) = \frac{\rho''_m(\omega_1, \omega_2)}{\mu(\vee N + h\Delta)} = \frac{\min\limits_{s \in \sigma_0 : s(\omega_1) \wedge \omega_2 = s(\omega_1)} W(\omega_1, s) + (\vee N + h\Delta)(\|\omega_2\| - \|\omega_1\|)}{\mu(\vee N + h\Delta)}, \tag{4}$$

если $\|\omega_1\| \leqslant \|\omega_2\|$. Аналогично – если $\|\omega_2\| \leqslant \|\omega_1\|$. Для метрики $\rho_m^{\mu}(\omega_1, \omega_2)$ и для любых масок ω_1 и ω_2 в рамках задачи с фиксированным максимальным весом маски справедливо неравенство

$$0 \leq \rho_m^{\mu}(\omega_1, \omega_2) \leq 1. \tag{5}$$

2.1.3. Метрика на частично определенных функциях из P_k^*. Введем меру сходства $\hat{\rho}$ с параметром $w_\lambda \in \mathbb{R}$ на множестве $\{0, 1, \ldots, k-1, \lambda\}$. Для этого доопределим модуль разности, действующий на множестве $\{0, 1, \ldots, k-1\}$, следующими правилами:

1) $\hat{\rho}(\lambda, x) = \hat{\rho}(x, \lambda) = w_\lambda > 0 \ \forall x \in \{0, 1, \ldots, k-1\}$,

2) $\hat{\rho}(\lambda, \lambda) = 0$.

В силу положительности параметра w_λ очевидна справедливость всех аксиом метрики, за исключением неравенства треугольника. Сформулируем следующее утверждение.

Теорема 5. *Описанное отображение $\hat{\rho}$ является метрикой тогда и только тогда, когда $k \leqslant 2w + 1$.*

Доказательство. *Необходимость.* Если $\forall x, y, z \in \{0, 1, \ldots, k-1, \lambda\}$ будет $\hat{\rho}(x, z) \leqslant \hat{\rho}(y, z) + \hat{\rho}(y, z)$, то $k - 1 = \hat{\rho}(0, k-1) \leqslant \hat{\rho}(0, \lambda) + \hat{\rho}(\lambda, k-1) = 2w_\lambda$ или $k \leqslant 2w_\lambda + 1$.

Достаточность. Пусть $k \leqslant 2w_\lambda + 1$. Тогда неравенство $\hat{\rho}(x, z) \leqslant \hat{\rho}(x, y) + \hat{\rho}(y, z)$ выполнено при следующих условиях:

1) $\forall x, y, z \in \{0, 1, \ldots, k-1\}$;

2) $y = \lambda \ \forall x, z \in \{0, 1, \ldots, k-1\}$, так как $\hat{\rho}(x, z) \leqslant k - 1 \leqslant w_\lambda = \hat{\rho}(x, y) + \hat{\rho}(y, z)$;

3) $x = \lambda\ \forall y \in \{0, 1, ..., k-1, \lambda\}$, $\forall z \in \{0, 1, ..., k-1\}$, так как $\hat{\rho}(x, z) = w_\lambda \leqslant \hat{\rho}(x, y) + \hat{\rho}(y, z)$;

4) $z = \lambda\ \forall y \in \{0, 1, ..., k-1, \lambda\}$, $\forall x \in \{0, 1, ..., k-1\}$ аналогично п. 3);

5) $x = \lambda$, $y = \lambda$, $z = \lambda$, так как $\hat{\rho}(x, z) = 0 \leqslant \hat{\rho}(x, y) + \hat{\rho}(y, z)$.

Следствие 1. Минимальное w_λ, при котором $\hat{\rho}$ является метрикой, равно $(k-1)/2$.

Теперь определим метрику на функциях, которые зависят от одинакового числа переменных. Для этого надо фактически ввести метрику на векторах фиксированной длины, координатами которых могут быть элементы множества $\{0, 1, ..., k-1, \lambda\}$, где λ обозначает отсутствие значения функции на данном наборе. Заметим, что здесь неявно предполагается, что фиксирован некоторый порядок переменных:

$$\rho_f(f, g) = \frac{\Sigma_i \hat{\rho}(f_i, g_i)}{k^{\|\omega\|+1}}, \tag{6}$$

где i "пробегает" по всем наборам значений переменных, f_i и g_i – соответственно, значения функций f и g на i-м наборе или символ λ, а $\|\omega\|$ – количество переменных.

Так как количество наборов равно $k^{\|\omega\|}$ и $\hat{\rho}(x, y) < k$, то для произвольных функций f и g выполнено неравенство

$$0 \leq \rho_f(f, g) \leq 1. \tag{7}$$

Знаменатель меры сходства функций позволяет нормировать данную функцию, что представляется более удобным для практического использования и одновременно делает меру инвариантной относительно добавления фиктивных переменных.

Так как в определении отображения ρ_f используется сумма координат, то справедлива

Теорема 6. *Минимальное w_λ, при котором ρ_f является метрикой, равно $(k-1)/2$. При $0 \leqslant w_\lambda \leqslant (k-1)/2$ выполнены все аксиомы метрики, за исключением неравенства треугольника.*

Наконец, можно ввести метрику на произвольных частично определенных функциях из P_k^*. Определим ее так же, как и для функций, которые зависят от одинакового числа переменных. Для этого добавим фиктивные переменные к функции, у которой их меньше. Но добавление фиктивных переменных определено неоднозначно, так как не задан порядок на переменных. Порядок задается с использованием введенной метрики на масках. Этот процесс описан ниже в определении меры сходства на закономерностях.

2.1.4. Мера сходства на закономерностях. Основываясь на содержательных соображениях, можно сформулировать два требования к мере сходства на закономерностях. Во-первых, она должна отражать близость масок, а во-вторых – близость функций.

Пусть $R_1 = (p, \omega_1, f)$, $R_2 = (p, \omega_2, g)$ – закономерности, порожденные k-значными временными рядами, т.е. функции f и g принадлежат P_k^*. Определим отображение ρ на закономерностях, используя понятия метрики на масках и метрики на частично определенных функциях, введенные ранее (формулы (4) и (6) соответственно):

$$\rho(R_1, R_2) = \kappa_m \rho_m^\mu(\omega_1, \omega_2) + \kappa_f \rho_f(f, g).$$

Здесь κ_m и κ_f – веса мер сходства, удовлетворяющие следующим условиям: $0 \leq \kappa_m \leq 1$, $0 \leq \kappa_f \leq 1$, $\kappa_m + \kappa_f = 1$.

Так как справедливы неравенства (5) и (7), то верно аналогичное неравенство для меры сходства закономерностей:

$$0 \leq \rho(R_1, R_2) \leq 1, \tag{8}$$

где R_1 и R_2 – произвольные закономерности.

Как уже упоминалось выше, метрика на функциях не задана однозначно, пока не определен порядок переменных. Сначала зададим порядок на множестве $\Omega = \{(1, 1), ..., (N, \Delta)\}$: $(a, b) \geqslant (c, d)$, если $a \geqslant c$ или если $b \geqslant d$ при $a = c$. Далее для упрощения выкладок предположим, что $\|\omega_1\| \leqslant \|\omega_2\|$. Пусть (i_1, j_1), ..., $(i_{\|\omega_1\|}, j_{\|\omega_1\|})$ – единичные элементы маски ω_1, расположенные в соответствии с введенным на Ω порядком. Пусть известна перестановка $s \in \sigma_0 : s(\omega_1) \wedge \omega_2 = s(\omega_1)$. Тогда порядок переменных функции f задается так: $(i_1, j_1) \leqslant (i_{\|\omega_1\|}, j_{\|\omega_1\|})$, а далее следуют $\|\omega_2\| - \|\omega_1\|$ фиктивных

переменных. Порядок переменных функции g задается так: $s(i_1, j_1) \leqslant \ldots \leqslant s(i_{\|\omega_1\|}, j_{\|\omega_1\|})$, а далее следуют оставшиеся единичные элементы маски ω_1 в соответствии с введенным на Ω порядком. В случае если $\|\omega_2\| \leqslant \|\omega_1\|$, порядок задается аналогично. Теперь введенное нами отображение на частично определенных функциях определено однозначно для всех функций. Следовательно, $\rho(R_1, R_2)$ определено для всех закономерностей, порожденных k-значными временными рядами. Нетрудно видеть, что для $\rho(R_1, R_2)$ выполнены все аксиомы метрики, за исключением неравенства треугольника. Выполнение аксиом очевидно. А неравенство треугольника не выполнено, так как в качестве примера можно привести закономерности с одинаковыми функциями и масками

$$x = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}, \quad y = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \quad \text{и} \quad z = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}.$$

Выше было показано, что для них неравенство треугольника не выполнено.

2.2. Разбиения

Отрезком $\mathfrak{S}^1$ на пучке временных рядов $\mathfrak{S}$ с *началом* $t_b \in \{0, 1, \ldots, T\}$ и *концом* $t_e \in \{0, 1, \ldots, T\}$ $(t_b < t_e)$ назовем матрицу $N \times \theta$, где $\theta = t_e - t_b + 1$, составленную из последовательных столбцов матрицы $\mathfrak{S}$, первым из которых является столбец с номером t_b, а последним – столбец с номером t_e. Величина θ называется *длиной* отрезка $\mathfrak{S}^1$. Множество отрезков $\mathfrak{S}^1, \ldots, \mathfrak{S}^m$ на пучке временных рядов $\mathfrak{S}$ с началами $t_b^1, \ldots, t_b^m$ и концами $t_e^1, \ldots, t_e^m$ соответственно называется *последовательностью отрезков*, если $\forall i, j \in \{1, 2, \ldots, m\}$ справедливо $\{(i < j) \Rightarrow ((t_b^i < t_b^j)\ \&\ (t_e^i < t_e^j))\}$. Последовательность отрезков $\mathfrak{S}^1, \mathfrak{S}^2, \ldots, \mathfrak{S}^m$ называется покрытием $\mathfrak{S}$, если $\forall \tau \in \{0, 1, \ldots, T\}\ \exists \mathfrak{S}^i$, $i \in \{1, 2, \ldots, m\} : t_b \leqslant \tau \leqslant t_e$. Покрытие $\mathfrak{S}$ называется *разбиением* временного ряда, если $t_b^1 = 0$, $t_e^m = T$ и $t_b^{i+1} = t_e^i + 1$ при $i \in \{1, 2, \ldots, m-1\}$.

2.3. Плавно меняющиеся закономерности

Пусть $\mathfrak{S}^1, \ldots, \mathfrak{S}^m$ – последовательность отрезков на пучке временных рядов $\mathfrak{S}$. Введем понятия изменяющейся закономерности и ее длины.

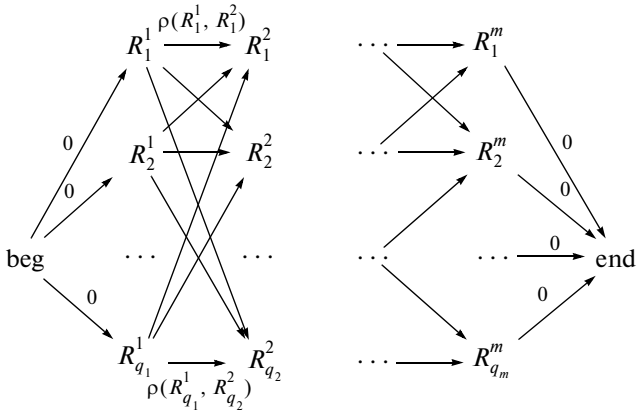
Изменяющейся закономерностью $\tilde{R}$ для последовательности отрезков $\mathfrak{S}^1, \ldots, \mathfrak{S}^m$ на пучке временных рядов $\mathfrak{S}$ называется система закономерностей $R^1, \ldots, R^m$, где каждая закономерность взаимно однозначно соответствует некоторому отрезку $\mathfrak{S}^i$, $i = 1, 2, \ldots, m$. Будем называть стационарные закономерности $R^1, \ldots, R^m$ *шагами*, которые *составляют* изменяющуюся закономерность $\tilde{R}$.

Длиной $l(\tilde{R})$ изменяющейся закономерности $\tilde{R}$ называется сумма мер сходства "соседних" шагов – закономерностей, составляющих меняющуюся закономерность:

$$l(\tilde{R}) = \sum_{j=1}^{m-1} \rho(R^j, R^{j+1}).$$

Пусть каждый из отрезков $\mathfrak{S}^j$, $j = 1, 2, \ldots, m$, разбит на две части: обучение $\mathfrak{S}^j_{\text{train}}$ и валидацию $\mathfrak{S}^j_{\text{valid}}$.

Тогда алгоритм поиска постоянных закономерностей, примененный к каждому из отрезков, порождает наборы закономерностей: $R_1^1, \ldots, R_{q_1}^1$ – на отрезке $\mathfrak{S}^1$, $\ldots$, $R_1^m, \ldots, R_{q_m}^m$ – на отрезке $\mathfrak{S}^m$. Для каждой закономерности R_i^j, $i = 1, 2, \ldots, q_j$, $j = 1, 2, \ldots, m$, определены значения показателей



Фиг. 1.

качества: достоверность на обучении $\mathrm{Conf}(R_i^j, \mathfrak{S}_{\mathrm{train}}^j)$, достоверность на валидации $\mathrm{Conf}(R_i^j, \mathfrak{S}_{\mathrm{valid}}^j)$, поддержка на обучении $\mathrm{Supp}(R_i^j, \mathfrak{S}_{\mathrm{train}}^j)$ и т.п.

Найденные закономерности можно представить в виде графа закономерностей (см. фиг. 1). Вершинами графа являются стационарные закономерности, найденные на каждом из отрезков, а также две дополнительные вершины: beg и end. С каждой вершиной ассоциированы показатели качества закономерности. Дугами на графе связаны закономерности соседних отрезков, что отражает факт возможного "превращения" одной закономерности в другую. С каждой дугой ассоциирован вес – мера сходства соответствующих закономерностей. Веса дуг, соединяющие закономерности крайних отрезков с вершинами beg и end, полагаются равными нулю.

Задача выделения наилучшей изменяющейся закономерности состоит в поиске пути между вершинами beg и end на ориентированном графе, который максимизирует показатели качества закономерностей вершин, входящих в него, и минимизирует суммарный вес ребер.

Эта задача сводится к стандартной задаче поиска кратчайшего пути на графе, если использовать в качестве веса вершины величину $(1 - Q_{\mathrm{step}})$, где Q_{step} – *функционал качества шага изменяющейся закономерности* $\tilde{R}$, который задается следующим образом:

$$\begin{gathered} Q_{\mathrm{step}}(R_i^j, R_l^{j+1}) = w_{\mathrm{conf}}\mathrm{Conf}(R_i^j, \mathfrak{S}_{\mathrm{valid}}^j) + w_{\mathrm{supp}}\mathrm{Supp}(R_i^j, \mathfrak{S}_{\mathrm{valid}}^j) + w_{\mathrm{similarity}}(1 - \rho(R_i^j, R_l^{j+1})), \\ Q_{\mathrm{step}}(\mathrm{beg}, R_i^j) = 0, \\ Q_{\mathrm{step}}(R_i^j, \mathrm{end}) = w_{\mathrm{conf}}\mathrm{Conf}(R_i^j, \mathfrak{S}_{\mathrm{valid}}^j) + w_{\mathrm{supp}}\mathrm{Supp}(R_i^j, \mathfrak{S}_{\mathrm{valid}}^j), \\ j = 1, 2, \ldots, m-1, \quad i = 1, 2, \ldots, q_j, \quad l = 1, 2, \ldots, q_{j+1}. \end{gathered} \tag{9}$$

Здесь $\mathrm{Conf}(R_i^j, \mathfrak{S}_{\mathrm{valid}}^j)$ и $\mathrm{Supp}(R_i^j, \mathfrak{S}_{\mathrm{valid}}^j)$ – показатели качества закономерности, $\rho(R_i^j, R_l^{j+1})$ – мера сходства закономерностей. Веса w_{conf}, w_{supp}, $w_{\mathrm{similarity}}$ функционала качества шага удовлетворяют следующим условиям:

$$0 \le w_{\mathrm{conf}} \le 1,$$

$$0 \le w_{\mathrm{supp}} \le 1,$$

$$0 \le w_{\mathrm{similarity}} \le 1,$$

$$w_{\mathrm{conf}} + w_{\mathrm{supp}} + w_{\mathrm{similarity}} = 1.$$

Так как справедливы неравенства (1), (2), (8), то для произвольных закономерностей R_1, R_2 верно неравенство

$$0 \le Q_{\mathrm{step}}(R_1, R_2) \le 1.$$

Таким образом, вес вершины $(1 - Q_{\text{step}})$ является неотрицательным и для решения задачи поиска кратчайшего пути на графе удобно использовать стандартные алгоритмы:

1) алгоритм Дейкстры (см. [23]) со сложностью $O(n^2)$, где n – число вершин графа;

2) алгоритм поиска кратчайшего расстояния в топологически отсортированном графе (см. [23]) со сложностью $O(n^2)$, где n – число вершин графа.

Изменяющуюся закономерность $\tilde{R}$ будем называть *плавно меняющейся*, если она составлена из закономерностей, лежащих на кратчайшем пути из вершины beg в вершину end и выполнено неравенство $w_{\text{similarity}} > 0$ для веса меры сходства закономерностей функционала качества шага.

2.4. Показатели качества плавно меняющихся закономерностей

Пусть $\tilde{R}_0$ – плавно меняющаяся закономерность, составленная из закономерностей $R_0^1, \ldots, R_0^m$, построенных, соответственно, на отрезках $\mathfrak{S}^1, \ldots, \mathfrak{S}^m$ пучка временных рядов $\mathfrak{S}$. Одним из основных показателей качества плавно меняющейся закономерности является ее длина $l(\tilde{R}_0)$, рассмотренная выше. Рассмотрим обобщение для случая плавно меняющихся закономерностей понятий достоверности и поддержки, введенных для постоянных закономерностей.

Достоверностью $\widetilde{\text{Conf}}(\tilde{R}_0, \mathfrak{S})$ *плавно меняющейся закономерности* $\tilde{R}_0$ на пучке временных рядов $\mathfrak{S}$ называется средневзвешеннная по длине отрезков достоверность закономерностей, составляющих плавно меняющуюся закономерность

$$\widetilde{\text{Conf}}(\tilde{R}_0, \mathfrak{S}) = \sum_{j=1}^{m} \frac{\theta_j}{T} \text{Conf}(R_0^j, \mathfrak{S}),$$

где $\theta_1, \ldots, \theta_m$ – длины отрезков $\mathfrak{S}^1, \ldots, \mathfrak{S}^m$ соответственно.

Аналогичным образом определяется *поддержка* $\widetilde{\text{Supp}}(\tilde{R}_0, \mathfrak{S})$ *плавно меняющейся закономерности* $\tilde{R}_0$ на пучке временных рядов $\mathfrak{S}$:

$$\widetilde{\text{Supp}}(\tilde{R}_0, \mathfrak{S}) = \sum_{j=1}^{m} \frac{\theta_j}{T} \text{Supp}(R_0^j, \mathfrak{S}).$$

Наконец, так как плавно меняющаяся закономерность является системой закономерностей, то для нее определено понятие полноты $C(\tilde{R}_0) = C(R_0^1, \ldots, R_0^m)$, введенное выше.

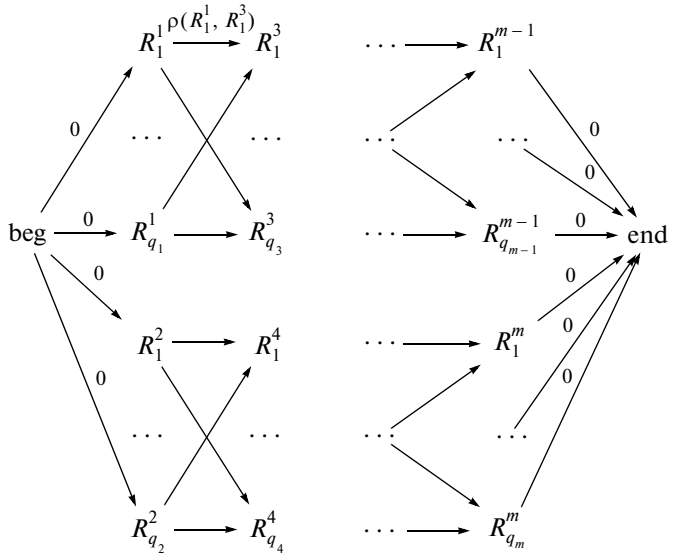
Поиск оптимальной плавно меняющейся закономерности на графе, где весом ребра является функционал качества шага изменяющейся закономерности, минимизирует следующую величину:

$$\sum_{j=1}^{m-1} [1 - Q_{\text{step}}(R_0^j, R_0^{j+1})] \longrightarrow \min.$$

В свою очередь, это означает, что плавно меняющаяся закономерность доставляет максимум функционала качества:

$$\sum_{j=1}^{m-1} Q_{\text{step}}(R_0^j, R_0^{j+1}) \longrightarrow \max.$$

Таким образом, плавно меняющаяся закономерность $\tilde{R}_0$ максимизирует достоверность $\widetilde{\text{Conf}}(\tilde{R}_0)$ и поддержку $\widetilde{\text{Supp}}(\tilde{R}_0)$ и минимизирует длину закономерности $l(\tilde{R}_0)$. Баланс между оптимумами определяется весами w_{conf}, w_{supp}, $w_{\text{similarity}}$ функционала качества шага изменяющейся закономерности.



Фиг. 2.

2.5. Поиск периодических закономерностей

Структура графа закономерностей не обязательно должна совпадать с изображенной на фиг. 1. Число ребер графа можно уменьшить, например, удалив те из них, которым приписан большой вес. И, наоборот, структуру графа можно усложнить, добавив ребра, соединяющие закономерности, полученные не на соседних отрезках. Такой подход опирается на то, что некоторые закономерности могут не проявляться в определенные моменты времени. Но, модифицируя структуру графа, крайне важно следить за делением пучка временных рядов на отрезки, так как оба этих процесса тесно взаимосвязаны и способны существенно повлиять на качество найденных закономерностей.

В зависимости от конкретных особенностей задачи можно формировать различные последовательности отрезков пучка временных рядов и составлять граф, на котором осуществляется поиск следов закономерностей. Например, во многих прикладных задачах имеет место непосредственная зависимость от времени года, месяца и т.п. В соответствии с этим естественным делением предлагается разбивать временнные ряды на отрезки. Но, чтобы иметь возможность проследить периодические закономерности, необходимо модифицировать граф закономерностей (см. фиг. 2).

Итак, варьируя последовательность отрезков и структуру графа закономерностей, мы можем подстраиваться под конкретные особенности рассматривамой задачи. Априорную информацию о закономерностях можно использовать при выборе отрезков и при конструировании графа закономерностей.

2.6. Сложность алгоритмов

При фиксированном числе рядов сложность $F(T)$ алгоритма поиска постоянных закономерностей линейно зависит от длины пучка рядов T при любом методе перебора масок. Это означает, что если в качестве последовательности отрезков ряда выбирается разбиение на m отрезков, то вычислительная сложность алгоритма поиска изменяющихся закономерностей составит $mF(T/m) + Cn^2 = F(T) + Cn^2$, где n – общее число порожденных закономерностей (оно зависит от T и m), а C – некоторая константа. Таким образом, в случае разбиения пучка рядов на отрезки сложность поиска изменяющихся закономерностей превосходит сложность поиска постоянных закономерностей ровно на сложность поиска кратчайшего пути на графе с числом вершин, равным числу порожденных закономерностей.

