

С. Г. МУЛЯРЧИК, С. С. БЕЛЯВСКИЙ

МНОГОУРОВНЕВОЕ ПРЕДОБУСЛОВЛИВАНИЕ АЛГЕБРАИЗОВАННЫХ СИСТЕМ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ ЭЛЛИПТИЧЕСКОГО ТИПА

ВВЕДЕНИЕ

Многие явления переноса в жидкой, твердой и газообразной средах описываются краевыми задачами для систем дифференциальных уравнений в частных производных (ДУЧП). Пространственная дискретизация такой системы с использованием конечно-разностного или конечно-элементного подхода и последующая линейаризация сводит краевую задачу к многократному решению систем линейных алгебраических уравнений (СЛАУ). По этой причине эффективность численного метода решения краевой задачи в значительной мере определяется эффективностью метода решения СЛАУ. Для матриц СЛАУ рассматриваемого класса физических задач характерен большой порядок (10^4 – 10^6), сильная разреженность, плохая обусловленность и часто отсутствие симметрии. Из-за больших размеров матриц решение СЛАУ прямыми методами достаточно проблематично [1]. Если при этом учесть, что такую процедуру приходится выполнять многократно и с высокой точностью, то прямые методы оказываются неприемлемыми. Известно [2], что итерационные методы решения СЛАУ с разреженной матрицей имеют целый ряд преимуществ перед прямыми методами, однако плохая обусловленность матриц снижает эффективность многих из них. Для повышения эффективности итерационных методов часто применяют предобуславливающую процедуру [3], т. е. некоторое предварительное преобразование СЛАУ, улучшающее спектральные свойства матрицы.

В настоящей работе рассматриваются два вида предобусловливателей. Первый вид представляет собой неполную факторизацию матрицы СЛАУ [1], применение которой улучшает спектральные свойства линейной системы [4, 5]. Второй вид является проекционным оператором. Его использование дает возможность уменьшить порядок решаемой системы, не ухудшая при этом спектральных свойств матрицы. Существуют различные способы построения такого рода операторов. Для сеточных уравнений они связаны чаще всего с раскрашиванием узлов – подразделением узлов сетки на группы [3]. На практике широко используются красно-черное и двухлинейное раскрашивания [6, 7]. Хотя существуют и другие подходы, например, в [8] предлагается производить подразделение узлов сетки, используя априорные сведения о поведении решения. Такое предположение сужает возможности применения метода.

В последнее время для построения предобуславливателей используется многосеточный (многоуровневый) метод [9–11]. Этот метод привлекает вложенную последовательность грубых сеток, на которых проводится некоторое число шагов релаксационного метода с целью исключения высокочастотных компонент решения. Многосеточный метод можно рассматривать как предобуславливающую процедуру при решении СЛАУ [10] методами псевдодвойственных направлений [11]. В рамках данного подхода широкое распространение получило двухуровневое предобуславливание [3, 6, 7, 12], хотя существуют эффективные процедуры, использующие три [13] и более сеток [14, 15]. Выбор сеток более низкого уровня осуществляется на