Таблица 3

Обозначение	Параметр
	Параметры генерации
K	количество значений в пучке
N	количество рядов в пучке
T	длина рядов
Δ_{gen}	максимальный отступ по времени
$\|\|\omega_1\|\|$	мощность первой маски
p_{gen}	индекс целевого ряда
m_{gen}	количество сегментов
ξ_{mask}	количество изменений маски при переходе к следующему отрезку
π_{mask}	вероятность каждого изменения маски при переходе к следующему отрезку
ξ_{func}	доля изменяемых значений функции при переходе к новому отрезку
π_{func}	вероятность каждого изменения функции при переходе к следующему отрезку
ε	уровень “белого шума” (доля значений целевого ряда, определяемых случайно)
	Параметры поиска стационарных закономерностей
p_{mine}	индекс целевого ряда
Δ_{mine}	максимальный отступ по времени
μ	максимальный вес маски
$\min \text{supp}_{\text{set}}$	минимальная поддержка набора
Valid	доля отрезка, которая используется для валидации закономерностей
	Фильтры базы знаний
$\text{conf}_{\min}$	минимальная достоверность на обучении закономерности для включения в базу знаний
$\text{err}_{\min}^{\text{valid}}$	максимальная ошибка на валидации
$\text{supp}_{\min}$	минимальная поддержка закономерности для включения в базу знаний

Таблица 4

Обозначение	Параметр
	Параметры поиска меняющихся закономерностей
m_{mine}	количество сегментов
v	стоимость перемещения аргумента по вертикали (используется в мере сходства масок)
h	стоимость перемещения аргумента по горизонтали (используется в мере сходства масок)
w_λ	расстояние до значения λ (используется в мере сходства функций)
$\boldsymbol{\varkappa}_{\text{mask}}$	вес меры сходства масок (используется в мере сходства закономерностей)
$\boldsymbol{\varkappa}_{\text{func}}$	вес меры сходства функций (используется в мере сходства закономерностей)
w_{conf}	вес меры, характеризующей точность закономерности (используется в функционале качества закономерностей)
w_{supp}	вес достоверности (используется в функционале качества закономерностей)
$w_{\text{similarity}}$	вес меры сходства закономерностей (используется в функционале качества закономерностей)

3. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

Разработанный алгоритм поиска плавно меняющихся закономерностей был использован на практике для анализа пучков временных рядов. Ниже приводятся результаты экспериментов на модельных и реальных временных рядах.

3.1. Примеры решения модельных задач

С целью испытания предложенного подхода для решения практических задач был подготовлен экспериментальный стенд. Стенд позволяет генерировать временные ряды и проводить по-

Таблица 5

Параметр	Значение в серии 1	Значение в серии 2
Параметры генерации		
K	4	4
N	10	10
T	1000	1000
Δ_{gen}	20	20
$\|\|\omega_1\|\|$	3	3
p_{gen}	0	0
m_{gen}	3	3
ξ_{mask}	1	1
π_{mask}	1	1
ξ_{func}	0.03	0.03
π_{func}	1	1
ε	изменяется	20%
Параметры поиска стационарных закономерностей		
p_{mine}	0	0
Δ_{mine}	20	20
μ	5	5
$\min \text{supp}_{\text{set}}$	0	0
Valid	20%	20%
Фильтры базы знаний		
$\text{conf}_{\min}$	0.3	0.3
$\text{err}_{\min}^{\text{valid}}$	1.0	1.0
$\text{supp}_{\min}$	0	0
Параметры поиска меняющихся закономерностей		
m_{mine}	3	3
v	1	1
h	1	1
w_λ	4	4
$\varkappa_{\text{mask}}$	0.5	0.5
$\varkappa_{\text{func}}$	0.5	0.5
w_{conf}	0.5	изменяется
w_{supp}	0	0
$w_{\text{similarity}}$	0.5	изменяется

иск стационарных и меняющихся закономерностей. С использованием стенда было проведено несколько серий экспериментов. Обозначения для параметров экспериментов представлены в табл. 3, 4.

Проводились две серии экспериментов на модельных рядах с целью выявить особенности для наиболее эффективного применения предложенных алгоритмов интеллектуального анализа временных рядов. Первая серия проводилась с целью определения эффективности алгоритма поиска постоянных закономерностей. Вторая серия экспериментов позволила оценить влияние меры сходства закономерностей при поиске плавно меняющихся закономерностей.

В первой серии модельных экспериментов было проведено исследование влияния уровня “белого шума” в моделируемых пучках временных рядов на качество распознавания. Для каждого значения уровня шума проводилась серия из 100 экспериментов. В каждом эксперименте ге-

Таблица 6

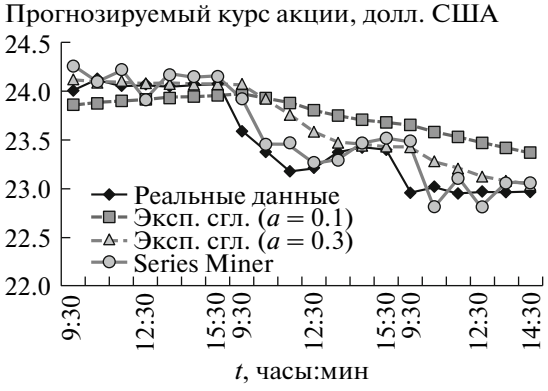
Уровень “белого шума” ε, %	Количество экспериментов	Доля успешных экспериментов, %
0	100	93
5	100	83
10	100	83
15	100	76
20	100	65
25	100	43
30	100	37
35	100	20
40	100	16
45	100	5
50	100	0

нерировалась случайная изменяющаяся закономерность, удовлетворяющая условиям на размерность маски и индекс целевого ряда. На основе закономерности генерировался пучок временных рядов, в котором затем происходил поиск изменяющихся закономерностей.

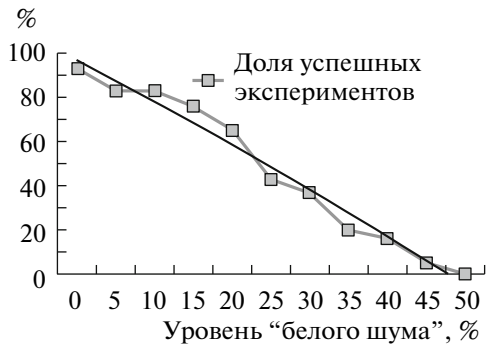
Критерии успешного эксперимента были определены следующим образом. Генерируемая и найденная стационарные закономерности называются *совпадающими*, если полностью совпадают их маски и доля различных значений функции не превышает 5% от общего числа значений. Изменяющиеся закономерности называются совпадающими, если совпадают все их соответствующие шаги – стационарные закономерности. Эксперимент признается успешным, если генерируемая и найденная изменяющиеся закономерности совпадают.

Значения параметров генерации и параметров алгоритмов поиска закономерностей представлены в табл. 5. Результаты моделирования в первой серии экспериментов представлены в табл. 6 и на фиг. 4. Результаты показывают, что алгоритм является достаточно стабильным и проводит вполне эффективный интеллектуальный анализ данных даже для зашумленных пучков временных рядов.

Во второй серии экспериментов исследовалось влияние меры сходства закономерностей на качество распознавания. С этой целью проводился поиск изменяющихся закономерностей для разных весов достоверности w_{conf} и меры сходства закономерностей $w_{\text{similarity}}$ функционала качества шага закономерности Q_{step}. При этом вес поддержки w_{supp} полагался равным нулю, что исключило влияние уровня поддержки на выбор оптимальной изменяющейся закономерности. Значения остальных параметров приведены в табл. 5.



Фиг. 3. Серия краткосрочных прогнозов ряда Business Objects.



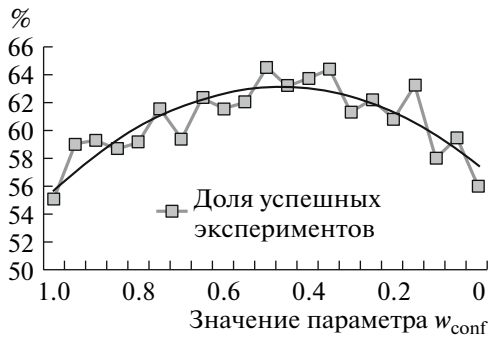
Фиг. 4. Качество распознавания при различном уровне “белого шума”.

Таблица 7

w_{conf}	w_{supp}	$w_{\text{similarity}}$	Количество экспериментов	Доля успешных экспериментов, %
1.00	0.00	0.00	1000	55.1
0.95	0.00	0.05	1000	59.0
0.90	0.00	0.10	1000	59.3
0.85	0.00	0.15	1000	58.7
0.80	0.00	0.20	1000	59.2
0.75	0.00	0.25	1000	61.6
0.70	0.00	0.30	1000	59.4
0.65	0.00	0.35	1000	62.4
0.60	0.00	0.40	1000	61.6
0.55	0.00	0.45	1000	62.1
0.50	0.00	0.50	1000	64.6
0.45	0.00	0.55	1000	63.3
0.40	0.00	0.60	1000	63.8
0.35	0.00	0.65	1000	64.5
0.30	0.00	0.70	1000	61.4
0.25	0.00	0.75	1000	62.3
0.20	0.00	0.80	1000	60.9
0.15	0.00	0.85	1000	63.4
0.10	0.00	0.90	1000	58.1
0.05	0.00	0.95	1000	59.6
0.00	0.00	1.00	1000	56.1

Комбинация значений весов $w_{\text{conf}} = 1$ и $w_{\text{similarity}} = 0$ соответствует алгоритму, при котором на каждом отрезке выбирается закономерность, обладающая максимальной достоверностью. Объединенные вместе данные закономерности составляют изменяющуюся закономерность. В обратном случае при комбинации значений весов $w_{\text{conf}} = 0$ и $w_{\text{similarity}} = 1$ ключевым параметром при выборе закономерностей, входящих в плавно меняющуюся закономерность, является их сходство. Таким образом, варьируя веса закономерности, можно оценить эффект от добавления меры сходства закономерностей в процесс распознавания.

В табл. 7 и на фиг. 5 приводятся значения доли успешных экспериментов при различных параметрах функционала качества закономерности. Из результатов видно, что добавление меры сходства закономерностей в функционал качества позволяет повысить точность распознавания.



Фиг. 5. Качество распознавания при различных весах функционала качества изменяющейся закономерности.

Таблица 8

Целевой ряд	Эксп. сгл. ($\alpha = 0.10$)	Эксп. сгл. ($\alpha = 0.30$)	Предложенный метод
Adobe	9.05×10^{-2}	7.32×10^{-2}	6.89×10^{-2}
BMC	13.24×10^{-3}	11.15×10^{-3}	9.72×10^{-3}
Business Objects	17.42×10^{-2}	7.19×10^{-2}	3.74×10^{-2}
Cognos	6.73×10^{-2}	3.08×10^{-2}	2.39×10^{-2}
Computer Associate	4.49×10^{-2}	2.87×10^{-2}	1.91×10^{-2}
Novell	7.62×10^{-3}	4.18×10^{-3}	2.54×10^{-3}
Oracle	8.61×10^{-3}	6.77×10^{-3}	5.45×10^{-3}
Peoplesoft	3.24×10^{-3}	2.94×10^{-3}	1.26×10^{-3}
Rational	5.19×10^{-2}	3.43×10^{-2}	2.06×10^{-2}

3.2. Примеры решения реальных задач

С целью сравнить предложенный в настоящей работе подход с другими методами была проведена серия экспериментов по краткосрочному прогнозированию временных рядов. Данными послужили курсы акций компаний Adobe, BMC, Business Objects, Cognos, Computer Associate, Novell, Oracle, Peoplesoft, Rational. Рассматривался средний почасовой курс акций в долларах за период с 13 мая 2002 г. по 10 декабря 2004 г. Средний почасовой курс получался как среднее арифметическое из четырех чисел: цены открытия (цены акции в начале часа), верхней цены (максимальной цены акции за час), нижней цены (минимальной цены акции за час), цены закрытия (цены акции в конце часа). Все девять временных рядов рассматривались как единый пучок, так как перечисленные выше компании работают в одной сфере разработки программного обеспечения для предприятий и не исключены взаимосвязи между поведением акций этих компаний.

Для прогнозирования цены акции помимо предложенного в настоящей работе метода применялось экспоненциальное сглаживание с параметрами α, равными 0.1 и 0.3. Исходные действительные временные ряды были преобразованы в 4-значные, где каждый элемент алфавита E_4 кодировал одно из значений: 0 – “большое падение”, 1 – “падение”, 2 – “рост”, 3 – “большой рост”. При этом диапазон действительных чисел, соответствующих каждому из значений, определялся на основе исходного временного ряда. Каждый из методов осуществил 20 прогнозов на один момент времени вперед. Средний квадрат ошибки каждого из методов представлен в табл. 8.

Как видно из табл. 8, при прогнозировании курса акций предложенный метод (обозначен как SeriesMiner) превосходит по качеству прогнозирования экспоненциальное сглаживание. Более подробно краткосрочные прогнозы для одного из рядов представлены на фиг. 3.

Закономерность, использованная при прогнозировании ряда Business Objects, такова:

BMC ($t-1$)	0	0	0	0	1	1	1	1	2	2	2	2	3	3	3	3
Business Objects ($t-1$)	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
Business Objects (t)	1	1	1	1	1	1	2	1	2	2	2	2	0	1	2	2

Здесь первые две строки задают значения переменных, а последняя – значение функции на соответствующем наборе. При этом 0 означает “большое падение”, 1 – “падение”, 2 – “рост”, 3 – “большой рост”. Приведенная закономерность была получена на последнем из отрезков пучка временных рядов, на котором происходило обучение. Для сравнения закономерность для того же целевого ряда на предпоследнем отрезке имеет вид

BMC ($t-1$)	0	0	0	0	1	1	1	1	2	2	2	2	3	3	3	3
Business Objects ($t-1$)	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
Business Objects (t)	1	1	**0**	1	1	1	2	1	2	**1**	2	2	0	1	2	2

Результаты испытаний показали, что предложенный в настоящей работе подход может быть достаточно эффективен при краткосрочном прогнозировании. Вместе с тем наибольший эф-

фект от применения метода достигается при анализе пучков временных рядов со значениями, которые изначально не могут быть представлены действительными числами, но кодируются числами k-значной логики.

4. ЗАКЛЮЧЕНИЕ

В настоящей работе предложен новый подход к поиску закономерностей в пучках k-значных временных рядов. Этот подход позволяет выявлять закономерности, которые подвергаются "плавным" структурным изменениям с течением времени. Для определения подобного рода изменений в работе предложена мера сходства закономерностей и описано ее применение как веса на графе закономерностей.

Найденные закономерности могут быть использованы как для прогнозирования следующих элементов пучка временных рядов, так и для анализа явления, описанного пучком, и для моделирования явления. Это делает возможным применение предложенного алгоритма в широком пласте задач прогнозирования временных рядов, а также в задачах изучения и описания процессов, которые могут быть представлены пучком временных рядов.

Предложенный в настоящей работе подход был реализован в программной системе и протестирован на модельных и реальных задачах. В работе описаны способы непосредственного практического использования разработанных методов анализа и прогноза пучков временных рядов. Испытания на модельных задачах показали адекватность метода, введенных мер сходства и функционалов качества. Краткосрочное прогнозирование реальных временных рядов свидетельствовало о достаточной точности найденных закономерностей.

Автор выражает благодарность своему учителю Константину Владимировичу Рудакову за постановку задачи и ценные советы, высказанные при написании настоящей работы.

СПИСОК ЛИТЕРАТУРЫ

1. *Андерсон Т.* Статистический анализ временных рядов. М.: Мир, 1976.
2. *Бокс Дж., Дженкинс Г.* Анализ временных рядов, прогноз и управление. М.: Мир, 1974.
3. *Engle R.F., Kroner K.F.* Multivariate simultaneous generalized ARCH // Econometric Theory. 1993. V. 11. P. 122–150.
4. *Hannan E.J.* Multiple time series. N.Y.: John Wiley and Sons, 1970.
5. *Лукашин Ю.П.* Адаптивные методы краткосрочного прогнозирования временных рядов. М.: Финансы и статистика, 2003.
6. *Morchen F., Ultsch A.* Mining hierarchical temporal patterns in multivariate time series // Proc. 27th Ann. German Conf. Artificial Intelligence, 2004. P. 127–140.
7. *Agrawal R., Imielinski T., Swamiet A.* Mining association rules between sets of items in large databases //Proc. Conf. Management of Data. Washington, USA. 1993. P. 207–216.
8. *Agrawal R., Srikant R.* Mining sequential patterns // Proc. 11th Internat. Conf. on Data Engng. Taipei: Taiwan, 1995. P. 3–14.
9. *Mannila H., Toivonen H., Verkamo A.I.* Discovery of frequent episodes in event sequences // Data Mining and Knowledge Discovery. 1997. V. 1. № 3. P. 259–289.
10. *Das G., Lin K., Mannila H. et al.* Rule discovery from time series // Proc. 4th Internat. Conf. Knowledge Discovery and Data Mining. New York, USA, 1998. P. 16–22.
11. *Sayal M.* Detecting time correlations in time-series data streams. Palo Alto: HP Labs, 2004.
12. *Morchen F., Ultsch A.* Optimizing time series discretization for knowledge discovery // Proc. 11th Internat. Conf. Knowledge Discovery and Data Mining. Chicago, USA. P. 660–665.
13. *Brown R.G.* Smoothing forecasting and prediction of discrete time series. New York: Prentice-Hall, 1963.
14. *Trigg D.W., Leach A.G.* Exponential smoothing with an adaptive response rate // Operat. Res. Quart. 1967. V. 18. № 1. P. 53–59.
15. *Engle R.F.* ARCH: Selected readings. Oxford: Oxford Univ. Press, 1995.
16. *Rao A.G., Shapiro A.* Adaptive smoothing using evolutionary spectra // Management Sci. 1970. V. 17. № 3. P. 208–218.
17. *Zadeh L.A., Ragazzini J.R.* An extension of Wiener's theory of prediction // J. Appl. Phys. 1950. V. 21. P. 645–655.
18. *Zadeh L.A., Ragazzini J.R.* The analysis of sampled-data systems // Applic. and Industry (AIEE). 1952. P. 225–234.

19. *Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И.* Методы и модели анализа данных: OLAP и Data Mining. СПб.: БХВ-Петербург, 2004.
20. *Журавлев Ю.И.* Избранные научные труды. М.: Магистр, 1998.
21. *Рудаков К.В.* Алгебраическая теория универсальных и локальных ограничений для алгоритмов распознавания: Дис.... докт. физ.-матем. наук. М.: ВЦ РАН, 1992.
22. *Pudil P., Ferri F.J., Novovicova J. et al.* Floating search methods for feature selection // Pattern Recognition Letts. 1994. V. 15. № 10. P. 1119–1125.
23. *Кристофидес Н.* Теория графов. Алгоритмический подход. М.: Мир, 1978.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2009, том 49, № 11, с. 2041–2058

УДК 519.71

О МЕРАХ СХОДСТВА И РАССТОЯНИЯХ МЕЖДУ ОБЪЕКТАМИ[1)]

© 2009 г. В. К. Леонтьев

(119333 Москва, ул. Вавилова, 40, ВЦ РАН)

e-mail: VKleontiev@mtu-net.ru

Поступила в редакцию 23.03.2009 г.

Рассматриваются различные метрики и функции от метрик. Приводится ряд результатов, относящихся к определению сходства и различия объектов для разных способов измерений. Библ. 17.

Ключевые слова: меры сходства, расстояние между объектами, мера Хаусдорфа, метрика Хемминга, метрика Левенштейна, метрика различия, функция близости, графы, булевы функции.

1. ВВЕДЕНИЕ

Понятие сходства и различия объектов является традиционным предметом исследования для ряда разделов дискретной математики. При этом часто для выражения этих понятий в количественной форме используется та или иная метрика. В теории кодирования с помощью расстояния моделируется мера искажения сообщений. В распознавании образов с помощью метрики устанавливается похожесть объектов. Проблема сходства и различия одинаково важна, казалось бы, для таких внешне далеких друг от друга областей, как геологоразведка, криминалистика, медицина и т.д. (см. [1]–[3]).

Мы рассматриваем различные способы сравнения объектов при помощи мер сходства и различия и приводим ряд результатов, призванных оценить целесообразность выбора тех или иных математических конструкций в разнообразных содержательных ситуациях. В статье не содержится никаких рецептов, но приводится много примеров. В целом содержание работы тяготеет не к чисто математическому осмыслению проблемы описания всех возможных метрик, а, скорее, к прикладной части этой проблемы, имеющей истоки в распознавании образов и теории информации.

Пусть M – произвольное конечное множество и $\rho(x, y)$ – произвольная метрика, заданная на M^2. Метрика ρ индуцирует расстояние Хаусдорфа между подмножествами M:

$$\rho_H(A, B) = \min_{\substack{x \in A \\ y \in B}} \rho(x, y). \tag{1}$$

Если $A \cap B \neq \varnothing$, то $\rho_H(A, B) \stackrel{\text{def}}{=} 0$.

В то же время “близость” подмножеств A и B можно “измерить” с помощью следующей величины:

$$R_f(A, B) = \sum_{\substack{x \in A \\ y \in B}} f(x, y). \tag{2}$$

Здесь $f(x, y)$ может быть, вообще говоря, произвольной функцией, свойства которой определяются условиями исследуемой проблемы и которая измеряет “близость” частей x и y объектов A и B. Отметим также, что на подмножествах M рассматриваются и другие функции, отличные от (1) и (2). Ниже приводится несколько примеров таких функций.

Пример 1: кодовое расстояние. Если $A \subseteq M$, то, по определению,

$$d(A) = \min_{x \neq y} \rho(x, y). \tag{3}$$

[1)] Работа выполнена при финансовой поддержке РФФИ (код проекта 08-01-00414).

Минимум в (3) берется по всем различным парам точек из подмножества A. Величина $d(A)$ называется *кодовым расстоянием* множества A. Фактически величина $d(A)$ характеризует меру различимости элементов подмножества A.

Пример 2: радиус покрытия. *Радиус покрытия* $r(A)$ подмножества $A \subseteq M$ определяется как минимальное число r такое, что шары радиуса $r(A)$ с центрами в точках множества A покрывают все множество M.

Пример 3: индекс Винера. *Индекс Винера* произвольного подмножества $A \subseteq M$ определяется следующим образом:

$$W(A) = \sum_{x, y \in A} \rho(x, y). \tag{4}$$

Функция (4) находит ряд серьезных приложений в химии (см. [2]). Индекс Винера характеризует совокупное сходство объектов класса A и тем самым определяет "внутреннюю близость" элементов одного таксона.

В формулах, приведенных выше, участвует расстояние $\rho(x, y)$, о котором мы пока ничего не сказали. Сейчас обсудим эту проблему с довольно общих позиций.

Если на M задано некоторое бинарное отношение, то разумно предположить, что мы имеем дело с конечным графом и тогда метрикой на M можно считать стандартное расстояние на графе, определяемое как минимальное число ребер, соединяющих две фиксированные вершины (в случае отсутствия пути между x и y это расстояние считается равным бесконечности). Отметим, что прием задания бинарного отношения на M и "превращения" этого множества в граф с естественной метрикой является довольно распространенным способом построения на конечном множестве метрического пространства. При этом могут возникать различные графы, зависящие от вида использованных бинарных отношений.

В общем виде расстояние на M – это вещественная функция $\rho(x, y)$, заданная на M^2 и удовлетворяющая следующим условиям:

$$\begin{gathered} \rho(x, y) \geqslant 0, \\ \rho(x, y) = \rho(y, x), \\ \text{если} \quad \rho(x, y) = 0, \quad \text{то} \quad x = y, \\ \rho(x, z) + \rho(z, y) \geqslant \rho(x, y). \end{gathered} \tag{5}$$

Система аксиом метрики (5) может варьироваться, и мы избрали наиболее распространенный случай.

Каждой метрике ρ на $M \times M$ соответствует квадратная матрица A_ρ порядка $n = |M|$, элементы которой – это расстояния между соответствующими точками множества M. Точнее, если $M = \{x_1, \ldots x_n\}$, то $\alpha_{ij} = \rho(x_i, x_j)$ и $A_\rho = \|\alpha_{ij}\|$. В соответствии с (5), матрица A_ρ удовлетворяет следующим условиям:

$$\alpha_{ij} \geqslant 0,$$

$$\alpha_{ii} = 0,$$

$$\alpha_{ij} = \alpha_{ji},$$

$$\alpha_{ij} + \alpha_{jk} \geqslant \alpha_{ik}, \quad \text{где} \quad i, j, k = \overline{1, n}.$$

Если матрица A_ρ удовлетворяет этим свойствам, то можно считать, что на множестве M задана метрика, определяемая этой матрицей.

Все такие матрицы A_ρ образуют выпуклый конус K в пространстве вещественнозначных матриц R_n. Так как конус K является конечно-порожденным, то абстрактное описание K может быть получено в терминах образующих элементов.

В целом глубокое и компетентное обсуждение проблемы описания метрик при разного рода ограничениях можно найти во многих работах чисто математического содержания. Мы же здесь перейдем к рассмотрению нескольких типов конкретных метрик, имеющих существенное значение в прикладных задачах.

2. МЕТРИКИ В B^n

2.1. Метрика Хемминга

Пусть B^n – множество вершин единичного n-мерного куба, или, что то же самое, множество двоичных наборов длины n. Формально

$$B^n = \{(x_1 \ldots x_n), x_i \in \{0, 1\}\}.$$

Самой известной метрикой в B^n является *расстояние Хемминга*, которое обозначим через $r_H(x, y)$. Формально

$$r_H(x, y) = \sum_{i=1}^{n} |x_i - y_i|, \tag{6}$$

где $x = (x_1 \ldots x_n)$, $y = (y_1 \ldots y_n)$.

Если на множестве B^n определить отношение смежности, соединив ребром любые две вершины, которые различаются ровно в одной компоненте, то мы получим граф $G = B^n$, расстояние между вершинами которого и есть расстояние Хемминга (6).

Если $\|x\|_H$ – норма вектора x или вес Хемминга точки x, т.е. число единичных компонент x, то

$$r_H(x, y) = \|x \oplus y\|_H, \tag{7}$$

где $\oplus$ – операция сложения по mod 2.

Таким образом, норма $\|x\|_H$ индуцирует расстояние Хемминга, согласованное с групповой операцией $\oplus$ на множестве B^n. С помощью нормы (7) на множестве B^n можно порождать другие метрики, “родственные” метрике Хемминга.

Пусть $N_n = \{1, 2, \ldots, n\}$ – отрезок натурального ряда и $\sigma = \{I_1, \ldots, I_m\}$ – произвольная система подмножеств, покрывающая N_n, т.е.

$$\bigcup_{k=1}^{m} I_k = N_n.$$

Каждое из подмножеств I_k “выделяет” некоторую систему координат точки $x = (x_1 \ldots x_n)$. Теперь положим

$$\|x\|_{I_k} = \sum_{j \in I_k} x_j$$

и определим норму $\|x\|^\sigma$ следующим образом:

$$\|x\|^\sigma = \max\{\|x\|_{I_1}, \ldots, \|x\|_{I_k}\}. \tag{8}$$

Утверждение 1. *Норма* (8) *порождает расстояние*

$$r_\sigma(x, y) = \|x \oplus y\|^\sigma.$$

Пример 4. Пусть $N_4 = \{1, 2, 3, 4\}$, $\sigma = \{I_1(1, 2), I_2(3, 4)\}$. Тогда

$$\|x\|_{I_1} = x_1 + x_2, \quad \|x\|_{I_2} = x_3 + x_4$$

и

$$\|x\|^\sigma = \max\{x_1 + x_2, x_3 + x_4\}.$$

Отсюда

$$\|1000\|_{I_1} = 1, \quad \|0101\|_{I_2} = 1, \quad r_\sigma(1000, 0101) = \|1101\|^\sigma = 2.$$

Пример 5. Если $I_n = \{1, 2, ..., n\}$, $\sigma = \{(i, j), 1 \leqslant i \leqslant j \leqslant n\}$, то

$$\|x\|^{\sigma} = \begin{cases} 0 & \text{при} \quad x = 0, \\ 1 & \text{при} \quad \|x\| = 1, \\ 2 & \text{при} \quad \|x\| \geqslant 2. \end{cases}$$

Отсюда

$$r_{\sigma}(x, y) = \begin{cases} 1, & \text{если} \quad r_H(x, y) = 1, \\ 2, & \text{если} \quad r_H(x, y) \geqslant 2. \end{cases}$$

Отметим, что метрика Хемминга обладает рядом свойств, делающих ее удобной и полезной в практических приложениях. Перечислим эти свойства, для того чтобы иметь возможность сравнивать метрику Хемминга с другими метриками.

Свойство 1. Инвариантность относительно сдвига:

$$r_H(x, y) = r_H(x + z, y + z).$$

Свойство 2. Инвариантность относительно перестановок координат:

$$r_H(x, y) = r_H(x_{\sigma}, y_{\sigma}),$$

где σ – перестановка из симметрической группы S_n и

$$x_{\sigma} = x_{\sigma(1)} \ldots x_{\sigma(n)}.$$

Свойство 3. Если $\circ$ – операция конкатенации на множестве слов B^*, то

$$r_H(x \circ y, u \circ v) = r_H(x, u) + r_H(y, v), \tag{9}$$

где $|x| = |u|$ и $|y| = |v|$.

Свойство (9) можно также записать в более общей форме:

$$r_H(x \circ y, u \circ v) = r_H(x_{\sigma} \circ y_{\tau}, u_{\sigma} \circ v_{\tau}),$$

где $\sigma \in S_{|x|}$, $\tau \in S_{|y|}$.

Ясно, что свойство 3 является следствием аддитивности нормы Хемминга

$$\|x \circ y\|_H = \|x\|_H + \|y\|_H.$$

Рассмотрим еще один известный способ введения нормы в B^n. Для $V \subseteq B^n$ положим

$$\|x\|_V = \min_{v \in B^n} \|x \oplus v\| = R_H(x, V).$$

Таким образом, $\|x\|_V$ – это расстояние от точки x до множества V.

Утверждение 2. *Если V – подгруппа B^n, то $\|x\|_V$ – псевдометрика.*

Доказательство. Нужно проверить лишь выполнение неравенства треугольника, так как симметричность расстояния $\rho_V(x, y) = \|x \oplus y\|_V$ очевидна. Имеем

$$\|x \oplus y\|_V = \min_{v \in V} \|x \oplus y \oplus v\| = \min_{v_1, v_2 \in V} \{\|x \oplus y \oplus v_1 \oplus v_2\|\} \leqslant \min_{v_1, v_2 \in V} \{\|x \oplus v_1\| + \|y \oplus v_2\|\} =$$
$$= \min_{v_1 \in V} \|x \oplus v_1\| + \min_{v_2 \in V} \|y \oplus v_2\| = \|x\|_V + \|y\|_V.$$

Пример 6. Если $V = \{(000), (111)\}$, то

$$\rho_V(000, 111) = 0, \quad \rho_V(100, 001) = 1, \quad \rho_V(110, 010) = 1.$$

Так как равенство $\rho_V(x, y) = 0$ не влечет $x = y$, то $\rho_V(x, y)$ – это лишь только псевдометрика. Приведенный выше вариант определения метрики называется *способом введения расстояния с помощью выпуклого тела*. Роль выпуклого тела в нашем случае играет группа V.

2.2. Метрика Левенштейна

Метрика Левенштейна была предложена в связи с изучением канала связи, в котором могут происходить ошибки типа “выпадения” и “вставки” символов (см. [4]).

Пусть $B = \{0, 1\}$ и B^* – множество всех слов конечной длины над алфавитом B. На множестве B^* определим отношение частичного порядка, положив

$$x \leqslant y, \tag{10}$$

если слово x может быть получено из слова y путем удаления некоторого числа букв.

Пример 7. Если $x = (010)$, $y = (110110)$, то $x < y$, так как x может быть получено из y удалением первой, второй и пятой букв.

Слова $x = (0110)$ и $y = (11011)$ являются “несравнимыми” по введенному частичному порядку. Соотношение

$$y \geqslant x$$

интерпретируется как существование некоторого множества вставок букв в слово x, которое переводит его в слово y.

Формально ситуация выглядит так. На множестве слов B^* построим граф (B^*, u), соединив два слова ребром, если одно из них может быть получено из другого путем удаления одной буквы. При этом очевидно, что вставкой одной буквы может быть осуществлена такая же трансформация. Полученный граф, очевидно, является аналогом диаграммы Хассе частичного порядка (10). В силу приведенных выше замечаний, этот граф является неориентированным и связным. Расстояние между вершинами x и y в этом графе может быть определено стандартным образом, и оно часто называется *расстоянием Левенштейна*, или r_L-метрикой (см. [4]). Содержательно $r_L(x, y)$ – это минимальное число вставок и выпадений, переводящих слово x в слово y.

Пример 8. Если $x = 1$, $y = 000$, то $r_L(x, y) = 4$ и соответствующий минимальный путь состоит из четырех ребер: (1, 01), (01, 001), (001, 00), (00, 000).

Пример 9. Если $x = 10101$, $y = 01010$, то $r_L(x, y) = 2$. В общем случае также если $x = (10)^n$ и $y = \bar{x}$, то $r_L(x, y) = 2$.

Расстоянию r_L может быть дана несколько другая интерпретация в терминах “подпоследовательностей” и “надпоследовательностей” (см. [4], [5]). Действительно, для любых двух слов x, $y \in B^*$ может быть определена *нижняя грань* длин слов z таких, что $x \leqslant z$, $y \leqslant z$. Эта нижняя грань обозначается через $M(x, y)$. Аналогично определяется *верхняя грань* длины слов z таких, что $z \leqslant x$, $z \leqslant y$. Эта грань обозначается через $m(x, y)$. Любое из слов z со свойством $x \leqslant z$, $y \leqslant z$ длины $|z| = M(x, y)$ называется наименьшей общей надпоследовательностью (н.о.н), а слово w со свойством $w \leqslant x$, $w \leqslant y$ и длиной $|w| = m(x, y)$ – наибольшей общей подпоследовательностью.

Числа $M(x, y)$ и $m(x, y)$ связаны следующим соотношением.

Лемма 1. *Имеет место соотношение*

$$M(x, y) + m(x, y) = |x| + |y|. \tag{11}$$

Из (11) при $|x| = |y| = n$ получим выражение для расстояния Левенштейна.

Лемма 2. *Имеет место соотношение*

$$r_L(x, y) = 2[M(x, y) - n]. \tag{12}$$

Равенство (11) является аналогом известного теоретико-числового соотношения

$$[a, b] \times (a, b) = ab,$$

где $[a, b]$ и (a, b) – наименьшее общее кратное и наибольший общий делитель натуральных чисел a и b.

Пример 10. Если $x = (000)$, $y = (111)$, то $M(x, y) = 6$ и $r_L(x, y) = 6$. Искомый кратчайший путь из x в y содержит 6 ребер: 3 вставки и 3 выпадения. Другой кратчайший путь имеет вид

$$111 \longrightarrow 11 \longrightarrow 101 \longrightarrow 10 \longrightarrow 100 \longrightarrow 00 \longrightarrow 000,$$

где вставки и выпадения чередуются.

Замечание 1. Доказательство того, что выражение (12) является метрикой, требует определенных усилий. В то же время графовая интерпретация r_L делает эти усилия излишними.

Свойства r_L-метрики.

1. Метрика r_L определена на любой паре слов $(x, y) \in B^* \times B^*$. Таким образом, расстояние может быть измерено и на словах разной длины. Это обстоятельство является существенным и выгодно отличает метрику r_L от метрики Хэмминга, где для измерения расстояния между словами разной длины необходимо предварительно их искусственно уравнять.

2. Справедливо соотношение $r_L(x, y) = r_L(\bar{x}, \bar{y})$. Здесь $\bar{x}$ – логическое дополнение слова x.

К сожалению, это соотношение является единственным общим инвариантом, сохраняющим расстояние Левенштейна. В связи с этим мощность шара радиуса t в метрике r_L зависит от центра этого шара (см. [5]).

3. Метрика r_L принимает только четные значения.

4. Если $x = (01)^n$ и $y = \bar{x}$, то $r_H(x, y) = 2n$ и $r_L(x, y) = 2$.

Таким образом, существуют пары слов, которые находятся друг от друга на максимальном расстоянии в метрике Хемминга и на минимальном расстоянии в метрике Левенштейна. С другой стороны, слова $x = 0^n$ и $y = \bar{x}$ находятся друг от друга максимально далеко в обеих метриках.

2.3. Метрика различия

Следующая метрика связана с выделением в двух объектах максимальной их "общей" части. Чем "массивнее" эта общая часть, тем "ближе" объекты. Ясно, что такого рода метрика отличается и от метрики Хемминга, и от метрики Левенштейна, если учесть ее содержательный смысл.

Пусть опять $B = \{0, 1\}$ и B^* – множество всех слов конечной длины над алфавитом B.

Определение 1. Слово $a \in B^*$ *различает* слова x и y, если выполнено одно из следующих условий:

$$\begin{aligned} &a \leqslant x, \quad a \not\leqslant y, \\ &a \leqslant y, \quad a \not\leqslant x. \end{aligned} \tag{13}$$

Отношение частичного порядка ($\leqslant$) в (13) понимается в смысле определения (10).

Таким образом, слово a различает слова x и y, если a является фрагментом одного и не является фрагментом другого. Любое слово a, обладающее свойством (13), называется *тестовым* для пары x, y. Множество всех таких слов обозначим через $T(x, y)$. Если $t(x, y)$ – минимальная из длин слов, входящих в $T(x, y)$, то любое из слов $a \in T(x, y)$ с длиной $|a| = t(x, y)$ называется *минимальным тестом* для пары x, y.

Замечание 2. Для $x = y$ положим, по определению, $t(x, y) = |x|$.

Пример 11. Если $x = 00110$, $y = 10010$, то $t(x, y) = 3$, так как любое слово длины два является фрагментом x и y, а слово $a = 100$ является лишь фрагментом слова y.

Пример 12. Если $x = 10^{n-1}$, $y = 0^n$, то $t(x, y) = 1$. При $x = 0^k1^{n-k}$, $y = 1^k0^{n-k}$ имеем $t(x, y) = 2$. Если же $x = (01)^n$, $y = (10)^n$, то $t(x, y) = n + 1$.

Ясно, что в общем случае справедливо неравенство

$$t(x, y) \leqslant \max\{|x|, |y|\}.$$

Лемма 3. *Функция $d_t(x, y) = n - t(x, y)$ является метрикой в B^n.*

Доказательство. Ясно, что $d_t(x, y) = d_t(y, x)$. Далее, если $d_t(x, y) = 0$, то множества $M_1 = \{(x_1, ..., ..., x_{n-1})(x_2, ..., x_n)\}$ и $M_2 = \{(y_1...y_{n-1})(y_2...y_n)\}$ совпадают, что означает равенство $x = y$. Осталось доказать неравенство треугольника, которое трансформируется в следующее:

$$t(x, z) + t(z, y) \leqslant n + t(x, y). \tag{14}$$

Пусть A, B, C – произвольные минимальные тесты для, соответственно, пар слов (x, z), (z, y), (x, y). Так как $t(u, v) \leqslant n$, то для доказательства (14) достаточно рассмотреть случай $|A| > |C|$, так как если $|A| \leqslant C$, то (14) выполняется, если $|A| > |C|$, то слово C не является минимальным тестом для пары (x, z), что означает следующее:

$$\text{(а) } C \leqslant X, \quad C \leqslant Z \quad \text{или} \quad \text{(б) } C \not\leqslant X, \quad C \not\leqslant Z.$$

С другой стороны, C – это минимальный тест для пары (x, y), что означает следующее:

$$\text{(в) } C \leqslant X, \quad C \not\leqslant Y \quad \text{или} \quad \text{(г) } C \leqslant Y, \quad C \not\leqslant X.$$

Из перечисленных вариантов “совместными” являются лишь следующие:

$$\text{(а), (в) } C \leqslant X, \quad C \leqslant Z, \quad C \not\leqslant Y \quad \text{или} \quad \text{(б), (г) } C \not\leqslant X, \quad C \not\leqslant Z, \quad C \leqslant Y.$$

Из (а), (в) следует, что слово C различает пару (z, y), т.е. выполняется неравенство $t(z, y) \leqslant |C| = t(x, y)$. Поэтому неравенство (14) выполнено. Если же справедливы (б), (г), то слово C различает пару (y, z) и вывод прежний.

Пример 13. Если $x = 0^n$, $y = 1^n$, то $t(x, y) = 1$ и $d_l(x, y) = n - 1$.

Пример 14. Если $x = (10)^n$ и $y = (01)^n$, то каждое из слов x и y содержит в качестве фрагментов все слова длины n. В то же время $1^n0 \in X$ и $1^n0 \notin Y$, т.е. $t(x, y) = n + 1$ и потому $d_l(x, y) = n - 1$.

Пример 15. Если $x = 100110$, $y = 011001$, то $d_l(x, y) = 2$. Если $x = 011011$, $y = 101001$, то $d_l(x, y) = 3$. Если $x = 101101110$, $y = 011011011$, то $d_l(x, y) = 5$.

Свойства метрики $d_l(x, y)$.

В [5] для величины $t(x, y)$ были получены верхние границы, из которых вытекают следующие неравенства:

$$\begin{aligned} d_l(x, y) &\geqslant [n/2] - 1, \\ d_l(x, y) &\geqslant n - m(x, y) - 1, \end{aligned} \tag{15}$$

где величина $m(x, y)$ – длина наибольшей общей подпоследовательности x и y.

Необычность метрики d_l состоит в том, что в B^n нет пар точек, находящихся на близком расстоянии, так как минимальное расстояние между точками в B^n, согласно (15), растет с ростом n.

Точки $x = 10^{n-1}$ и $y = 0^n$, находящиеся максимально близко по Хеммингу, находятся на расстоянии $d_l(x, y) = n - 1$ – максимально далеко в смысле расстояния d_l. Это же точки находятся максимально близко в метрике r_L.

3. ОКРЕСТНОСТИ

Понятие *окрестности* является традиционно важным для многих разделов дискретной математики (см. [5]–[9]). В терминах этого понятия формулируется ряд результатов, являющихся ключевыми как в теории локальных алгоритмов (см. [6]), так и в других областях дискретного анализа.

3.1. Метрическое понятие окрестности

Если множество M является метрическим пространством с метрикой ρ, то ε-*окрестностью* точки $x \in M$ называется множество точек из M, удовлетворяющих условию $\rho(a, x) \leqslant \varepsilon$. Обычно это множество обозначается через $Q_\varepsilon(x)$. В непрерывном случае традиционным образом окрестности является шар. В дискретном случае этот шар может выглядеть весьма причудливо. Если в метрике Хемминга все шары одинакового радиуса равномощны и изометричны, то для двух других рассматриваемых выше метрик эти качества шаров в значительной степени утрачены.

Так, мощность шара радиуса единица $Q_L(x)$ с центром в точке $x \in B^n$ ограничена следующими асимптотическими оценками:

$$n \lesssim Q_L(x) \lesssim n^2.$$

Точное значение этой величины приводится в [5].

Для случая метрики d_l ситуация выглядит еще более экзотически. Так, окрестность любого порядка $t \leqslant n - 2$ у точки $x = 0^n$ содержит только x, а окрестность порядка $t = n - 1$ совпадает с B^n.

Аналогичным образом определяется понятие ε-окрестности подмножества $A \subseteq M$.

Определение 2. ε-*окрестностью множества* A называется совокупность всех точек из M, находящихся на расстоянии не более чем ε от A в метрике Хаусдорфа. Формально

$$Q_\varepsilon(A) = \{x \in M : R_H(A, x) \leqslant \varepsilon\}.$$

Пример 16. Если $M = B^n$ и A есть k-мерная грань в B^n, то при ε целом имеем (см. [10])

$$\left|I_\varepsilon(A)\right| = 2^k \sum_{i=0}^{k} \binom{n-k}{i}.$$

Таким образом, мощность окрестности любой k-мерной грани одна и та же. В данном случае метрика Хаусдорфа порождается метрикой Хемминга.

Пример 17. Если опять $M = B^n$ и A – шар радиуса t с центром в точке a, то $O_\varepsilon(A)$ – это шар радиуса $t + \varepsilon$ с центром в точке a, т.е. $O_\varepsilon(S_t(a)) = S_{t+\varepsilon}(a)$.

В симметрических пространствах с понятием окрестности тесно связано понятие ε-энтропии, являющееся ключевым во многих задачах теории приближений и общей теории представимости функций одного класса в виде суперпозиций функций другого класса (см. [9]).

Отметим также, что в конечных графах с обычным понятием расстояния некоторые привычные геометрические объекты могут выглядеть довольно экзотически: шар может иметь более одного центра и т.д. Так возникают проблемы нахождения числа различных шаров заданного радиуса в различных типах графов. Многие интересные подробности на эту тему можно найти в [11]–[13].

3.2. Окрестности и преобразования

Понятие окрестности может быть введено и при отсутствии метрики и далеко не единственным способом (см. [7]). Мы не пытаемся здесь исследовать проблему окрестностей в аксиоматическом плане и рассматриваем лишь отдельные примеры, имеющие известный содержательный смысл.

Пусть задано множество M и группа преобразований T, действующая на M в традиционном смысле этого слова (см. [14]). Далее в группе T выбирается некоторое семейство преобразований G и с каждым элементом $x \in M$ связывается транзитивное множество

$$S^1(x) = \{gx : g \in G\}.$$

Множество $S^1(x)$ не является орбитой в классическом смысле этого слова, так как G, вообще говоря, не является подгруппой T.

Определение 3. *Окрестностью первого порядка элемента $x \in M$* называется следующее множество:

$$O^1(x, G) = x \cup S^1(x).$$

Замечание 3. Если G – группа, то $O^1(x, G)$ – орбита элемента x.

Окрестность второго порядка определяется как объединение окрестностей первого порядка всех элементов, входящих в $O^1(x, G)$, т.е.

$$O^2(x, G) = \bigcup_{y \in O^1(x, G)} O^1(y, G). \tag{16}$$

Окрестности высших порядков определяются аналогичным образом.

1. Преобразование “сдвига”. Пусть G – абелева группа с аддитивной операцией. Для произвольного элемента $a \in G$ определим преобразование элементов группы G следующим образом:

$$T_a(x) = x + a,$$

где $x \in G$. Таким образом, $\{T_a(x)\}$ – это группа преобразований группы G в себя. В предыдущих обозначениях $M = G$ и $T = \{T_a(x)\}$.

Выделим теперь в группе G произвольное A и рассмотрим семейство преобразований T_A, индуцированных элементами множества A.

Определение 4 (см. [15]). *Окрестностью $O^1(x, A)$ элемента $X \in G$* называется следующее множество:

$$O^1(x, A) = \{g + x, g \in A\} \cup X.$$

Окрестности следующих порядков определяются индуктивно в соответствии с формулой (16).

Для описания окрестностей важно понять, как устроены так называемые *аддитивные* множества. Положим, по определению, что для $V \subseteq G$

$$V^2 = \{v_i + v_j,\ v_i,\ v_j \in V\}$$

и V^2 – аддитивное множество порядка 2. По определению окрестности второго порядка точки $x \in G$, можно записать

$$O^2(x, V) = \{x + v,\ v \in V^2\}. \tag{17}$$

Равенство (17) показывает роль аддитивных множеств при построении окрестностей высших порядков.

Пример 18. Если G' – подгруппа группы G, то окрестность – это либо сама подгруппа, либо смежный класс по этой подгруппе.

Пусть

$$AB = \{a + b,\ a \in A,\ b \in B\}.$$

Справедливо следующее простое утверждение.

Лемма 4. *Имеет место соотношение*

$$(A \cup B)^2 = A^2 \cup AB \cup B^2.$$

Если A, B – подгруппы группы G, то $A^2 = A$, $B^2 = B$. Отсюда получаем

Следствие 1. Если A, B – подгруппы группы G, то

$$(A \cup B)^2 = A \cup B \cup AB.$$

В общем случае справедливы оценки

$$|AB| \leqslant |A| \cdot |B|, \quad |AB| \geqslant \max\{|A|, |B|\}. \tag{18}$$

Говоря о точности границ в (18), отметим следующее обстоятельство. Если A – произвольное подмножество группы G, а $B = \{0\}$, $|AB| = |A|$ и, в самом общем случае, оценки (18) точны.

Переходя к функции $|A^2|$, которая для нас представляет особый интерес, можно отметить следующие границы:

$$|A| \leqslant |A^2| \leqslant \binom{|A|}{2}.$$

Выберем теперь в качестве группы G куб B^n, который будем рассматривать как линейное векторное пространство над полем из двух элементов $F_2 = \{0, 1\}$. Отметим следующее простое утверждение.

Пусть $A = \{a_1 \ldots a_m\} \subseteq B^n$.

Лемма 5. *Если в множестве A каждые четыре вектора линейно независимы, то* $|A^2| = \binom{m}{2}$.

Доказательство. Соотношение линейной зависимости вида

$$a_i + a_j = a_s + a_r \tag{19}$$

означает, что пары (a_i, a_j) и (a_s, a_r) определяют один и тот же элемент множества A^2. Отсутствие соотношений вида (19) в множестве A говорит о том, что каждая пара (a_i, a_j) определяет "свой" элемент множества A^2, что и требовалось доказать.

Условие (19) легко интерпретировать в терминах теории алгебраических кодов (см. [12]).

Рассмотрим матрицу H_A, столбцами которой являются все элементы множества A. Если групповой код V имеет матрицу H_A в качестве проверочной, то кодовое расстояние $d(V)$ этого кода удовлетворяет неравенству $d(V) \geqslant 5$. Таким образом, используя стандартные аргументы из теории кодирования, нетрудно получить следующий результат.

Теорема 1. *Существует константа C_0 такая, что для любого числа $m_0(n) \leqslant C_0 \times 2^{n/2}$ справедливо равенство* $\max\limits_{|A| = m} |A^2| = \binom{m}{2}$ *при* $m \leqslant m_0(n)$.

Отметим, что мощность окрестности второго порядка играет существенную роль при оценке мощности оптимальных кодов для аддитивного канала связи (см. [15]).

2. Преобразование циклического сдвига. Пусть $M = B^n$ и T – группа преобразований циклического сдвига, действующая на B^n. Тогда окрестность $O^1(x, T)$ точки x – это следующее транзитивное множество:

$$O^1(x, T) = \{y, y = T^i(x), i = \overline{1, n}\}.$$

При этом мощность окрестности $|O^1(x, T)|$ равна индексу подгруппы

$$\mathrm{Stab}x = \{g \in T : gx = x\}.$$

Окрестности высших порядков точки x совпадают с $O^1(x, T)$, и разложение вида

$$B^n = \bigcup_{x \in B^n} O^1(x, T)$$

является разбиением B^n.

3.3. Окрестности по пересечению

Одним из естественных понятий “близости” является наличие общих элементов. Так возникает определение окрестности, связанное с операций пересечения. В дискретной математике наиболее существенные применения понятие окрестности, связанное с операцией пересечения, нашло в теории локальных алгоритмов (см. [6]). При этом каждый локальный алгоритм характеризуется двумя параметрами: величиной окрестности, изучаемой на каждом шаге, и числом признаков, запоминаемых о каждом элементе. В применении теории локальных алгоритмов к задаче минимизации булевых функций “базовыми” множествами являются грани единичного n-мерного куба, и “структура” пересечений граней играет существенную роль в этой теории (см. [6]).

1. Пересечение граней в B^n. Приведем некоторые результаты, полученные в [10].

Пусть $I_k = \{\mathfrak{M}_1, ..., \mathfrak{M}_{R_k}\}$ – семейство всех k-мерных граней в B^n. На произведении $I_k \times I_m$ определим случайную величину

$$t_{ij} = |\mathfrak{M}_i \cap \mathfrak{M}_j|,$$

где $\mathfrak{M}_i \in I_k$, $\mathfrak{M}_j \in I_m$. Каждой паре ($\mathfrak{M}_i \cap \mathfrak{M}_j$) припишем одну и ту же вероятность (равномерное распределение): $p = (R_k \cdot R_m)^{-1}$.

Пусть $F_{k,m}(z)$ – производящая функция этой случайной величины, т.е.

$$F_{k,m}(z) = p\sum_{i,j} z^{t_{ij}}.$$

Теорема 2 (см. [10]). *Справедлива формула*

$$F_{k,m}(z) = 1 + \frac{1}{R_k}\sum_{s = m+n-n}\left[2^{m-s}\binom{m}{s}\binom{n-m}{k-s}(z^{2^s} - 1)\right],$$

где $m \leqslant k$.

В качестве следствия этой теоремы легко получить выражение для математического ожидания мощности пересечения k-мерной и m-мерной граней B^n.

Следствие 2. Справедливо соотношение

$$\frac{1}{R_k R_m}\sum_{i,j} t_{ij} = 2^{m+k-n}.$$

2. Окрестности, фрагменты, опорные множества. Пусть опять $B = \{0, 1\}$ и $a \in B^*$.

Определение 5. Множество всех слов $x \in B^*$ таких, что

$$x \leqslant a$$

и $|x| = \lambda$, называется λ-*окрестностью* слова a.

Это понятие окрестности рассматривается в литературе в связи с использованием проблемы о восстановлении или реконструкции слова по его частям (см. [5]). При этом интересными оказываются как задача о реконструкции по окрестности, так и задача о реконструкции по мультиокрестности. Последнее связано с числом вхождений фрагмента x в слово a, которое называется *кратностью* $t_x(a)$.

В то же время если задана система признаков $I = \{s_1 \ldots s_m\}$ и определена некоторая система опорных множеств $V_1, \ldots, V_r \subseteq 2^I$ (см. [16]), то для этой системы $V = \{V_1 \ldots V_r\}$ можно ввести понятие V-эквивалентности объектов A и B, положив $A \approx B$, если и только если эти объекты неразличимы по системе опорных множеств V. Это понятие позволяет определить окрестность объекта $A - O_V^1(A)$, как класс объектов, эквивалентных A по системе опорных множеств V.

4. РАССТОЯНИЕ МЕЖДУ МНОЖЕСТВАМИ

Если M – конечное множество с метрикой ρ, то, по (1), расстояние по Хаусдорфу между подмножествами $A, B \subseteq M$ определяется формулой

$$\rho_H(A, B) = \min_{\substack{a \in A \\ b \in B}} \rho(a, b)$$

и при этом $\rho_H(A, B) = 0$, если $A \cap B \neq \varnothing$.

Таким образом, каждое расстояние на M индуцирует расстояние Хаусдорфа на семействе 2^M.

Пример 19. Пусть $M = N_n = \{1, 2, \ldots, n\}$ и $\rho(i, j) = |i - j|$.

Каждое из множества $A \subseteq M$ будем представлять в упорядоченном виде $A = \{a_1 < \ldots < a_m\}$. Тогда при $A = \{a_1 < \ldots < a_m\}$, $B = \{b_1 < \ldots < b_k\}$ имеем

$$\rho(A, B) = \min\{|a_1 - b_1|, |b_1 - a_m|, |a_m - b_k|\}$$

при $a_1 < b_1 < a_m < b_k$ и

$$\rho_H(A, B) = |a_m - b_1|$$

при $a_m < b_1$.

Если же рассматривать в качестве подмножеств N_n только семейство интервалов, то это семейство имеет вид $I_n = \{(i, j),\ i < j\}$ и

$$\rho_H[(i_1, j_1), (i_2, j_2)] = \begin{cases} |j_1 - i_2| & \text{при} \quad j_1 < i_2, \\ |j_2 - i_1| & \text{при} \quad j_2 < i_1. \end{cases}$$

Пусть $N_m * N_n$ – целочисленная решетка размеров $m \times n$ на плоскости. Будем рассматривать эту решетку как граф, ребра которого соединяют соседние вершины решетки. В качестве семейства подмножеств выберем множество всех простых путей этого графа, а расстояние между любыми двумя путями – это расстояние по Хаусдорфу между множествами вершин графа, входящих в эти пути.

4.1. Расстояние между подмножествами в B^n

В кубе B^n существуют “фигуры” (подмножества), играющие существенную роль в ряде разделов дискретного анализа. Такими фигурами являются грани B^n и шары в B^n. В этом пункте приводится ряд результатов, относящихся к расстоянию между гранями и шарами.

1. Расстояния между шарами в B^n.

Определение 6. Пусть $M \subseteq B^n$ и $a \in M$. Точка a называется *внутренней точкой множества M*, если шар радиуса 1 с центром в точке a принадлежит M. В противном случае a называется *граничной точкой множества M*.

Множество всех граничных точек множества M называется его *границей* и обозначается через $\Gamma(M)$.

Пример 20. Если M – счетчик четности, т.е. $M = \{x : \|x\| \equiv 0(\mathrm{mod}2)\}$, тогда все точки M являются граничными, т.е. $M = \Gamma(M)$.

Пример 21. Если M – шар радиуса t, то $\Gamma(M)$ – сфера радиуса t.

Справедливо следующее общее утверждение, относящееся к вычислению расстояний между подмножествами B^n.

Лемма 6. *Если $A \cap B \neq \varnothing$, то справедливо соотношение*

$$R_H(A, B) = R_H(\Gamma(A), \Gamma(B)).$$

Ясно, что $|\Gamma(A)| \leqslant |A|$ и, как показывает пример 20, это неравенство является достижимым. Любое множество в B^n имеет границу, но не любое подмножество $A \subseteq B^n$ может служить границей.

Например, шар радиуса единица не может служить границей ни для какого множества из B^n.

Прямым следствием леммы 6 является

Лемма 7. *Если $S_t(a)$ – шар радиуса t с центром в точке $a \in B^n$, то*

$$R_H(S_r(a), S_R(b)) = \max\{0, r_H(a, b) - r - R\}.$$

Рассмотрим теперь производящую функцию распределения расстояний между шарами радиуса t и B^n, т.е.

$$F_t(z) = \frac{1}{2^{2n}} \sum_{a, b \in B^n} z^{R_H(a, b)}.$$

Теорема 3. *Справедливо соотношение*

$$F_t(z) = \frac{S_n^{2t}}{2^n} + \frac{z^{-2t}}{2^n} \sum_{m = 2t+1} \binom{n}{m} z^m, \quad \text{где} \quad S_n^r = \sum_{k=0}^{r} \binom{n}{m}.$$

Доказательство. Из определения $F_t(z)$ используя лемму 7, получаем

$$F_t(z) = \frac{1}{2^{2n}} \sum_{\substack{a, b \in B^n \\ r_H(a, b) \leqslant 2t}} z^{R_H(a, b)} + \frac{1}{2^{2n}} \sum_{\substack{a, b \in B^n \\ r_H(a, b) \geqslant 2t+1}} z^{R_H(a, b)} =$$

$$= \frac{S_n^{2t}}{2^n} + \frac{1}{2^{2n}} \sum_{r_H(a, b) \geqslant 2t+1} z^{R_H(a, b)} = \frac{S_n^{2t}}{2^n} + \frac{z^{-2t}}{2^n} \sum_{m = 2t+1}^{n} \sum_{\substack{a, b \\ r_H(a, b) = m}} z^m = \frac{S_n^{2t}}{2^n} + \frac{z^{-2t}}{2^n} \sum_{m = 2t+1} \binom{n}{m} z^m.$$

Из последней теоремы, используя традиционные вероятностные методы, нетрудно получить следующие утверждения.

Следствие 3. Если $t = o(n)$, то случайная величина $\xi = R_H(a, b)$ имеет параметры

$$M\xi = \frac{n}{2} - 2t - o(1), \quad D\xi = \frac{n}{4} + o(1).$$

Следствие 4. Если $t = o(n)$, то почти все пары шаров радиуса t находятся на расстоянии $r = n/2 - 2t - o(1)$ один от другого.

2. Расстояния между гранями в B^n. Достаточно подробно эта задача рассмотрена в [10].

Мы приведем только один результат. Пусть I_p, I_q – грани размерности p и q из B^n. Рассмотрим случайную величину $\xi_{p,q} = R_H(I_p, I_q)$ – расстояние между гранями I_p и I_q. Пусть $F_{p,q}(z)$ – функция распределения этой случайной величины, т.е.

$$F_{p,q}(z) = \frac{1}{2^{2n-p-q}\binom{n}{p}\binom{n}{q}} \sum_{I_p, I_q} z^{R_H(I_p, I_q)}.$$

Теорема 4 (см. [6]). *Справедлива формула*

$$F_{p,q}(z) = \frac{1}{\binom{n}{p}2^{n-p}} \frac{1}{2\pi i} \oint\limits_{|u|=\rho} \frac{(u+2)^q (u+z+1)^{n-q}}{u^{p+1}} du,$$

здесь $\rho < 1$.

Следствие 5 (см. [6]). Для математического ожидания случайной величины $\xi_{p,q}$ выполняется соотношение

$$\mathbf{M}(\xi_{p,q}) = \frac{n}{2}(1 - p/n)(1 - q/n).$$

5. ФУНКЦИЯ БЛИЗОСТИ

Если A, B – некоторые подмножества M и $f(x, y)$ – произвольная функция вида

$$M \times M \longrightarrow R,$$

то выражение вида

$$R_f(A, B) = \sum_{\substack{x \in A \\ y \in B}} f(x, y) \tag{20}$$

называется функцией близости подмножеств A и B.

Для упрощения изложения будем считать, что в формуле (20) множества A и B либо не пересекаются, либо совпадают и суммирование в (20) ведется по всем парам $(x, y) \in A \times B$.

Замечание 4. Если заранее предположить, что $f(x, x) = 0$ для всех $x \in \mathbf{M}$, то ограничения на множества A и B могут быть сняты.

Иногда в качестве A выбирается таблица обучения (см. [16]), а в качестве B – описание нового объекта, который необходимо протестировать на предмет его принадлежности к классу A. Тогда функция $R_f(A, B)$ характеризует суммарную близость объекта B к классу A.

Если A – множество вершин графа $G = (A, U)$ и $f(x, y)$ – функция расстояния, заданная на этом графе, то выражение

$$R_f(A, A) = \sum_{(x, y) \in U} f(x, y) = W_f(G)$$

называется *индексом Винера графа* G (см. [2]).

Нетрудно видеть, что индекс Винера для произвольного подмножества A

$$W_f(A) = \sum_{(x, y) \in A^2} f(x, y)$$

связан с функцией близости следующим образом.

Лемма 8. *Справедливо равенство*

$$R_f(A, B) = W_f(A \cup B) - W_f(A) - W_f(B). \tag{21}$$

Равенство (21) показывает, что при исследовании функций близости достаточно ограничиться индексом Винера.

Примеры вычисления индекса Винера многочисленны и разнообразны. Ограничимся несколькими частными случаями, связанными с метрикой Хэмминга.

1. Случай $M = B^n$ и $f(x, y) = r_H(x, y)$.

Следующее известное утверждение используется во многих задачах, связанных с помехоустойчивым кодированием (см. [12]).

Пусть $A = \{a_1 \ldots a_m\} \subseteq B^n$ и H_A – матрица, строками которой являются элементы множества A. Пусть i-й столбец H_A содержит k_i единичных элементов и $K_A = (k_1 \ldots k_n)$ – столбцовый вектор матрицы H_A.

Теорема 5. *Справедливо соотношение*

$$W_{\rho_H}(A) = 2\sum_{i=1}^{n} k_i(m - k_i). \tag{22}$$

Лемма 8 позволяет легко вычислить индекс Винера следующих фигур в B^n (см. [17]):

а) если A есть k-мерная грань в B^n, то

$$W_{\rho_H}(A) = k \times 2^k;$$

б) если A – сфера радиуса t в B^n, то

$$W_{\rho_H}(A) = 2n\binom{n-1}{k-1}\binom{n-1}{k};$$

в) если A – сфера радиуса t в B^n, то

$$W_{\rho_H}(A) = 2nS_{n-1}^{t-1}(S_n^{t-1} + n + 1),$$

где $S_n^r = \sum_{i=0}^{r}\binom{n}{i}$ – число точек в шаре радиуса r в B^n.

Из (21) и (22) вытекает следующее соотношение для функции близости $R_{r_H}(A, B)$ двух подмножеств из B^n.

Пусть K_A, K_B – столбцовые векторы матрицы H_A и H_B соответственно:

$$\|A\| = \sum_{i=1}^{n} K_i(A), \quad \|B\| = \sum_{i=1}^{n} K_i(B).$$

Лемма 9. *Справедливо соотношение*

$$R_{r_H}(A, B) = |B|\|A\| + |A|\|B\| - 2(K_A, K_B).$$

Здесь (K_A, K_B) *– обычное скалярное произведение векторов* K_A *и* K_B.

Следствие 6. Пусть $A = B_p^n$, $B = B_q^n$ суть p-й и q-й слои куба. Тогда

$$R_{r_H}(A, B) = 2\binom{n}{p}\binom{n}{q}\left[p + q - \frac{2pq}{n}\right].$$

Следствие 7. Пусть $R_{r_н}(A, b)$ – суммарная близость объекта $b = (\beta_1...\beta_n) \in B^n$ к классу A. Тогда справедлива формула

$$R_{r_н}(A, b) = \|A\| + m\|B\| - \sum_{i=1}^{n} \beta_i K_i(A).$$

Здесь $|A| = m$.

Пусть $M = B^n$, $x = (x_1...x_n)$, $y = (y_1...y_n)$ и $f(x, y)$ имеет следующий вид:

$$f(x, y) = \sum_{i=1}^{n} \varphi(x_i, y_i),$$

где $\varphi(a, b)$ – произвольная вещественнозначная функция, определенная на B^2. Таким образом $\varphi(a, b)$ может принимать только четыре различных значения.

Рассмотрим произвольное подмножество $A = (x^1...x^m)$ из B^n и определим индекс Винера на A, положив

$$W_f(A) = \sum_{s, r} f(x^s, x^r).$$

Лемма 10. *Справедливо равенство*

$$W_f(A) = \varphi(0, 0)\sum_{i=1}^{n} k_i^2 + [\varphi(0, 1) + \varphi(1, 0)]\sum_{i=1}^{n} k_i(m - k_i) + \varphi(1, 1)\sum_{i=1}^{n} (m - k_i)^2, \tag{23}$$

где $k = (k_1...k_n)$ – столбцовый вектор матрицы H_A, соответствующей множеству A и $m = |A|$.

Доказательство. По определению, имеем

$$W_f(A) = \sum_{i=1}^{n} \sum_{s, r} \varphi(x_i^s, x_i^r).$$

Рассмотрим матрицу H_A:

$$H_A = \begin{Vmatrix} x_1^1 & \dots & x_n^1 \\ & & \\ x_1^m & \dots & x_n^m \end{Vmatrix}.$$

Разобьем все пары элементов i-го столбца матрицы H_A на четыре класса V_{00}^i, V_{01}^i, V_{10}^i, V_{11}^i, отнеся к классу $V_{\alpha\beta}$ все пары вида $(\alpha\beta)$. Тогда

$$\sum_{(s, r)} \varphi(x_i^s, x_i^r) = \varphi(0, 0)\left|V_{00}^i\right| + \varphi(0, 1)\left|V_{01}^i\right| + \varphi(1, 0)\left|V_{10}^i\right| + \varphi(1, 1)\left|V_{11}^i\right|$$

и, соответственно,

$$W_f(A) = \sum_{(\alpha, \beta) \in B^2} \varphi(\alpha, \beta)\sum_{i=1}^{n} \left|V_{\alpha\beta}^i\right|.$$

Если в i-м столбце матрицы H_A содержится k_i единиц, то

$$\left|V_{00}^i\right| = k_i^2, \quad \left|V_{01}^i\right| = \left|V_{10}^i\right| = k_i(m-k_i), \quad \left|V_{11}^i\right| = (m-k_i)^2.$$

Отсюда следует соотношение (23).

Следствие 8. Если $\varphi(0, 0) = \varphi(1, 1) = 0$, а $\varphi(0, 1) = \varphi(1, 0) = 1$, то

$$W_f(A) = 2\sum_{i=1}^{n} k_i(m-k_i),$$

что представляет классическую формулу (22).

6. РАССТОЯНИЕ МЕЖДУ ФОРМУЛАМИ

Рассмотрим класс P_2^n булевых функций, зависящих не более чем от n переменных. Каждая из функций этого класса может быть задана таблицей или формулой в каком-нибудь полном или не полном базисе (см. [14]). При любом таком задании искажение отдельных букв может привести к другой функции, которая существенно отличается от первоначальной. Например, функция f может быть воспринята как $\bar{f}$ и этот переход является фатальным, так как он искажает смысл исходной информации. В связи с этим обстоятельством возникает проблема построения методов задания функций, которые были бы не слишком чувствительны по отношению к такого рода искажениям.

Рассмотрим сначала класс формул в базисе конъюнкция, дизъюнкция, отрицание, т.е. класс дизъюнктивных нормальных форм (ДНФ). Будем кодировать ДНФ от n переменных словами длины $2n$ в алфавите $B = \{0, 1\}$ следующим образом.

Пусть $X_n = \{x_1\bar{x}_1 \ldots x_n\bar{x}_n\}$ – упорядоченный алфавит булевых переменных. Каждой элементарной конъюнкции $\mathfrak{A}$ сопоставим слово из нулей и единиц следующим способом: 1 характеризует вхождение буквы в $\mathfrak{A}$, а 0 – ее отсутствие в этой конъюнкции. Код конъюнкции обозначим через $C(\mathfrak{A})$.

Пример 22. В алфавите переменных $X_3 = (x_1\bar{x}_1x_2\bar{x}_2x_3\bar{x}_3)$ конъюнкция $\mathfrak{A} = x_1\bar{x}_3$ имеет код $C(\mathfrak{A}) = (100001)$.

Замечание 5. Пустой конъюнкции сопоставим код, являющийся словом из одних нулей.

Определение 7 (см. [17]). Расстоянием между элементарными конъюнкциями $\mathfrak{A}$ и $\mathfrak{M}$ называется расстояние Хэмминга между кодами этих конъюнкций, т.е.

$$r_H(\mathfrak{A}, \mathfrak{M}) = r_H(C(\mathfrak{A}), C(\mathfrak{M})).$$

Фактически расстояние $r_H(\mathfrak{A}, \mathfrak{M})$ между элементарными конъюнкциями $\mathfrak{A}$ и $\mathfrak{M}$ – это минимальное число вставок и выпадений букв, требующихся для преобразования одной конъюнкции в другую.

Пример 23. Если $X_3 = (x_1\bar{x}_1x_2\bar{x}_2x_3\bar{x}_3)$, $\mathfrak{A} = x_1\bar{x}_3$, $\mathfrak{M} = x_1x_2x_3$, то $C(\mathfrak{A}) = (100001)$, $C(\mathfrak{M}) = (101010)$ и $r_H(\mathfrak{A}, \mathfrak{M}) = 3$. При этом ясно, что $\mathfrak{A}$ можно преобразовать в $\mathfrak{M}$ путем вставки x_2, удаления $\bar{x}_3$, вставки x_3.

Определим теперь расстояние между двумя ДНФ. Так как пустой конъюнкции отвечает слово из одних нулей, то можно считать, что заданные ДНФ Z_1 и Z_2 содержат одинаковое число элементарных конъюнкций.

Пусть $Z_1 = \bigvee_{i=1}^{m} u_i$, $Z_2 = \bigvee_{i=1}^{m} v_i$.

Определение 8. *Расстоянием между* ДНФ Z_1 и Z_2 называется следующая величина:

$$r_L(Z_1, Z_2) = \min_{g \in S_m} \sum_{i=1}^{m} r_H(C(\nu_i), C(\nu_{g(i)})). \tag{24}$$

Минимум в (24) берется по всем g перестановкам из симметрической группы S_m.

В стандартных терминах теории графов задача (24) нахождения расстояния между двумя ДНФ – это задача о построении минимального паросочетания в полном взвешенном двудольном графе, ребрам которого приписаны веса, равные расстояниям между соответствующими элементарными конъюнкциями.

Пример 24. Пусть $Z_1 = \bar{x}_1\bar{x}_2 \vee \bar{x}_1x_2 \vee x_1x_2$, $Z_2 = \bar{x}_1 \vee x_2$. Тогда $C(Z_1) = \{(0101), (0110), (1010)\}$, $C(Z_2) = \{(0100), (0010), (0000)\}$. Легко проверить, что

$$r_H(Z_1, Z_2) = r_H[(1010), (0000)] + r_H[(0110), (0010)] + r_H[(0101), (0100)] = 4.$$

Пример 25. Пусть $Z_1 = \bigvee_{i=1}^{n} x_i$, $Z_2 = \bar{x}_1 \ldots \bar{x}_n$; тогда прямое вычисление показывает, что $r_H(Z_1, Z_2) = 2n$.

Пусть $f(x_1 \ldots x_n)$ – произвольная булева функция и Z_f – множество всех ДНФ реализующих функцию f. Если $\nu \in Z_f$, то искажение формулы ν может привести к ДНФ $\nu' \notin Z_f$.

Определение 9. Число

$$t(f) = R_H(Z_f, Z_{\bar{f}})$$

называется *уровнем устойчивости* функции f.

Содержательно число $t(f)$ указывает на тот показатель минимальных изменений формул из Z_f, которые могут перевести функцию f в ее отрицание $\bar{f}$. Хрестоматийный пример рассматриваемой ситуации содержится в высказывании “Казнить нельзя, помиловать”. Если перенести запятую, то смысл высказывания изменится на противоположный.

Пример 26. Если f – дизъюнкция, то $\bar{f} = \bar{x}_1 \ldots \bar{x}_n$. Согласно примеру 25, если формула для f имеет вид $Z_1 = x_1 \vee \ldots \vee x_n$, то при $Z_2 = \bar{x}_1 \ldots \bar{x}_n$

$$r_L(z_1, z_2) = 2n.$$

Таким образом $r_L(f, \bar{f}) \leqslant 2n$. С другой стороны, $Z_{\bar{f}} = \{\bar{x}_1 \ldots \bar{x}_n\}$ и $|Z_{\bar{f}}| = 1$. Минимальная и кратчайшая ДНФ для дизъюнкции – это z_1. Поэтому $z_L(Z_f, Z_{\bar{f}}) \geqslant r_L(z_1, z_2) = 2n$. Отсюда следует, что $t(f) = 2n$.

СПИСОК ЛИТЕРАТУРЫ

1. *Воронин Ю.А.* Начала теории сходства Новосибирск: ВЦСО, 1991. С. 3–127.
2. *Добрынин А.А., Гутман И.* Индекс Винера для деревьев и графов гексагональных систем // Дискретный анализ и иссл. операций. Сер. 2. 1998. Т. 5. № 2. С. 34–60.
3. *Леонтьев В.К., Мовсисян Г.Л.* О некоторых аспектах понятия информация // Компьюлог. 2005. № 1 (67). С. 17–21.
4. *Левенштейн В.И.* Элементы теории кодирования // Дискретная матем. и матем. вопросы кибернетики. М.: Наука, 1974. С. 207–305.
5. *Леонтьев В.К.* Задачи восстановления слов по фрагментам и их приложения // Дискретный анализ и иссл. операций. 1995. Т. 2. № 2. С. 26–48.
6. *Журавлев Ю.И.* Теоретико-множественные методы в алгебре логики // Пробл. кибернетики. 1962. № 8.
7. *Журавлев Ю.И.* Окрестности в задачах дискретной математики. Избр. труды. М.: Магистр, 1988.
8. *Бурдюк В.Я.* Дискретные метрические пространства. Днепропетровск: ДГУ, 1982. С. 1–99.
9. *Витушкин А.Г.* Оценка сложности задачи табулирования. М.: Физматгиз, 1959.

10. *Леонтьев В.К.* О гранях единичного n-мерного куба // Ж. вычисл. матем. и матем. физ. 2008. Т. 48. № 6. С. 1126–1139.
11. *Евдокимов А.А.* Локально-изометрические вложения графов и свойства продолжения метрики // Сибирский ж. иссл. операций. 1994. Т. 1. № 1. С. 5–12.
12. *Мак-Вильямс Ф. Дж., Слоэн Н.Дж.* Теория кодов, исправляющих ошибки. М.: Связь, 1979. С. 1–744.
13. *Федоряева Т.И.* Векторы разнообразия шаров для графов и оценки их компонент // Дискретный анализ и иссл. операций. Сер. 1. 2006. Т. 14. № 2. С. 47–67.
14. *Биркгоф Г., Барти Т.* Современная прикладная алгебра. М.: Мир, 1976.
15. *Леонтьев В.К., Мовсисян Г.Л.* Об аддитивных каналах связи // Докл. НАН Армении. 2004. Т. 104. № 1. С. 23–28.
16. *Журавлев Ю.И.* Об алгебраическом подходе к решению задач распознавания или классификации // Пробл. кибернетики. 1978. № 33.
17. *Леонтьев В.К.* Избранные задачи комбинаторного анализа. М.: Изд-во МГТУ им. Н.Э. Баумана, 2001.

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2009, том 49, № 11, с. 2059–2065

УДК 519.71

О СЛОЖНОСТИ НЕКОТОРЫХ ЗАДАЧ ПОИСКА ПОДМНОЖЕСТВ ВЕКТОРОВ И КЛАСТЕРНОГО АНАЛИЗА[1)]

© 2009 г. А. В. Кельманов, А. В. Пяткин

(630090 Новосибирск, пр-т Акад. Коптюга, 4, Ин-т матем. СО РАН)

e-mail: {kelm, artem}@math.nsc.ru

Поступила в редакцию 21.10.2008 г.

Доказана NP-полнота дискретных экстремальных задач, к которым сводятся некоторые варианты проблемы поиска подмножеств векторов и кластерного анализа. Библ. 16.

Ключевые слова: дискретная экстремальная задача, сложность, NP-полнота, поиск подмножеств, кластерный анализ, распознавание образов.

ВВЕДЕНИЕ

Объектом исследования являются проблемы оптимизации в задачах анализа данных и распознавания образов. Предмет исследования – дискретные экстремальные задачи, к которым сводятся некоторые варианты проблемы поиска подмножеств векторов и кластерного анализа. Цель работы – исследование сложности этих ранее не изученных задач.

К рассмотренным далее экстремальным задачам сводятся важные содержательные задачи, возникающие в приложениях, связанных с off-line анализом и распознаванием массивов зашумленных структурированных данных (числовых последовательностей, временных рядов, сигналов), включающих повторяющиеся, чередующиеся или перемежающиеся информационно значимые блоки – векторы или фрагменты одинаковой размерности, в случае когда места расположения этих векторов (фрагментов) в массиве неизвестны. Ситуации, в которых требуется решение этих задач, характерны, в частности, для электронной разведки и дистанционного зондирования, геофизики и биометрики, медицинской и технической диагностики, обработки изображений и речевых сигналов, радиолокации, гидроакустики, телекоммуникации, криминалистики, поиска по мультимедийным базам данных, обработки результатов эксперимента и др. (см., например, [1]–[5] и цитированные там работы).

Сущность проблемы состоит в обнаружении информационно значимых векторов (или фрагментов) в зашумленных массивах данных, оценивании этих векторов и принятии решения (классификации) по результатам обнаружения и оценивания. В данной работе показано, что экстремальные задачи, к которым сводятся некоторые важные для приложений варианты этой проблемы, относятся к классу труднорешаемых задач. Эти задачи являются результатом сведения соответствующих формализованных содержательных задач помехоустойчивого анализа данных и распознавания образов, сформулированных в виде задач максимизации функционала правдоподобия (в случае когда помеха аддитивна и является гауссовской последовательностью независимых одинаково распределенных случайных величин, см. [6]) или в виде задач среднеквадратического приближения (см. [7]).

1. БАЗОВЫЕ ЗАДАЧИ ПОИСКА ПОДМНОЖЕСТВА ВЕКТОРОВ

Приведем известные факты о простейших редуцированных экстремальных задачах, напрямую связанных с решением упомянутой проблемы. Сформулируем эти задачи в форме задач верификации свойств.

[1)] Работа выполнена при финансовой поддержке РФФИ (коды проектов 07-07-00022, 08-01-00516 и 09-01-00032).

Задача MLSVS (максимум длины суммы векторов из подмножества фиксированной мощности)

Дано: множество $V = \{\mathbf{v}_1, ..., \mathbf{v}_L\}$ векторов из $\mathbb{R}^q$, натуральное число M и положительное число K. Вопрос: существует ли такое подмножество $B \subseteq V$, мощность которого равна M, что имеет место неравенство

$$\left\| \sum_{\mathbf{v} \in B} \mathbf{v} \right\| \geq K?$$

К этой задаче сводится поиск в множестве векторов евклидова пространства подмножества векторов, похожих между собой по критерию минимума суммы квадратов уклонений, при условии, что мощность искомого подмножества задана (см. [6]–[11]).

Теорема 1 (см. [8], [9]). *Задача MLSVS* NP-*полна.*

При доказательстве к частному случаю этой задачи сводится NP-полная задача CLIQUE (см. [12]).

Задача MALSSVS (максимум среднего значения квадрата длины суммы векторов из подмножества)

Дано: множество $V = \{\mathbf{v}_1, ..., \mathbf{v}_L\}$ векторов из $\mathbb{R}^q$ и положительное число K. Вопрос: существует ли непустое подмножество $B \subseteq V$ такое, что имеет место неравенство

$$\frac{1}{|B|}\left\| \sum_{\mathbf{v} \in B} \mathbf{v} \right\|^2 \geq K?$$

К этой задаче сводится поиск в множестве векторов евклидова пространства подмножества векторов, похожих между собой по критерию минимума суммы квадратов уклонений, когда мощность искомого подмножества неизвестна (см. [6]–[11]).

Теорема 2 (см. [10], [11]). *Задача MALSSVS* NP-*полна.*

При доказательстве к частному случаю этой задачи сводится NP-полная задача 3-SAT (см. [12]).

Основным результатом настоящей работы является установление статуса NP-полноты для четырех задач, сформулированных ниже в разд. 2.

2. ЗАДАЧИ ПОИСКА ПОДМНОЖЕСТВ ВЕКТОРОВ И КЛАСТЕРНОГО АНАЛИЗА

Проанализируем сначала две задачи, которые можно трактовать как обобщения задач MLSVS и MALSSVS на случай поиска нескольких подмножеств. Чтобы пояснить истоки этих задач, рассмотрим следующую модель анализа данных.

Пусть векторная последовательность $\mathbf{x}_n \in \mathbb{R}^q$, $n \in \mathbb{N}$, где $\mathbb{N} = \{1, 2, ..., N\}$, обладает свойством

$$\mathbf{x}_n = \begin{cases} \mathbf{w}_1, & n \in \mathbb{M}_1, \\ \ldots, & \ldots, \\ \mathbf{w}_J, & n \in \mathbb{M}_J, \\ \mathbf{0}, & n \in \mathbb{N} \setminus \bigcup\limits_{j=1}^{J} \mathbb{M}_j, \end{cases} \tag{2.1}$$

где $\bigcup_{j=1}^{J} \mathbb{M}_j \subseteq \mathbb{N}$, причем $\mathbb{M}_i \cap \mathbb{M}_j = \varnothing$, если $i \neq j$.

Положим $|\mathbb{M}_j| = M_j, j = 1, 2, ..., J$. В формуле (2.1) векторы $\mathbf{w}_j \in \mathbb{R}^q$ будем интерпретировать как информационно-значимые векторы, а M_j – как число их повторов. Рассмотрим аддитивную модель помехи (ошибок наблюдения). Доступной для обработки будем считать последовательность

$$\mathbf{v}_n = \mathbf{x}_n + \mathbf{e}_n, \quad n \in \mathbb{N},$$

где $\mathbf{e}_n$ – вектор помехи (ошибки), независимый от вектора $\mathbf{x}_n$. Положим

$$S(\mathbb{M}_1, \ldots, \mathbb{M}_J, \mathbf{w}_1, \ldots, \mathbf{w}_J) = \sum_{n \in \mathbb{N}} \|\mathbf{v}_n - \mathbf{x}_n\|^2 \tag{2.2}$$

и рассмотрим следующую задачу среднеквадратического приближения.

Дана последовательность $\mathbf{v}_n \in \mathbb{R}^q$, $n \in \mathbb{N}$. Найти: семейство $\{\mathbb{M}_1, ..., \mathbb{M}_J\}$ непустых непересекающихся подмножеств множества $\mathbb{N}$ и совокупность векторов $\{\mathbf{w}_1, ..., \mathbf{w}_J\}$, минимизирующих $S(\cdot)$.

Эта задача соответствует сформулированной во Введении проблеме совместного обнаружения и оценивания по критерию минимума суммы квадратов уклонений ненулевых неизвестных информационно-значимых векторов, повторяющихся в ненаблюдаемой последовательности (2.1). Задачу можно также трактовать, как поиск семейства непересекающихся подмножеств векторов, похожих в среднеквадратическом. Нетрудно установить, что к аналогичной формулировке можно прийти, если считать, что вектор $\mathbf{e}_n$ есть выборка из q-мерного нормального распределения с параметрами $(\mathbf{0}, \sigma^2\mathbf{I})$, где $\mathbf{I}$ – единичная матрица, а в качестве критерия решения использовать традиционный для статистики максимум функционала правдоподобия.

Заметим, что в частном случае, когда семейство $\{\mathbb{M}_1, ..., \mathbb{M}_J\}$ известно, эта задача вырождается в классическую задачу оценивания средних значений векторов $\mathbf{w}_1, ..., \mathbf{w}_J$. В рассматриваемом общем случае минимум функционала (2.2) по независимым неизвестным векторам $\mathbf{w}_1, ..., \mathbf{w}_J$ находится аналитически. Нетрудно убедиться, что этот минимум доставляется векторами $\bar{\mathbf{w}}_j = \sum_{n \in \mathbb{M}_j} \mathbf{v}_n / M_j$, $j = 1, 2, ..., J$, и равен

$$S_{\min}(\mathbb{M}_1, \ldots, \mathbb{M}_J) = \sum_{n \in \mathbb{N}} \|\mathbf{v}_n\|^2 - \sum_{j=1}^{J} \frac{1}{M_j} \left\| \sum_{n \in \mathbb{M}_j} \mathbf{v}_n \right\|^2. \tag{2.3}$$

Первый член в правой части равенства (2.3) является константой. Поэтому в форме верификации свойств имеем следующие редуцированные дискретные экстремальные задачи.

Задача *J*-SSAF (поиск J таких непересекающихся подмножеств векторов, что сумма средних значений квадратов длин сумм векторов из этих подмножеств максимальна при условии, что мощности подмножеств фиксированы)

Дано: множество $\tilde{V} = \{\mathbf{v}_1, ..., \mathbf{v}_N\}$ векторов из $\mathbb{R}^q$ и положительное число $\tilde{K}$. Вопрос: существует ли такое семейство $\{B_1, ..., B_J\}$ непустых непересекающихся подмножеств множества $\tilde{V}$, имеющих фиксированные мощности $M_1, ..., M_J$, что имеет место неравенство

$$\sum_{j=1}^{J} \frac{1}{M_j} \left\| \sum_{\mathbf{v} \in B_j} \mathbf{v} \right\|^2 \geq \tilde{K}?$$

Задача *J*-SSANF (поиск J таких непересекающихся подмножеств векторов, что сумма средних значений квадратов длин сумм векторов из этих подмножеств максимальна при условии, что мощности подмножеств не фиксированы)

Дано: множество $\tilde{V} = \{\mathbf{v}_1, ..., \mathbf{v}_N\}$ векторов из $\mathbb{R}^q$ и положительное число $\tilde{K}$. Вопрос: существует ли такое семейство $\{B_1, ..., B_J\}$ непустых непересекающихся подмножеств множества $\tilde{V}$, что имеет место неравенство

$$\sum_{j=1}^{J} \frac{1}{|B_j|} \left\| \sum_{\mathbf{v} \in B_j} \mathbf{v} \right\|^2 \geq \tilde{K}? \tag{2.4}$$

Теорема 3. *Задача J-SSAF* NP-*полна.*

Доказательство проводится аналогично приведенному ниже доказательству теоремы 4.

Теорема 4. *Задача J-SSANF* NP-*полна.*

Доказательство. Принадлежность этой задачи к классу NP очевидна. Если J является частью входа, то достаточно заметить, что частным случаем этой задачи (когда $J = 1$) является рассмотренная выше в разд. 1 NP-полная задача MALSSVS.

Пусть J не является частью входа. Покажем NP-полноту задачи для $J = 2$. В общем случае сведение J-SSANF к $(J + 1)$-SSANF строится аналогично.

Рассмотрим множество $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_L\}$ векторов из $\mathbb{R}^q$ и положительное число K из задачи MALSSVS. Обозначим через a наибольшую по модулю координату векторов из V. Без ограничения общности будем считать, что $a > 0$.

Положим в задаче 2-SSANF множество $\tilde{V} = V \cup \{\mathbf{x}\}$ и число $\tilde{K} = K + \|\mathbf{x}\|^2$, где $\mathbf{x} = (La + 1, La + 1, \ldots, La + 1) \in \mathbb{R}^q$. Тогда легко видеть, что если в задаче MALSSVS подмножество B существует, то и в задаче 2-SSANF существуют подмножества $B_1 = B$ и $B_2 = \{\mathbf{x}\}$, удовлетворяющие неравенству (2.4). В самом деле, имеем

$$\left\|\sum_{\mathbf{v}\in B_1}\mathbf{v}\right\|^2/|B_1| + \left\|\sum_{\mathbf{v}\in B_2}\mathbf{v}\right\|^2/|B_2| = \left\|\sum_{\mathbf{v}\in B}\mathbf{v}\right\|^2/|B| + \|\mathbf{x}\|^2 \geq K + \|\mathbf{x}\|^2 = \tilde{K}.$$

Чтобы доказать обратную импликацию, покажем сначала, что для всякого $X \subseteq \tilde{V}$ выполняется неравенство

$$\left\|\sum_{\mathbf{v}\in X}\mathbf{v}\right\|^2/|X| \leq \|\mathbf{x}\|^2. \tag{2.5}$$

Действительно, если $X = \{\mathbf{x}\}$, то, очевидно, в (2.5) имеет место равенство. Далее, если $\mathbf{x} \notin X$, то

$$\left\|\sum_{\mathbf{v}\in X}\mathbf{v}\right\|^2/|X| \leq qL^2a^2 < q(La + 1)^2 = \|\mathbf{x}\|^2.$$

Если же $\mathbf{x} \in X$ и $|X| \geq 2$, то, учитывая, что $|X| \leq L + 1$, имеем

$$\left\|\sum_{\mathbf{v}\in X}\mathbf{v}\right\|^2/|X| \leq q[(|X| - 1)a + (La + 1)]^2/|X| < q(La + 1)^2(\sqrt{|X|}/L + 1/\sqrt{|X|})^2 \leq$$

$$\leq \|\mathbf{x}\|^2(\max\{(\sqrt{2}/L + 1/\sqrt{2}, \sqrt{L + 1}/L + 1/\sqrt{L + 1})\})^2 \leq \|\mathbf{x}\|^2$$

при $L > 4$.

Пусть теперь для семейства подмножеств $\{B_1, B_2\}$ в задаче 2-SSANF выполнено неравенство (2.4). Без ограничения общности будем считать, что $\mathbf{x} \notin B_1$. Тогда из (2.5) следует, что

$$\left\|\sum_{\mathbf{v}\in B_1}\mathbf{v}\right\|^2/|B_1| \geq \left\|\sum_{\mathbf{v}\in B_1}\mathbf{v}\right\|^2/|B_1| + \left\|\sum_{\mathbf{v}\in B_2}\mathbf{v}\right\|^2/|B_2| - \|\mathbf{x}\|^2 \geq \tilde{K} - \|\mathbf{x}\|^2 = K,$$

т.е. ответ в задаче MALSSVS также положителен: подмножество $B = B_1$ существует. Теорема 4 доказана.

Перед анализом сложности двух следующих задач напомним широко известную (см., например, [13]–[15]) задачу MSSC – кластеризация (разбиение) множества векторов евклидова пространства на кластеры или подмножества по критерию минимума суммы квадратов. В форме верификации свойств эта задача формулируется в следующем виде.

Задача MSSC

Дано: множество $V = \{\mathbf{v}_1, ..., \mathbf{v}_N\}$ векторов из $\mathbb{R}^q$, натуральное число $J > 1$ и положительное число K. Вопрос: существует ли такое разбиение множества V на непустые подмножества (кластеры) $C_1, ..., C_J$, что имеет место неравенство

$$\sum_{j=1}^{J} \sum_{\mathbf{v} \in C_J} \|\mathbf{v} - \bar{\mathbf{w}}_j\|^2 \le K,$$

где $\bar{\mathbf{w}}_j = \Sigma_{\mathbf{v} \in C_j} \mathbf{v} / |C_j|$, $j = 1, 2, ..., J$, – центры кластеров?

Нетрудно заметить, что эта задача возникает в том случае, когда вместо последовательности вида (2.1) рассматривается последовательность, обладающая свойством

$$\mathbf{x}_n = \begin{cases} \mathbf{w}_1, & n \in \mathbb{M}_1, \\ \ldots, & \ldots, \\ \mathbf{w}_J, & n \in \mathbb{M}_J, \end{cases}$$

где $\bigcup_{j=1}^{J} \mathbb{M}_j = \mathbb{N}$, причем $\mathbb{M}_i \cap \mathbb{M}_j = \varnothing$, если $i \ne j$.

На протяжении многих лет задача MSSC в оптимизационном варианте бездоказательно считалась NP-трудной. В опубликованном недавно (см. [13]) доказательстве ее NP-полноты впоследствии была обнаружена ошибка (см. [14]). Корректное доказательство труднорешаемости этой задачи для случая, когда $J = 2$, опубликовано совсем недавно в [15]. Ниже показана NP-полнота двух важных специальных случаев задачи MSSC.

Задача *J*-MSSCF (кластеризация по критерию минимума суммы квадратов при условии, что мощности кластеров фиксированы, и центр одного из кластеров определять не требуется)

Дано: множество $\tilde{V} = \{\mathbf{v}_1, ..., \mathbf{v}_N\}$ векторов из $\mathbb{R}^q$, натуральное число $J > 1$ и положительное число $\tilde{K}$. Вопрос: существует ли такое разбиение множества $\tilde{V}$ на непустые подмножества (кластеры) $C_1, ..., C_J$, имеющих фиксированные мощности $M_1, ..., M_J$, что имеет место неравенство

$$\sum_{j=1}^{J-1} \sum_{\mathbf{v} \in C_j} \|\mathbf{v} - \bar{\mathbf{w}}_j\|^2 + \sum_{\mathbf{v} \in C_J} \|\mathbf{v}\|^2 \le \tilde{K},$$

где $\bar{\mathbf{w}}_j = \Sigma_{\mathbf{v} \in C_j} \mathbf{v} / |M_j|$, $j = 1, 2, ..., J-1$, – центры кластеров?

Теорема 5. *Задача J-MSSCF* NP-*полна.*

К этой задаче сводится (см. далее) задача *J*-SSAF.

Задача *J*-MSSCNF (кластеризация по критерию минимума суммы квадратов при условии, что мощности кластеров не фиксированы, и центр одного из кластеров определять не требуется)

Дано: множество $\tilde{V} = \{\mathbf{v}_1, ..., \mathbf{v}_N\}$ векторов из $\mathbb{R}^q$, натуральное число $J > 1$ и положительное число $\tilde{K}$. Вопрос: существует ли такое разбиение множества $\tilde{V}$ на непустые подмножества (кластеры) $C_1, ..., C_J$, что имеет место неравенство

$$\sum_{j=1}^{J-1} \sum_{\mathbf{v} \in C_j} \|\mathbf{v} - \bar{\mathbf{w}}_j\|^2 + \sum_{\mathbf{v} \in C_J} \|\mathbf{v}\|^2 \le \tilde{K},$$

где $\bar{\mathbf{w}}_j = \Sigma_{\mathbf{v} \in C_j} \mathbf{v} / |C_j|$, $j = 1, 2, ..., J-1$, – центры кластеров?

К этой задаче сводится задача *J*-SSANF. Справедлива

Теорема 6. *Задача J-MSSCNF* NP-*полна.*

Доказательство. Доказательство NP-полноты задачи J-MSSCNF продемонстрируем для случая двух кластеров ($J = 2$). Пусть, для определенности, в задаче 2-MSSCNF центр второго кластера определять не требуется. Тогда для целевой функции задачи 2-MSSCNF имеем цепочку равенств

$$\sum_{\mathbf{v} \in C_1} \|\mathbf{v} - \bar{\mathbf{w}}_1\|^2 + \sum_{\mathbf{v} \in C_2} \|\mathbf{v}\|^2 = \sum_{\mathbf{v} \in C_1} \|\mathbf{v}\|^2 + |C_1|\|\bar{\mathbf{w}}_1\|^2 - 2\sum_{\mathbf{v} \in C_1} (\mathbf{v}, \bar{\mathbf{w}}_1) + \sum_{\mathbf{v} \in C_2} \|\mathbf{v}\|^2 =$$

$$= \sum_{\mathbf{v} \in V} \|\mathbf{v}\|^2 + \frac{\left\|\Sigma_{\mathbf{v} \in C_1}\mathbf{v}\right\|^2}{|C_1|} - 2\frac{\Sigma_{\mathbf{v} \in C_1}(\mathbf{v}, \Sigma_{\mathbf{u} \in C_1}\mathbf{u})}{|C_1|} = \sum_{\mathbf{v} \in V} \|\mathbf{v}\|^2 - \frac{\left\|\Sigma_{\mathbf{v} \in C_1}\mathbf{v}\right\|^2}{|C_1|}.$$

Отсюда следует, что частный случай задачи J-SSANF (т.е. NP-полная задача MALSSVS) сводится к задаче 2-MSSCNF. Причем разбиение семейства векторов $\tilde{V}$ на кластеры C_1 и C_2 в задаче 2-MSSCNF существует тогда и только тогда, когда в задаче MALSSVS при $K = = \Sigma_{\mathbf{v} \in V}\|\mathbf{v}\|^2 - \tilde{K}$ существует соответствующее подмножество векторов B. Теорема 6 доказана.

Из приведенной цепочки равенств аналогичным образом легко установить, что частный случай задачи J-SSAF (т.е. NP-полная задача MLSVS) сводится к частному случаю задачи J-MSSCF. При этом разбиение семейства векторов $\tilde{V}$ на кластеры C_1 и C_2 в задаче J-MSSCF существует тогда и только тогда, когда в задаче MLSVS при $K^2/|B| = \Sigma_{\mathbf{v} \in V}\|\mathbf{v}\|^2 - \tilde{K}$ существует соответствующее подмножество векторов B. Отсюда следует справедливость теоремы 5.

Остается заметить, что задачи J-MSSCF и J-MSSCNF можно трактовать как задачи помехоустойчивого кластерного анализа. В этой трактовке векторы, соответствующие аддитивному шуму или помехе, ассоциируются с тем кластером, центр которого определять не требуется.

ЗАКЛЮЧЕНИЕ

Рассмотренные экстремальные задачи являются простейшими в семействе всевозможных задач, к которым сводятся проблемы off-line анализа и распознавания зашумленных числовых последовательностей, включающих какие-либо структуры над информационно значимыми векторами (фрагментами). Это семейство в настоящее время включает несколько сотен задач (см. [7], [16]). Для части задач из этого семейства установлена полиномиальная разрешимость. Для другой части доказана NP-трудность. Однако большинство задач из этого семейства не изучены, статус их сложности не выяснен и какие-либо алгоритмы с гарантированными оценками точности для их решения неизвестны. Масштабность проблемы можно оценить, заглянув в интернет (см. [16]). Значимость изложенных в данной работе результатов состоит в том, что они, как базовые, позволяют устанавливать доказуемый статус алгоритмической сложности других экстремальных задач из этого семейства (и вместе с этим статус сложности соответствующих задач анализа данных и распознавания образов), используя известную (см. [12]) технику полиномиальной сводимости.

СПИСОК ЛИТЕРАТУРЫ

1. *Кельманов А.В., Хамидуллин С.А.* Апостериорное обнаружение заданного числа одинаковых подпоследовательностей в квазипериодической последовательности // Ж. вычисл. матем. и матем. физ. 2001. Т. 41. № 5. С. 807–820.
2. *Kel'manov A.V., Jeon B.* A posteriori joint detection and discrimination of pulses in a quasiperiodic pulse train // IEEE Transactions on Signal Processing. 2004. V. 52. № 3. P. 1–12.
3. *Кельманов А.В., Михайлова Л.В.* Совместное обнаружение в квазипериодической последовательности заданного числа фрагментов из эталонного набора и ее разбиение на участки, включающие серии одинаковых фрагментов // Ж. вычисл. матем. и матем. физ. 2006. Т. 46. № 1. С. 172–189.
4. *Кельманов А.В., Михайлова Л.В.* Апостериорное обнаружение квазипериодических фрагментов из эталонного набора в числовой последовательности // Ж. вычисл. матем. и матем. физ. 2008. Т. 48. № 5. С. 168–184.
5. *Кельманов А.В., Михайлова Л.В., Хамидуллин С.А.* Апостериорное обнаружение в квазипериодической последовательности повторяющегося набора эталонных фрагментов // Ж. вычисл. матем. и матем. физ. 2008. Т. 48. № 12. С. 168–184.

6. *Кельманов А.В.* Полиномиально разрешимые и NP-трудные варианты задачи оптимального обнаружения в числовой последовательности повторяющегося фрагмента // Материалы Рос. конф. “Дискретная оптимизация и исследование операций”. Владивосток, 7–14 сентября 2007. Новосибирск: ИМ СО РАН, 2007. http://math.nsc.ru/conference/door07/DOOR_abstracts.pdf. С. 46–50.
7. *Кельманов А.В.* Проблема off-line обнаружения повторяющегося фрагмента в числовой последовательности // Тр. ИММ УРО РАН. Екатеринбург. 2008. Т. 14. № 2. С. 81–88.
8. *Гимади Э.Х., Кельманов А.В., Кельманова М.А., Хамидуллин С.А.* Апостериорное обнаружение в числовой последовательности квазипериодического фрагмента при заданном числе повторов // Сибирский журнал индустр. матем. 2006. Т. 9. № 1(25). С. 55–74.
9. *Бабурин А.Е., Гимади Э.Х., Глебов Н.И., Пяткин А.В.* Задача отыскания подмножества векторов с максимальным суммарным весом // Дискретный анализ и иссл. операций. Сер. 2. 2007. Т. 14. № 1. С. 32–42.
10. *Кельманов А.В., Пяткин А.В.* О сложности одного из вариантов задачи выбора подмножества “похожих” векторов // Докл. РАН. 2008. Т. 421. № 5. С. 590–592.
11. *Кельманов А.В., Пяткин А.В.* Об одном варианте задачи выбора подмножества векторов // Дискретный анализ и иссл. операций. 2008. Т. 15. № 5. С. 25–40.
12. *Garey M.R., Johnson D.S.* Computers and intractability: a guide to the theory of NP-completeness. Freeman, San Francisco, CA, 1979.
13. *Drineas P., Frieze A., Kannan R., Vinay V.* Clustering large graphs via the singular value decomposition // Machine Learning. 2004. V. 56. P. 9–33.
14. *Aloise D., Hansen P.* On the complexity of minimum sum-of-squares clustering // Les Cahiers du GERAD, G-2007-50. 2007. 12 p.
15. *Aloise D., Deshpande A., Hansen P., Popat P.* NP-Hardness of Euclidean sum-of-squares clustering // Les Cahiers du GERAD, G-2008-33. 2008. 4 p.
16. http://math.nsc.ru/~serge/qpsl/

ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2009, том 49, № 11, с. 2066–2080

УДК 519.71

ЭФФЕКТИВНЫЙ МЕТОД ОТБОРА ПРИЗНАКОВ В ЛИНЕЙНОЙ РЕГРЕССИИ С ПОМОЩЬЮ ОБОБЩЕНИЯ ИНФОРМАЦИОННОГО КРИТЕРИЯ АКАИКЕ[1)]

© 2009 г. Д. П. Ветров*, Д. А. Кропотов, Н. О. Пташко***

*(*119992 Москва, Ленинские горы, МГУ ВМиК;*
***119333 Москва, ул. Вавилова, 40, ВЦ РАН)*
e-mail: vetrovd@yandex.ru; dkropotov@yandex.ru; ptashko@inbox.ru

Поступила в редакцию 12.05.2009 г.

Предлагается метод отбора признаков для линейной регрессии с помощью обобщения информационного критерия Акаике. Использование классического информационного критерия Акаике (ИКА) для отбора признаков связано с полным перебором по всем подмножествам признаков, что приводит к неоправданно большим вычислительным и временным затратам. Предлагается новый информационный критерий, который является непрерывным обобщением ИКА. В результате задача отбора признаков сводится к задаче гладкой оптимизации. Выводится эффективная процедура решения полученной задачи оптимизации. Экспериментальные исследования показывают, что разработанный метод действительно позволяет быстро и эффективно отбирать признаки в линейной регрессии. В экспериментах новая процедура также сравнивается с методом релевантных векторов, который является методом отбора признаков на основе байесовского подхода. Показано, что обе процедуры близки по результатам. Основное отличие нового метода состоит в том, что некоторые коэффициенты регуляризации становятся тождественно равными нулю. Это позволяет избежать эффекта переупрощения модели, который характерен для метода релевантных векторов. Также рассматривается специальный случай (так называемая недиагональная регуляризация), в котором оба метода оказываются идентичными. Библ. 18. Фиг. 4. Табл. 2.

Ключевые слова: распознавание образов, линейная регрессия, отбор признаков, информационный критерий Акаике.

1. ВВЕДЕНИЕ

Байесовские методы широко используются для построения процедур автоматического выбора модели. Метод релевантных векторов (МРВ), предложенный в [1], является одним из примеров применения байесовской парадигмы в задаче линейной регрессии. В МРВ с каждым весом регрессора в линейном решающем правиле связывается индивидуальный коэффициент регуляризации (L2-регуляризация). Коэффициенты регуляризации подбираются автоматически путем максимизации правдоподобия модели (обоснованности). В результате такой процедуры, известной также как АОР (автоматическое определение релевантности (см. [2])), большинство коэффициентов регуляризации становятся равными бесконечности, что соответствует обнулению весов регрессоров и удалению соответствующих им регрессоров из модели. L1-Регуляризация, использующая общий коэффициент регуляризации (см. [3]) либо использующая несобственное априорное распределение Джеффри с последующим интегрированием по коэффициентам регуляризации, также показывают высокую разреженность (число нулевых весов), сохраняя при этом хорошую обобщающую способность (см. [4]–[6]).

Известный информационный критерий Байеса–Шварца (см. [7]) может быть рассмотрен как грубая аппроксимация логарифма маргинального правдоподобия (обоснованности, см. [8]). Также для решения задач выбора моделей широко применяется информационный критерий Акаике (ИКА, см. [9]), предлагающий альтернативный подход, основанный на теории информации. Несмотря на то, что этот метод был изначально предложен для выбора из конечного числа моделей, он может быть расширен на случай континуального семейства моделей. В работе пред-

[1)] Работа выполнена при финансовой поддержке РФФИ (коды проектов 08-01-00405, 08-01-90016, 08-01-90427, 07-01-00211).

ложено подобное обобщение информационного критерия Акаике для выбора модели с дальнейшим применением в задаче линейной регрессии. Подбор коэффициентов регуляризации, связанных индивидуально с каждым весом, производится путем максимизации непрерывного аналога критерия Акаике (ОИКА). Особый интерес представляет разреженность решений, получаемых с помощью нового метода, и сравнение нового метода с классическим МРВ.

В разд. 2 приведен вывод непрерывного аналога критерия Акаике (ОИКА – обобщенный информационный критерий Акаике). В разд. 3 показано применение критерия ОИКА к задаче обобщенной линейной регрессии и выведены итеративные формулы пересчета коэффициентов регуляризации. В разд. 4 представлены результаты экспериментов и проведено сравнение ОИКА с классическим МРВ.

2. ОБОБЩЕНИЕ ИНФОРМАЦИОННОГО КРИТЕРИЯ АКАИКЕ

Опишем расширение ИКА на непрерывный случай. Пусть задана обучающая выборка $Z = (\mathbf{z}_1, ..., \mathbf{z}_n)$, $\mathbf{z} \in \mathbb{R}^d$. Требуется восстановить неизвестную плотность распределения $p(\mathbf{x})$ на элементах множества X, где Z и X – выборки длины n из одного вероятностного распределения.

Для описания общей схемы поиска $p(\mathbf{x})$ введем понятие модели.

Определение 1. *Вероятностной моделью алгоритмов восстановления плотностей* назовем тройку $\langle \Omega, p(X|\mathbf{w}), p(\mathbf{w})\rangle$, где $\Omega = \{\mathbf{w}\}$ – значения параметров плотностей распределения, $p(X|\mathbf{w}) = \prod_{i=1}^{n} p(\mathbf{x}_i|\mathbf{w})$ – функция правдоподобия выборки X при фиксированном значении $\mathbf{w}$ и $p(\mathbf{w})$ – априорное распределение на $\mathbf{w}$.

Предположим, что априорное распределение зависит от некоторого параметра A, т.е. может быть записано в виде $p(\mathbf{w}|A)$. Тогда, варьируя A, получаем параметрическое семейство моделей алгоритмов восстановления плотностей $\{\langle \Omega, p(X|\mathbf{w}), p(\mathbf{w}|A)\rangle, A \in \mathcal{A}\}$. В этом случае A называется параметром вероятностной модели.

Определение 2. Назовем *байесовской оценкой параметра* $\mathbf{w}$ значение $\mathbf{w}_{MP}(Z, A)$, максимизирующее величину регуляризованного правдоподобия, т.е.

$$\mathbf{w}_{MP}(Z, A) \stackrel{\Delta}{=} \arg\max_{\mathbf{w}} p(Z|\mathbf{w})p(\mathbf{w}|A).$$

Заметим, что байесовская оценка зависит как от обучающей выборки Z, так и от параметра вероятностной модели A. Пусть

$$\mathbf{w}_n^*(A) \stackrel{\Delta}{=} \mathbb{E}_Z \mathbf{w}_{MP}(Z, A),$$

$$C_n(A) \stackrel{\Delta}{=} \mathbb{E}_Z[\mathbf{w}_{MP}(Z, A) - \mathbf{w}_n^*(A)][\mathbf{w}_{MP}(Z, A) - \mathbf{w}_n^*(A)]^{\mathrm{T}},$$

где математические ожидания берутся по всем выборкам длины n из данного распределения $p(\mathbf{x})$.

Определение 3. *Матрицей Фишера* назовем выражение вида

$$F \stackrel{\Delta}{=} -\int \nabla_{\mathbf{w}} \nabla_{\mathbf{w}} \log p(\mathbf{x}|\mathbf{w}) p(\mathbf{x}) d\mathbf{x}.$$

Пусть $F_n = nF$. Заметим, что $F_n = \mathbb{E}_X \nabla_{\mathbf{w}} \nabla_{\mathbf{w}} \log p(X|\mathbf{w}) = \nabla_{\mathbf{w}} \nabla_{\mathbf{w}} \mathbb{E}_X \log p(X|\mathbf{w})$. В дальнейшем под символом ∇ будем понимать $\nabla_{\mathbf{w}}$.

В случае фиксированного значения параметра A (т.е. фиксированной вероятностной модели) в качестве оценки $p(\mathbf{x})$ будем использовать $p(\mathbf{x}|\mathbf{w}_{MP}(Z, A))$. Параметр модели может быть подобран, следуя идее Акаике, путем максимизации информации Кульбака по A:

$$\mathbb{E}_X \mathbb{E}_Z \log p(X|\mathbf{w}_{MP}(Z, A)) = \iint p(Z)p(X)\log p(X|\mathbf{w}_{MP}(Z, A)) dX dZ \longrightarrow \max_A. \tag{2.1}$$

Итак, задача обучения состоит в нахождении значения параметра вероятностной модели A, оптимального в смысле критерия (2.1). Аналитически вычислить данный интеграл не удается. Сформулируем условия, накладываемые на семейство вероятностных моделей, при выполнении которых данное выражение можно упростить.

Теорема 1. *Пусть A – симметричная неотрицательно-определенная квадратная матрица действительных чисел. Пусть при любом A для вероятностной модели $\Omega(A)$ справедливо следующее*:

1) $p(X|\mathbf{w}) = \prod_{i=1}^{n} p(\mathbf{x}_i|\mathbf{w})$, *т.е. объекты обучающей выборки – это независимые, одинаково распределенные случайные величины;*

2) $\log p(X|\mathbf{w})$ – *квадратичная функция по* $\mathbf{w}$;

3) $p(\mathbf{w}|A) = \mathcal{N}(\mathbf{w}|\mathbf{0}, A^{-1})$, *т.е* $\mathbf{w}$ *распределен нормально с центром в нуле и матрицей ковариации* A^{-1};

4) *случайные величины* $\mathbf{w}_{MP}(X, A)$ *и* $\nabla\nabla\log p(X|\mathbf{w}_{MP}(X, A))$ *независимы.*

Тогда верно соотношение

$$\mathbb{E}_X\mathbb{E}_Z\log p(X|\mathbf{w}_{MP}(Z, A)) = \mathbb{E}_X\log p(X|\mathbf{w}_{MP}(X, A)) - \operatorname{tr}(F_n + A)C_n(A). \tag{2.2}$$

Для упрощения записи будем опускать зависимость $\mathbf{w}_{MP}$, $\mathbf{w}_n^*$ и C_n от A (подразумевая ее). Используя теорему о замене переменных под знаком интеграла Лебега (см. [10, гл. II, п. 6]), можем переписать исходное выражение в виде

$$\mathbb{E}_X\mathbb{E}_Z\log p(X|\mathbf{w}_{MP}(Z)) = \mathbb{E}_{w_{MP}}\mathbb{E}_X\log p(X|\mathbf{w}_{MP}). \tag{2.3}$$

Раскладывая внутреннее математическое ожидание в ряд Тейлора в точке $\mathbf{w}_n^*$, получаем

$$\begin{aligned}\mathbb{E}_{\mathbf{w}_{MP}}\mathbb{E}_X\log p(X|\mathbf{w}_{MP}) &= \mathbb{E}_{\mathbf{w}_{MP}}\mathbb{E}_X\log p(X|\mathbf{w}_n^*) + \mathbb{E}_{\mathbf{w}_{MP}}\mathbb{E}_X[\nabla\log p(X|\mathbf{w}_n^*)]^{\mathrm{T}}(\mathbf{w}_{MP} - \mathbf{w}_n^*) + \\ &+ \frac{1}{2}\mathbb{E}_{\mathbf{w}_{MP}}(\mathbf{w}_{MP} - \mathbf{w}_n^*)^{\mathrm{T}}[\nabla\nabla\mathbb{E}_X\log p(X|\mathbf{w}_n^*)](\mathbf{w}_{MP} - \mathbf{w}_n^*) = \mathbb{E}_X\log p(X|\mathbf{w}_n^*) - \frac{1}{2}\operatorname{tr}F_nC_n.\end{aligned} \tag{2.4}$$

Для оценки первого слагаемого в (2.4) разложим в ряд Тейлора $\log p(X|\mathbf{w}_n^*)$ в точке $\mathbf{w}_{MP}(X)$. Тогда имеем

$$\begin{aligned}\mathbb{E}_X\log p(X|\mathbf{w}_n^*) &= \mathbb{E}_X\log p(X|\mathbf{w}_{MP}(X)) + \mathbb{E}_X\{[\nabla\log p(X|\mathbf{w}_{MP}(X))]^{\mathrm{T}}[\mathbf{w}_n^* - \mathbf{w}_{MP}(X)]\} + \\ &+ \frac{1}{2}\mathbb{E}_X\{[\mathbf{w}_n^* - \mathbf{w}_{MP}(X)]^{\mathrm{T}}\nabla\nabla\log p(X|\mathbf{w}_{MP}(X))[\mathbf{w}_n^* - \mathbf{w}_{MP}(X)]\}.\end{aligned} \tag{2.5}$$

Так как $\log p(X|\mathbf{w})$ квадратична по $\mathbf{w}$, то легко показать, что

$$\nabla\log p(X|\mathbf{w}_{MP}) = A\mathbf{w}_{MP}.$$

Используя данный факт и условие независимости случайных величин $\mathbf{w}_{MP}(X)$ и $\nabla\nabla\log p(X|\mathbf{w}_{MP}(X))$, можно упростить два последних слагаемых в (2.5):

$$\begin{aligned}&\mathbb{E}_X[\nabla\log p(X|\mathbf{w}_{MP}(X))]^{\mathrm{T}}[\mathbf{w}_n^* - \mathbf{w}_{MP}(X)] + \\ &+ \frac{1}{2}\mathbb{E}_X[\mathbf{w}_n^* - \mathbf{w}_{MP}(X)]^{\mathrm{T}}\nabla\nabla\log p(X|\mathbf{w}_{MP}(X))[\mathbf{w}_n^* - \mathbf{w}_{MP}(X)] = \\ &= \mathbb{E}_X\mathbf{w}_{MP}^{\mathrm{T}}(X)A[\mathbf{w}_n^* - \mathbf{w}_{MP}(X)] + \\ &+ \frac{1}{2}\operatorname{tr}\{\mathbb{E}_X\nabla\nabla\log p(X|\mathbf{w}_{MP}(X))\mathbb{E}_X[\mathbf{w}_n^* - \mathbf{w}_{MP}(X)]^{\mathrm{T}}[\mathbf{w}_n^* - \mathbf{w}_{MP}(X)]\} = -\operatorname{tr}AC_n - \frac{1}{2}\operatorname{tr}F_nC_n.\end{aligned}$$

Подставляя полученное выражение в (2.5) и объединяя результат с (2.4), получаем

$$\begin{aligned}&\mathbb{E}_Z\mathbb{E}_X\log p(X|\mathbf{w}_{MP}(Z, A)) = \\ &= \mathbb{E}_X\log p(X|\mathbf{w}_{MP}(X, A)) - \operatorname{tr}AC_n - \frac{1}{2}\operatorname{tr}F_nC_n - \frac{1}{2}\operatorname{tr}F_nC_n = \\ &= \mathbb{E}_X\log p(X|\mathbf{w}_{MP}(X, A)) - \operatorname{tr}(F_n + A)C_n.\end{aligned} \tag{2.6}$$

Теорема доказана.

Следствие 1. При использовании в методах распознавания критерий (2.1) может быть приближенно вычислен по формуле

$$\mathbb{E}_X\mathbb{E}_Z \log p(X|\mathbf{w}_{MP}(Z, A)) \approx \log p(Z|\mathbf{w}_{MP}(Z, A)) - \operatorname{tr}(H(Z) + A)^{-1}H(Z), \tag{2.7}$$

где $H(Z) = \nabla\nabla \log p(Z|\mathbf{w}) = \Sigma_{i=1}^{n} \nabla\nabla \log p(\mathbf{z}_i|\mathbf{w})$ – гессиан логарифма правдоподобия.

Покажем, что $C_n \approx (F_n + A)^{-1}F_n(F_n + A)^{-1}$. Обозначим через $\mathbf{w}_{ML}(X)$ оценку максимального правдоподобия на выборке X. Известно (см. [11]), что $\mathbf{w}_{ML} \sim \mathcal{N}(\mathbf{w}_{ML}|\mathbf{w}_*, F_n^{-1})$, где $\mathbf{w}_* = \operatorname{argmax} \int p(\mathbf{x}) \log p(\mathbf{x}|\mathbf{w}) d\mathbf{x}$. При условии квадратичности $\log p(\mathbf{x}|\mathbf{w})$ по $\mathbf{w}$ легко показать, что

$$\mathbf{w}_{MP}(X, A) = [H(X) + A]^{-1}H(X)\mathbf{w}_{ML}(X), \tag{2.8}$$

где $H(X) = \Sigma_{i=1}^{n} \nabla\nabla \log p(\mathbf{x}_i|\mathbf{w})$.

С учетом $\mathbb{E}\nabla\nabla \log p(\mathbf{x}_i|\mathbf{w}) = F$, используя закон больших чисел (см. [10, гл. III, п. 3]), записываем

$$\forall \varepsilon > 0 \quad P\left(\frac{|H(X) - F_n|}{n} \geq \varepsilon\right) \longrightarrow 0 \quad \text{при} \quad n \longrightarrow \infty. \tag{2.9}$$

Рассмотрим множество $I_n(\varepsilon) = \left\{ X \middle| \frac{|H(X) - F_n|}{n} \geq \varepsilon \right\}$. Из (2.9) следует, что $P(I_n(\varepsilon)) \longrightarrow 0$ при $n \longrightarrow \infty$; при этом на множестве $\mathbb{R}^{n \times d} \setminus I_n(\varepsilon)$ справедливо представление

$$\frac{H(X)}{n} = \frac{F_n}{n} + \delta_1(n), \quad \text{где} \quad \|\delta_1(n)\| \longrightarrow 0 \quad \text{при} \quad n \longrightarrow \infty.$$

При фиксированных $\varepsilon > 0$ и $n > 0$ имеем

$$\mathbb{E}\mathbf{w}_{MP}(X) = \int\limits_{\mathbb{R}^{n \times d} \setminus I_n(\varepsilon)} \mathbf{w}_{MP}(X)p(X)dX + \int\limits_{I_n(\varepsilon)} \mathbf{w}_{MP}(X)p(X)dX. \tag{2.10}$$

Предполагая ограниченность $\mathbf{w}_{MP}(X)$, применяем теорему о среднем ко второму слагаемому в (2.10):

$$\mathbb{E}\mathbf{w}_{MP}(X) = \int\limits_{\mathbb{R}^{n \times d} \setminus I_n(\varepsilon)} [H(X) + A]^{-1}H(X)\mathbf{w}_{ML}(X)p(X)dX + L_{\mathbf{w}_{MP}}P(I_n(\varepsilon)), \tag{2.11}$$

где $L_{\mathbf{w}_{MP}}$ – некоторая положительная константа. Заметим, что при достаточно больших значениях n справедлива оценка $\|\delta_1(n)\| \leq \|F_n/n + A/n\|^{-1}$; тогда верно (см. [12, гл. 5, п. 6]) следующее разложение: $[F_n/n + A/n + \delta_1(n)]^{-1} = (F_n/n + A/n)^{-1} + \delta_2(n)$, где $\|\delta_2(n)\| \longrightarrow 0$ при $n \longrightarrow \infty$. Далее,

$$\int\limits_{\mathbb{R}^{n \times d} \setminus I_n(\varepsilon)} [H(X) + A]^{-1}H(X)\mathbf{w}_{ML}(X)p(X)dX =$$

$$= \int\limits_{\mathbb{R}^{n \times d} \setminus I_n(\varepsilon)} \left[\frac{H(X)}{n} + \frac{A}{n}\right]^{-1} \frac{H(X)}{n}\mathbf{w}_{ML}(X)p(X)dX =$$

$$= \int\limits_{\mathbb{R}^{n \times d} \setminus I_n(\varepsilon)} \left(\frac{F_n}{n} + \frac{A}{n} + \delta_1(n)\right)^{-1}\left(\frac{F_n}{n} + \delta_1(n)\right)\mathbf{w}_{ML}(X)p(X)dX = \tag{2.12}$$

$$= \int\limits_{\mathbb{R}^{n\times d}\setminus I_n(\varepsilon)} \left[\left(\frac{F_n}{n}+\frac{A}{n}\right)^{-1}+\delta_2(n)\right]\left(\frac{F_n}{n}+\delta_1(n)\right)\mathbf{w}_{ML}(X)p(X)dX =$$

$$= \int\limits_{\mathbb{R}^{n\times d}\setminus I_n(\varepsilon)} (F_n+A)^{-1}F_n\mathbf{w}_{ML}(X)p(X)dX + \int\limits_{\mathbb{R}^{n\times d}\setminus I_n(\varepsilon)} \delta_3(n)p(X)dX,$$

где $\|\delta_3(n)\| \longrightarrow 0$ при $n \longrightarrow \infty$. Итак, получаем

$$\mathbb{E}\mathbf{w}_{MP}(X) = (F_n+A)^{-1}F_n \int\limits_{\mathbb{R}^{n\times d}\setminus I_n(\varepsilon)} \mathbf{w}_{ML}(X)p(X)dX + \int\limits_{\mathbb{R}^{n\times d}\setminus I_n(\varepsilon)} \delta_3(n)p(X)dX + L_{\mathbf{w}_{MP}}P(I_n(\varepsilon)). \tag{2.13}$$

В силу того что $L_{\mathbf{w}_{MP}}P(I_n(\varepsilon))$ – положительная константа, $P(I_n(\varepsilon)) \longrightarrow 0$ и $\|\delta_3(n)\| \longrightarrow 0$ при $n \longrightarrow \infty$, при увеличении объема выборки множество $\mathbb{R}^{n\times d}\setminus I_n(\varepsilon)$ будет стремиться к множеству $\mathbb{R}^{n\times d}$, а два последних слагаемых в (2.13) – к нулю. Отбрасывая два последних слагаемых, получаем следующее приближение к $\mathbb{E}\mathbf{w}_{MP}(X)$:

$$\mathbb{E}\mathbf{w}_{MP}(X) \approx (F_n+A)^{-1}F_n\mathbb{E}\mathbf{w}_{ML}(X).$$

Аналогично можем провести приближение и для $\mathbb{E}\mathbf{w}_{MP}(X)\mathbf{w}_{MP}^{\mathrm{T}}(X)$:

$$\mathbb{E}\mathbf{w}_{MP}(X)\mathbf{w}_{MP}^{\mathrm{T}}(X) \approx (F_n+A)^{-1}F_n\mathbb{E}\mathbf{w}_{ML}(X)\mathbf{w}_{ML}^{\mathrm{T}}(X)F_n(F_n+A)^{-1}.$$

Учитывая, что $F_n^{-1} = [\mathbb{E}\mathbf{w}_{ML}(X)\mathbf{w}_{ML}^{\mathrm{T}}(X) - (\mathbb{E}\mathbf{w}_{ML}(X))(\mathbb{E}\mathbf{w}_{ML}(X))^{\mathrm{T}}]$, получаем

$$\begin{gathered} C_n = \mathbb{E}\mathbf{w}_{MP}(X)\mathbf{w}_{MP}^{\mathrm{T}}(X) - (\mathbb{E}\mathbf{w}_{MP}(X))(\mathbb{E}\mathbf{w}_{MP}(X))^{\mathrm{T}} \approx \\ \approx (F_n+A)^{-1}F_n[\mathbb{E}\mathbf{w}_{ML}(X)\mathbf{w}_{ML}^{\mathrm{T}}(X) - (\mathbb{E}\mathbf{w}_{ML}(X))(\mathbb{E}\mathbf{w}_{ML}(X))^{\mathrm{T}}]F_n(F_n+A)^{-1} = \\ = (F_n+A)^{-1}F_nF_n^{-1}F_n(F_n+A)^{-1} = (F_n+A)^{-1}F_n(F_n+A)^{-1}. \end{gathered} \tag{2.14}$$

Подставив приближенное значение C_n во второе слагаемое (2.2), получим

$$\mathbb{E}_X\mathbb{E}_Z\log p(X|\mathbf{w}_{MP}(Z,A)) = \mathbb{E}_X\log p(X|\mathbf{w}_{MP}(X,A)) - \mathrm{tr}(F_n+A)^{-1}F_n.$$

Подставляя в полученное выражение вместо X обучающую выборку Z и вместо матрицы F_n ее несмещенную оценку $H(Z)$, получаем утверждение следствия.

Заметим, что если A – диагональная матрица с элементами, равными либо 0, либо $+\infty$, то $\mathrm{tr}(F_n(F_n+A)^{-1}) = k$ – количество нулевых диагональных элементов A; при этом (2.7) становится с точностью до бесконечно малой константы эквивалентно классическому информационному критерию Акаике. Подобное непрерывное расширение критерия (ОИКА) может также быть рассмотрено, как частный случай девиантного информационного критерия, описанного в [13].

3. ПРИМЕНЕНИЕ ОИКА К ЗАДАЧЕ ОБОБЩЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

3.1. Постановка задачи оптимизации

Рассмотрим классическую задачу обобщенной линейной регрессии. Пусть $(X, \mathbf{t}) = \{(\mathbf{x}_1, t_1), \ldots$ $\ldots (\mathbf{x}_n, t_n)\}$ – обучающая выборка, где $\mathbf{x}_i = (x_i^1, \ldots, x_i^d) \in \mathbb{R}^d$ – вектор наблюдаемых признаков объекта, а $t_i \in \mathbb{R}$ – значение зависимой переменной. Зафиксируем некоторое множество базисных функций $\{\phi_1(\mathbf{x}), \ldots, \phi_m(\mathbf{x})\}$, $\phi_j : \mathbb{R}^d \longrightarrow \mathbb{R}^d$. Требуется найти вектор весов $\mathbf{w} \in \mathbb{R}$ такой, чтобы функция

$$y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}) = \sum_{j=1}^{m} w_j\phi_j(\mathbf{x})$$

приближала значения переменной t в объектах обучающей выборки X. Пусть $\Phi = (\phi_{ij})_{n\times m} = (\phi_j(\mathbf{x}_i))_{n\times m}$ – матрица базисных функций, вычисленных для каждого объекта обучающей выборки. Классический подход к обучению линейной регрессии состоит в оптимизации регуляризованного правдоподобия

$$\mathbf{w}_{MP} = \arg\max_{\mathbf{w}} p(\mathbf{t}|X, \mathbf{w})p(\mathbf{w}|\alpha), \tag{3.1}$$

где

$$p(\mathbf{t}|X, \mathbf{w}) = \frac{1}{\sqrt{(2\pi)^n}\sigma^n}\exp\left(-\frac{1}{2\sigma^2}\|\Phi\mathbf{w} - \mathbf{t}\|^2\right) \tag{3.2}$$

есть функция правдоподобия,

$$p(\mathbf{w}|\alpha) = \sqrt{\left(\frac{\alpha}{2\pi}\right)^m}\exp\left(-\frac{\alpha}{2}\|\mathbf{w}\|^2\right)$$

есть априорное распределение на веса. Априорное распределение имеет смысл регуляризатора, штрафующего большие значения $\mathbf{w}$. Более общий случай семейства регуляризаторов рассмотрен в методе релевантных векторов (MPB, см. [1]), где для каждого веса w_j вводится свой коэффициент регуляризации, а априорное распределение имеет вид

$$p(\mathbf{w}|\alpha) = \prod_{j=1}^{m}\sqrt{\frac{\alpha_j}{2\pi}}\exp\left(-\frac{\alpha_j}{2}w_j^2\right) = \frac{\sqrt{\det(A)}}{(2\pi)^{m/2}}\exp\left(-\frac{1}{2}\mathbf{w}^{\mathrm{T}}A\mathbf{w}\right), \tag{3.3}$$

где $A = \mathrm{diag}(\alpha_1, \ldots, \alpha_m)$ – матрица регуляризации, $\alpha_j \geq 0$. Такое априорное распределение позволяет проводить выбор базисных функций. В случае если $\alpha_j = 0$, то никаких дополнительных ограничений на значение веса $w_{MP,j}$ не накладывается и его значение совпадает с точкой максимума правдоподобия $w_{ML,j}$. Если параметр регуляризации α_j стремится к плюс бесконечности, то соответствующая базисная функция $\phi_j(\cdot)$ исключается из модели, так как ее вес $w_{MP,j} = 0$. Таким образом, априорное распределение (3.3) вместе с функцией правдоподобия (3.2) и методом байесовского оценивания (3.1) позволяет решать задачу селекции базисных функций. Данная задача переходит в задачу отбора признаков, если в качестве базисных функций выбираются исходные признаки $\phi_j(\mathbf{x}) = x^j$. Если в качестве базисных функций выбираются ядровые или потенциальные функции с центром в объектах обучающей выборки $\phi_j(\mathbf{x}) = K(\mathbf{x}_j, \mathbf{x})$, то данный подход позволяет отбирать релевантные объекты.

Заметим, что задачу восстановления регрессии в статистической постановке можно рассматривать как частный случай задачи восстановления плотностей. Используя введенные функции, сформулируем процедуру обучения (подбора $\mathbf{w}$ и α) в терминах вероятностных моделей алгоритмов восстановления плотностей. Параметрическое семейство вероятностных моделей может быть записано в виде

$$\{\langle \mathbb{R}^m, P(\mathbf{t}|X, \mathbf{w}), p(\mathbf{w}|\alpha)\rangle, \alpha \in \mathbb{R}^m\}. \tag{3.4}$$

При фиксированной вероятностной модели α в качестве решения задачи выбираем $\mathbf{w}_{MP}(X, \alpha)$ – байесовскую оценку вектора весов $\mathbf{w}$. Рассмотрим далее способ подбора α.

Условия теоремы 1 для семейства вероятностных моделей (3.4) выполнены. Поэтому для выбора наилучшей модели α воспользуемся следствием теоремы 1

$$\begin{aligned}\alpha = \arg\max f(\alpha) &= \arg\max\{\log p(\mathbf{t}|X, \mathbf{w}_{MP}) - \mathrm{tr}[H(H + A)^{-1}]\} = \\ &= \arg\max\{\mathscr{L}(\mathbf{w}_{MP}) - \mathrm{tr}[H(H + A)^{-1}]\}.\end{aligned} \tag{3.5}$$

Здесь $H = -\nabla\nabla\log p(\mathbf{t}|X, \mathbf{w}) = \sigma^{-2}\Phi^{\mathrm{T}}\Phi$.

3.2. Процедура оптимизации

Поиск решения задачи оптимизации (3.5) будем проводить с помощью покоординатного спуска – отдельно по каждой компоненте α_j. Для вывода итеративных уравнений пересчета α_j воспользуемся тождеством блочного матричного обращения:

$$\begin{pmatrix} P & Q \\ R & S \end{pmatrix}^{-1} = \begin{pmatrix} P^{-1} + P^{-1}QBRP^{-1} & -P^{-1}QB \\ -BRP^{-1} & B \end{pmatrix}. \tag{3.6}$$

Здесь $P \in \mathbb{R}^{p\times p}$, $Q \in \mathbb{R}^{p\times q}$, $R \in \mathbb{R}^{q\times p}$, $S \in \mathbb{R}^{q\times q}$ – некоторые матрицы, а $B = (S - RP^{-1}Q)^{-1}$ – дополнение Шура.

Далее применим это тождество к матрице $(H + A)$, представленной в следующем виде:

$$(H + A) = \begin{pmatrix} P & \mathbf{q} \\ \mathbf{q}^{\mathrm{T}} & h_{mm} + \alpha_m \end{pmatrix}.$$

Для простоты изложения, не ограничивая общности, будем выводить итеративные уравнения для α_m. Пересчет остальных компонент вектора $\boldsymbol{\alpha}$ производится аналогичным образом. Используя (3.6), получаем

$$(H + A)^{-1} = \begin{pmatrix} P^{-1} + \beta_m P^{-1}\mathbf{q}\mathbf{q}^{\mathrm{T}}P^{-1} & -\beta_m P^{-1}\mathbf{q} \\ -\beta_m \mathbf{q}^{\mathrm{T}}P^{-1} & \beta_m \end{pmatrix},$$

где $\beta_m = (h_{mm} + \alpha_m - \mathbf{q}^{\mathrm{T}}P^{-1}\mathbf{q})^{-1}$ – скалярное дополнение Шура. Тогда $\mathrm{tr}[H(H + A)^{-1}]$ может быть выражено как функция от α_m, при условии, что остальные α_j фиксированы:

$$\mathrm{tr}[H(H + A)^{-1}] = \mathrm{tr}(H_{(m)}P^{-1}) + \beta_m[\mathbf{q}^{\mathrm{T}}P^{-1}(H_{(m)}P^{-1}\mathbf{q} - 2\mathbf{q}^{\mathrm{T}}P^{-1}\mathbf{q} + h_{mm}].$$

Здесь нижний индекс $_{(m)}$ означает вектор (или матрицу), у которого удалена m-я строка (и столбец).

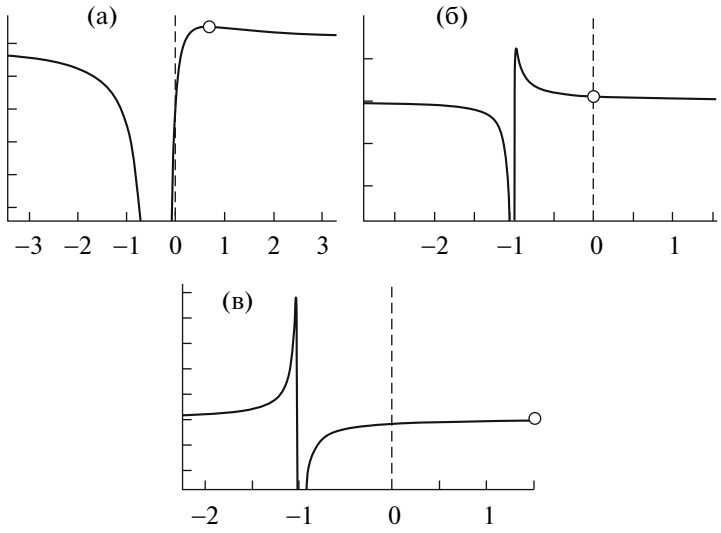
Заметим, что значение $\mathrm{tr}(H_{(m)}P^{-1})$ в точности равно значению $\mathrm{tr}[H(H + A)^{-1}]$ при $\alpha_m = +\infty$, т.е. при удалении из модели m-й базисной функции.

Рассмотрим разницу между точкой максимума апостериорного распределения $\mathbf{w}_{MP}$ и точкой максимума апостериорного распределения при бесконечно большом значении коэффициента регуляризации для m-й базисной функции (при этом значения остальных компонент вектора $\boldsymbol{\alpha}$ остаются неизменными) $\mathbf{w}^*_{MP} = \mathbf{w}_{MP}\big|_{\alpha_m = +\infty} \in \mathbb{R}^m$. Пусть $\boldsymbol{\psi} = H\mathbf{w}_{ML}$. Используя соотношение $\mathbf{w}_{MP} = (H + A)^{-1}H\mathbf{w}_{ML}$, получаем

$$\begin{aligned} \mathbf{w}_{MP} - \mathbf{w}^*_{MP} &= [(H + A)^{-1} - (H + A)^{-1}\big|_{\alpha_m = +\infty}]\boldsymbol{\psi} = \\ &= \left[\begin{pmatrix} P^{-1} + \beta_m P^{-1}\mathbf{q}\mathbf{q}^{\mathrm{T}}P^{-1} & -\beta_m P^{-1}\mathbf{q} \\ -\beta_m \mathbf{q}^{\mathrm{T}}P^{-1} & \beta_m \end{pmatrix} - \begin{pmatrix} P^{-1} & \mathbf{0} \\ \mathbf{0}^{\mathrm{T}} & 0 \end{pmatrix}\right]\begin{pmatrix} \boldsymbol{\psi}_{(m)} \\ \psi_m \end{pmatrix} = \\ &= \beta_m \begin{pmatrix} P^{-1}\mathbf{q}\mathbf{q}^{\mathrm{T}}P\boldsymbol{\psi}_{(m)} - \psi_m P^{-1}\mathbf{q} \\ -\mathbf{q}^{\mathrm{T}}P^{-1}\boldsymbol{\psi}_{(m)} + \psi_m \end{pmatrix} = \beta_m \boldsymbol{\xi}_m. \end{aligned} \tag{3.7}$$

Рассмотрим разность между значениями логарифма правдоподобия в точках $\mathbf{w}_{MP}$ и $\mathbf{w}^*_{MP}$:

$$\mathscr{L}(\mathbf{w}_{MP}) - \mathscr{L}(\mathbf{w}^*_{MP}) = \beta_m \nabla\mathscr{L}(\mathbf{w}^*_{MP})^{\mathrm{T}}\boldsymbol{\xi}_m - \frac{\beta_m^2}{2}\boldsymbol{\xi}_m^{\mathrm{T}}H\boldsymbol{\xi}_m = \beta_m \boldsymbol{\zeta}_m^{\mathrm{T}}\boldsymbol{\xi}_m - \frac{\beta_m^2}{2}\boldsymbol{\xi}_m^{\mathrm{T}}H\boldsymbol{\xi}_m.$$



Фиг. 1.

Используя (2.8), записываем градиент в виде

$$\zeta_m = \nabla\mathscr{L}(\mathbf{w}^*_{MP}) = -H(\mathbf{w}^*_{MP} - \mathbf{w}_{ML}) =$$

$$= -[H(H+A)^{-1}\big|_{\alpha_m=+\infty} - I]\psi = -\begin{pmatrix} (H_{(m)}P^{-1} - I)\psi_{(m)} \\ \mathbf{q}^{\mathrm{T}}P^{-1}\psi_{(m)} - \psi_m \end{pmatrix}. \tag{3.8}$$

В результате значение критерия ОИКА f как функции от β_m при фиксированных α_j, $j \neq m$ представляется в следующем виде:

$$f(\beta_m) = f(0) - \frac{1}{2}\beta_m^2\xi_m^{\mathrm{T}}H\xi_m + \beta_m\zeta_m^{\mathrm{T}}\xi_m - \beta_m\mathbf{q}^{\mathrm{T}}P^{-1}H_{(m)}P^{-1}\mathbf{q} + 2\beta_m\mathbf{q}^{\mathrm{T}}P^{-1}\mathbf{q} - \beta_m h_{mm}. \tag{3.9}$$

Критерий квадратичен по β_m и, следовательно, имеет единственный максимум, который вычисляется аналитически по формуле

$$\beta^*_m = \frac{\zeta_m^{\mathrm{T}}\xi_m - \mathbf{q}^{\mathrm{T}}P^{-1}H_{(m)}P^{-1}\mathbf{q} + 2\mathbf{q}^{\mathrm{T}}P^{-1}\mathbf{q} - h_{mm}}{\xi_m^{\mathrm{T}}H\xi_m} = \frac{b}{a}.$$

Используя выражения для ξ_m (3.7) и ζ_m (3.8), значения a и b вычисляем следующим образом:

$$a = (\mathbf{q}^{\mathrm{T}}P^{-1}\psi_{(m)} - \psi_m)^2(\mathbf{q}^{\mathrm{T}}P^{-1}H_{(m)}P^{-1}\mathbf{q} - 2\mathbf{q}^{\mathrm{T}}P^{-1}\mathbf{q} + h_{mm}), \tag{3.10}$$

$$\begin{aligned} b = &-(\psi_{(m)}^{\mathrm{T}}P^{-1}H_{(m)}P^{-1}\mathbf{q} - 2\psi_{(m)}^{\mathrm{T}}P^{-1}\mathbf{q} + \psi_m)(\mathbf{q}^{\mathrm{T}}P^{-1}\psi_{(m)} - \psi_m) - \\ &- \mathbf{q}^{\mathrm{T}}P^{-1}H_{(m)}P^{-1}\mathbf{q} + 2\mathbf{q}^{\mathrm{T}}P^{-1}\mathbf{q} - h_{mm}. \end{aligned} \tag{3.11}$$

Перейдем от вспомогательной переменной β_m к исходной α_m:

$$\beta_m = (h_{mm} + \alpha_m - \mathbf{q}^{\mathrm{T}}P^{-1}\mathbf{q})^{-1};$$

следовательно,

$$\alpha^*_m = \mathbf{q}^{\mathrm{T}}P^{-1}\mathbf{q} - h_{mm} + \frac{1}{\beta^*_m}.$$

При использовании последнего выражения необходимо учитывать также, что $\alpha_m \geq 0$. Зависимость ОИКА от α_m имеет характерную форму ириса, показанную на фиг. 1. В случае (а) максимум

достигается для положительного α_j. В случае (б) критерий монотонно убывает в области $\alpha_j \geq 0$ и, следовательно, оптимальное значение $\alpha_j = 0$. В случае (в) оптимальное неотрицательное значение α_j равно $+\infty$. Критерий равен минус бесконечности, когда матрица $H + A$ вырождена, т.е. $\alpha_m = \mathbf{q}^{\mathrm{T}} P^{-1}\mathbf{q} - h_{mm}$. Согласно свойству дополнения Шура, $\mathbf{q}^{\mathrm{T}} P^{-1}\mathbf{q} - h_{mm} \leq 0$, т.е. "стебель" всегда соответствует неположительным α_m. В зависимости от взаиморасположения точки максимума и "стебля" значение α_m пересчитывается разными способами:

$$\alpha_m^{(\mathrm{new})} = \begin{cases} \alpha_m^*, & \alpha_m^* \geq 0, \\ 0, & \mathbf{q}^{\mathrm{T}} P^{-1}\mathbf{q} - h_{mm} < \alpha_m^* < 0, \\ +\infty, & \alpha_m^* < \mathbf{q}^{\mathrm{T}} P^{-1}\mathbf{q} - h_{mm}. \end{cases} \tag{3.12}$$

Выражения для α_j при $j \neq m$ аналогичны.

Для подбора значения параметра σ^2 продифференцируем ОИКА по σ^{-2}. Приравнивая производную к нулю, получаем следующую формулу пересчета:

$$\sigma^{2(\mathrm{new})} = \frac{\left\|\mathbf{t} - \Phi\mathbf{w}_{MP}\right\|^2}{n - 2\mathrm{tr}(A(H+A)^{-1}H(H+A)^{-1})}. \tag{3.13}$$

Итак, доказана

Теорема 2. *Справедливы соотношения*

$$\arg\max_{\alpha_j \geq 0} f(\alpha_j) = \alpha_j^{(\mathrm{new})}, \quad j = 1, 2, \ldots, m,$$

$$\arg\max_{\sigma^2 \geq 0} f(\sigma^2) = \sigma^{2(\mathrm{new})},$$

где $\alpha_j^{(\mathrm{new})}$ *и* $\sigma^{2(\mathrm{new})}$ *рассчитываются по формулам* (3.12) *и* (3.13) *соответственно. Формула* (3.12) *подразумевает, что оптимизация критерия производится поочередно по каждой из* $\alpha_j, j = 1, 2, \ldots, m$, *при фиксированных остальных компонентах* α.

Используя полученный результат, можно построить итерационный процесс оптимизации критерия. На каждом шаге оптимизируется тот параметр α_j, который обеспечивает максимальный прирост критерия (см. Алгоритм 1). Такой алгоритм схож с методом обучения MPB, предложенным в [14].

Алгоритм 1 ОИКА MPB

вход Обучающая выборка $(X, \mathbf{t}) = \{\mathbf{x}_i, t_i\}_{i=1}^n$, $\mathbf{x}_i \in \mathbb{R}^d$, $t_i \in \mathbb{R}$, множество базисных функций $\{\phi_j(\mathbf{x})\}_{j=1}^m$.

Инициализировать $\alpha_j = +\inf\ \forall j = 1, 2, \ldots, m$, $\sigma^2 = \sigma_0^2$, $\Phi = \{\phi_j(\mathbf{x})_i\}_{i,j=1}^{n,m}$ и $A = \mathrm{diag}(\alpha_1, \ldots, \alpha_m)$.

Найти максимум логарифма правдоподобия $\mathbf{w}_{ML} = (\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}}\mathbf{t}$.

повторять

Вычислить $H = \sigma^{-2}\Phi^{\mathrm{T}}\Phi$ и $\Psi = H\mathbf{w}_{ML}$.

для $j = 1, 2, \ldots, m$ **цикл**

Вычислить $H_{(j)}$, $P^{-1} = (H_{(j)} + A_{(j)})^{-1}$ и $\mathbf{q}$, т.е. j-й столбец H без j-го элемента.

Вычислить a и b, используя выражения (3.10) и (3.11).

Вычислить оптимальное значение $\alpha_j^* = \mathbf{q}^{\mathrm{T}} P^{-1}\mathbf{q} - h_{jj} + a/b$ и текущее приращение ОИКА $\Delta_j = b^2/a$.

если $\alpha_j^* < 0$ **тогда**

если $\alpha_j^* > \mathbf{q}^{\mathrm{T}} P^{-1}\mathbf{q} - h_{jj}$ **тогда**

$\alpha_j^* = 0, \beta_0 = 1/(h_{jj} - \mathbf{q}^{\mathrm{T}} P^{-1}\mathbf{q}), \Delta_j = -\beta_0^2/(2a) + \beta_0 b.$

иначе

$\alpha_j^* = +\infty, \Delta_j = 0.$

конец если

конец если

если $\alpha_j \neq +\infty$ **тогда**

$\beta_{\mathrm{old}} = 1/(h_{jj} - \mathbf{q}^{\mathrm{T}} P^{-1}\mathbf{q} + \alpha_j)$

$\Delta_j^{\mathrm{old}} = -\beta_{\mathrm{old}}^{\mathrm{old}}/(2a) + \beta_{\mathrm{old}} b, \Delta_j = \Delta_j - \Delta_j^{\mathrm{old}}.$

конец если

конец если

Найти $j^* = \arg\max_j \Delta_j$ и установить $\alpha_{j^*} = \alpha_{j^*}^*$.

Вычислить $A = \mathrm{diag}(\alpha_1, .., \alpha_m)$, $\mathbf{w}_{MP} = (H + A)^{-1} H \mathbf{w}_{ML}$ и пересчитать σ^2, используя (3.13).

пока процесс не сошелся

выход Решающее правило для нового объекта $\mathbf{x}$: $f(\mathbf{x}) = \sum_{j=1}^{m} w_{MP,j}\phi_j(\mathbf{x})$

3.3. Недиагональная регуляризация

Альтернативный подход для определения коэффициентов регуляризации α предложен в методе релевантных векторов (MPB, см. [15]). В этом методе используется байесовская парадигма для оценивания параметров и коэффициенты регуляризации находятся с помощью максимизации правдоподобия модели (обоснованности):

$$EV(\alpha) = \int p(\mathbf{t}|X, \mathbf{w}) p(\mathbf{w}|\alpha) d\mathbf{w} \longrightarrow \max_{\alpha}. \tag{3.14}$$

Здесь функция правдоподобия и априорное распределение выбираются, как и раньше, по формулам (3.2) и (3.3). Обоснованность максимизируется с помощью итерационной процедуры, в которой на каждом шаге подынтегральная функция аппроксимируется нормальным распределением.

Другой способ оптимизации обоснованности предложен в рамках подхода недиагональной регуляризации (см. [16]). В этом случае предполагается, что матрица регуляризации A является диагональной в базисе из собственных векторов гессиана логарифма правдоподобия. Тогда можно перейти к новым переменным $\mathbf{u}$, являющимся линейными комбинациями $\mathbf{w}$, таким, что в новом базисе матрицы H, A и $H + A$ станут диагональными. Этот переход существенно упрощает процесс оптимизации обоснованности, так как многомерный интеграл (3.14) переходит в произведение одномерных интегралов, каждый из которых зависит от своего параметра регуляризации α_j. В результате оптимизационный процесс сходится за одну итерацию, так как оптимальные значения коэффициентов регуляризации не зависят друг от друга. Более того, в отличие от MPB и ридж-регрессии, данная процедура инвариантна относительно линейных преобразований базисных функций (т.е. при любом невырожденном линейном преобразовании $\phi(\mathbf{x})$ результат обучения регрессии остается неизменным).

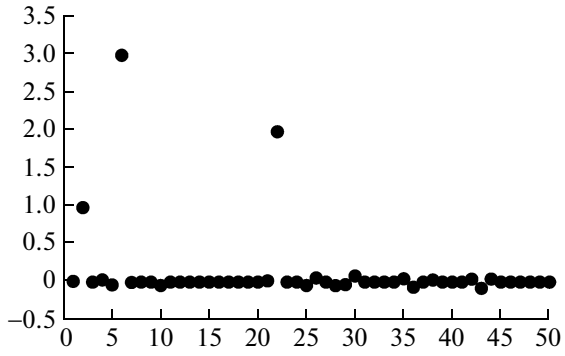
Экспериментальное сравнение классического метода релевантных векторов и предлагаемого в работе метода на основе обобщения информационного критерия Акаике приведено ниже. Показано, что в случае недиагональной регуляризации оба подхода оказываются эквивалентными.

Предполагая, что матрица $H + A$ диагональна и, следовательно, $\mathbf{q} = \mathbf{0}$, получаем

$$\beta_m = \frac{1}{h_{mm} + \alpha_m},$$

$$\xi_m = \begin{pmatrix} \mathbf{0} \\ h_{mm} u_{ML,m} \end{pmatrix},$$

$$\zeta_m^{\mathrm{T}} \xi_m = (h_{mm} u_{ML,m})^2.$$



Фиг. 2.

Оптимальное значение β определяется следующим выражением:

$$\beta_m^* = \frac{(h_{mm}u_{ML,m})^2 - h_{mm}}{h_{mm}^3 u_{ML,m}^2}.$$

Отсюда имеем

$$\alpha_m = \begin{cases} \dfrac{h_{mm}}{h_{mm}u_{ML,m}^2 - 1}, & h_{mm}u_{ML,m}^2 > 1, \\ +\infty, & 0 < h_{mm}u_{ML,m}^2 < 1. \end{cases} \tag{3.15}$$

Случай, соответствующий $\alpha_j = 0$ (см. фиг. 1б), невозможен, так как $h_{mm}u_{ML,m}^2$ всегда неотрицательно. Таким образом, получены те же выражения для α_j, как и в случае оптимизации обоснованности при настройке коэффициентов регуляризации, связанных с собственными векторами гессиана (см. [16]). Следовательно, недиагональная регуляризация с помощью критерия Акаике эквивалентна недиагональной байесовской регуляризации для задачи восстановления регрессии.

4. ЭКСПЕРИМЕНТЫ И ОБСУЖДЕНИЕ

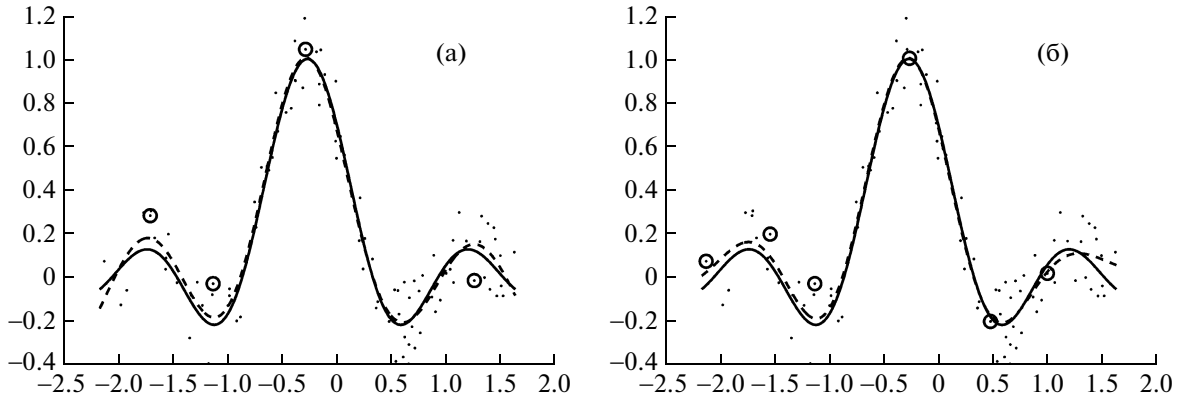
4.1. Отбор признаков

Информационный критерий Акаике широко используется для отбора регрессоров в классической линейной регрессии. Тем не менее эта задача является чрезвычайно трудоемкой из-за необходимости перебирать всевозможные подмножества регрессоров. Предлагаемый метод (ОИКА) значительно облегчает этот процесс, так как при этом задача дискретной оптимизации сводится к задаче гладкой оптимизации, для которой удается построить эффективную итерационную процедуру.

Рассмотрим модельную задачу регрессии с 49 признаками, имеющими стандартные гауссовские распределения, 100 объектами и значением целевой переменной

$$t = x_2 + 3x_6 + 2x_{22} + \varepsilon,$$

где $\varepsilon \sim \mathcal{N}(\varepsilon|0, 0.5)$. Запустив ОИКА с $\phi_j(\mathbf{x}) = x^j$, получим 15 релевантных признаков, из которых 12 имеют коэффициенты регуляризации $\alpha_j > 100$, среди остальных $\alpha_2 = 0.93$, $\alpha_6 = 0.10$ и $\alpha_{22} = 0$. Соответствующие веса показаны на фиг. 2. Легко видеть, что только три веса значительно отличаются от нуля и близки к истинным значениям.



Фиг. 3.

4.2. Функция Sinc

ОИКА обладает свойством разреженности и может рассматриваться как альтернатива МРВ, где коэффициенты регуляризации подбираются с использованием принципа максимальной обоснованности:

$$\alpha = \arg\max \int p(\mathbf{t}|X, \mathbf{w}) p(\mathbf{w}|\alpha) d\mathbf{w}.$$

Различие между двумя методами можно проследить на модельной задаче – зашумленной функции $\frac{\sin(x)}{x}$ с равномерным шумом на отрезке [–0.2, 0.2] и 100 объектами в выборке. На фиг. 3 представлены регрессии, полученные методами ОИКА (график а) и МРВ (график б). Базисные функции $\phi_j(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{x}_j - \mathbf{x}\|^2}{2\sigma^2}\right), j = 1, 2, ..., n$, где $\mathbf{x}_j$ – объекты обучающей выборки; параметр ширины гауссианы $\sigma = 0.4$. Истинная зависимость показана сплошной линией, пунктирная линия соответствует прогнозируемым значениям, релевантные объекты – кружочки. МРВ и ОИКА выделяют 6 и 4 релевантных объектов соответственно. Из графика видно, что регрессия, полученная с помощью МРВ, в целом больше прижимается к нулю, особенно на концах отрезка. Это можно объяснить тем фактом, что все коэффициенты регуляризации отличны от нуля, поэтому даже релевантные базисные функции подвергаются небольшой регуляризации. С другой стороны, решение ОИКА получается более разреженным. При этом двум из четырех базисных функций соответствуют строго нулевые коэффициенты регуляризации. Таким образом, по сравнению с ОИКА, в МРВ наблюдается эффект “недообучения” (переупрощения модели), который часто отмечается для данного метода (см. [17]).

4.3. Сравнительная оценка

Было проведено сравнение методов МРВ, ОИКА и линейной ридж-регрессии (ЛР) на 11 задачах, взятых из хранилища UCI[2)] и Regression Toolbox by Heikki Hyotyniemi[3)].

Во всех методах были установлены следующие параметры. Количество базисных функций $m = n + 1$, $\phi_j(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}_j\|^2}{2\sigma}\right)$ и $\phi_{n+1}(\mathbf{x}) = 1$. Параметр ширины σ подбирался с использованием

[2)] http://archive.ics.uci.edu/ml/
[3)] http://www.control.hut.fi/Hyotyniemi/publications/01_ report125/RegrToolbox

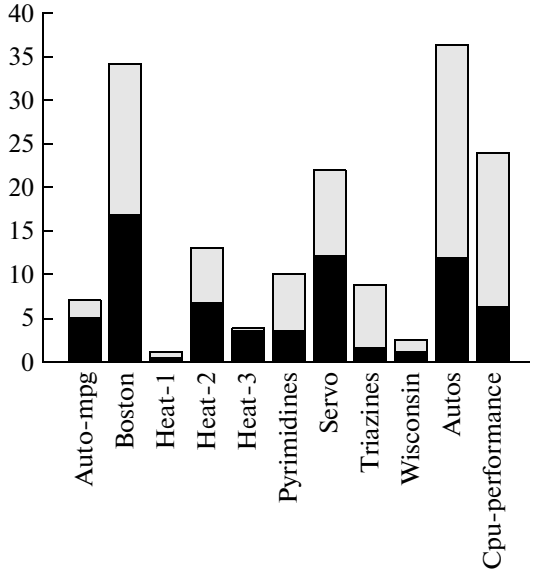
Таблица 1. Корень из среднего квадрата отклонения для различных алгоритмов

Задача	ОИКА	MPB	ЛР
Auto-mpg	2.95 ± 0.06	*2.93 ± 0.04*	**2.92 ± 0.03**
Boston	*3.78 ± 0.22*	**3.75 ± 0.19**	3.86 ± 0.11
HeatExchange-1	*7.90 ± 0.09*	**7.88 ± 0.10**	9.16 ± 0.68
HeatExchange-2	*8.75 ± 0.97*	9.27 ± 1.07	**8.35 ± 1.17**
HeatExchange-3	**0.80 ± 0.07**	*0.82 ± 0.04*	0.84 ± 0.05
Pyrimidines	**0.10 ± 0.01**	**0.10 ± 0.01**	0.11 ± 0.01
Servo	*0.91 ± 0.07*	0.95 ± 0.06	**0.90 ± 0.02**
Triazines	**0.16 ± 0.01**	*0.17 ± 0.00*	*0.17 ± 0.01*
Wisconsin (wdbc)	*25.80 ± 2.31*	**25.27 ± 1.60**	29.18 ± 4.52
Autos	**0.33 ± 0.06**	**0.33 ± 0.02**	0.46 ± 0.04
Cpu-performance	**0.36 ± 0.04**	*0.40 ± 0.04*	0.48 ± 0.20
Ранг	19.00	20.50	26.50
Шрифтовая легенда	**Место1**	*Место2*	Место3

Таблица 2. Разреженность различных алгоритмов (число релевантных объектов)

Задача	ОИКА	MPB	ЛР
Auto-mpg	7.10 ± 4.94	8.60 ± 2.99	199.00 ± 0.00
Boston	34.00 ± 5.24	24.50 ± 3.02	253.00 ± 0.00
HeatExchange-1	1.20 ± 0.45	5.60 ± 5.47	45.00 ± 0.00
HeatExchange-2	13.10 ± 10.17	10.10 ± 4.39	45.00 ± 0.00
HeatExchange-3	3.90 ± 2.86	2.60 ± 0.22	45.00 ± 0.00
Pyrimidines	10.10 ± 2.97	9.80 ± 5.53	37.00 ± 0.00
Servo	22.00 ± 11.80	15.80 ± 3.47	83.50 ± 0.00
Triazines	8.90 ± 5.48	31.30 ± 19.84	93.00 ± 0.00
Wisconsin (wdbc)	2.60 ± 1.39	8.30 ± 4.96	23.50 ± 0.00
Autos	36.30 ± 6.23	23.80 ± 3.27	100.50 ± 0.00
Cpu-performance	23.90 ± 0.89	26.70 ± 3.03	104.50 ± 0.00

кросс-валидации на основе пятикратного разбиения обучающей выборки на два подмножества (5x2-fold cross-validation) (см. [18]). Для каждой обучающей выборки СКО также оценивалось с использованием 5x2-fold кросс-валидации.



Фиг. 4.

Для ридж-регрессии во всех задачах коэффициенты регуляризации были установлены равными 10^{-6}. Для ОИКА и MPB дополнительно вычислялась разреженность (количество ненулевых весов). В табл. 1 и 2 отражены результаты экспериментов. На фиг. 4 проиллюстрировано число релевантных объектов для ОИКА в различных задачах. Черная часть столбца соответствует числу релевантных объектов с нулевыми коэффициентами регуляризации.

ЗАКЛЮЧЕНИЕ

Заметим, что, как и в MPB, большинство α_j в ОИКА стремятся к бесконечности, обеспечивая, таким образом, разреженность получаемого решения. Более того, во многих случаях методы дают близкие результаты. Основным выводом данной работы является тот факт, что информационный критерий Акаике может быть использован для проведения процедуры автоматического определения релевантности наравне с байесовскими методами.

В отличие от MPB, в случае ОИКА некоторые коэффициенты регуляризации становятся тождественно равными нулю. Подход, основанный на использовании ОИКА, перспективен для решения задачи отбора признаков в линейной регрессии, традиционно решаемой с помощью информационного критерия Акаике. Вместо проведения вычислительно сложной процедуры полного перебора при решении дискретной задачи отбора признаков становится возможным переход к непрерывной задаче гладкой оптимизации и использование ОИКА.

Результаты экспериментов позволяют сделать вывод, что байесовское обучение и информационный подход имеют много общего и, возможно, являются двумя косвенными характеристиками одного и того же явления.

Одним из направлений будущих исследований является применение ОИКА к задаче классификации. Одним из возможных путей здесь является сведение задачи классификации к задаче регрессии (см. [15]).

Авторы выражают признательность В.В. Моттлю за ценные замечания и обсуждение работы.

СПИСОК ЛИТЕРАТУРЫ

1. *Tipping M.E.* The relevance vector machine // Advances Neural Information Processing Systems. 2000. V. 12. P. 652–658.
2. *MacKay D.J.C.* The evidence framework applied to classification networks // Neural Comput. 1992. V. 4. P. 720–736.
3. *Tibshirani R.* Regression shrinkage and selection via the lasso // J. Roy. Stat. Soc. 1996. V. 58. P. 267–288.
4. *Figueiredo M.* Adaptive sparseness for supervised learning // IEEE Trans. Pattern Analys. Mach. Intelligence. 2003. V. 25. P. 1150–1159.
5. *Williams P.M.* Bayesian regularization and pruning using a laplace prior // Neural Comput. 1995. V. 7. P. 117–143.
6. *Cawley G.C., Talbot N.L.C., Girolami M.* Sparse multinomial logistic regression via bayesian l1 regularisation // Advances Neural Informat. Processing Systems. 2007. V. 19. P. 209–216.
7. *Schwarz G.* Estimating the dimension of a model // Ann. Statistics. 1978. V. 6. P. 461–464.
8. *Bishop C.M.* Pattern recognition and machine learning. New York: Springer, 2006.
9. *Akaike H.* A new look at statistical model identification // IEEE Trans. Automatic Control. 1974. V. 25. P. 461–464.
10. *Ширяев А.Н.* Вероятность. М.: Наука, 1979.
11. *Боровков А.А.* Математическая статистика. М.: Физматлит, 2007.
12. *Хорн Р., Джонсон Ч.* Матричный анализ. М.: Мир, 1989.
13. *Spiegelhalter D., Best N., Carlin B., van der Linde A.* Bayesian measures of model complexity and fit // J. Roy. Statist. Soc. 2002. V. 64. P. 583–640.
14. *Faul A.C., Tipping M.E.* Analysis of sparse bayesian learning // Advances Neural Informat. Processing Systems. 2002. V. 14. P. 383–389.
15. *Tipping M.E.* Sparse bayesian learning and the relevance vector machines // J. Mach. Learning Res. 2001. V. 1. P. 211–244.

16. *Kropotov D.A., Vetrov D.P.* On one method of non-diagonal regularization in sparse bayesian learning // Proc. 24th Internat. Conf. Mach. Learning. Corvalis: Omnipress, 2007. P. 457–464.

17. *Qi Y., Minka T., Picard R., Ghahramani Z.* Predictive automatic relevance determination by expectation propagation // Proc. 21st Internat. Conf. Mach. Learning. Banff: Omnipress, 2004. P. 671–678.

18. *Dietterich T.G.* Approximate statistical tests for comparing supervised classification learning algorithms // Neural Comput. 1998. V. 10. P. 1895–1924.

Сдано в набор 10.06.2009 г. Подписано к печати 23.09.2009 г. Формат бумаги $60 \times 88^{1}/_{8}$
Цифровая печать Усл. печ. л. 22.0 Усл. кр.-отт. 6.0 тыс. Уч.-изд. л. 14.0 Бум. л. 11.0
Тираж 269 экз. Зак. 728

Учредители: Российская академия наук, Вычислительный центр им. А.А. Дородницына РАН

Издатель: Академиздатцентр “Наука”, 117997, Москва, Профсоюзная ул., 90
Оригинал-макет подготовлен МАИК “Наука/Интерпериодика”
Отпечатано в ППП “Типография “Наука”, 121099, Москва, Шубинский пер., 6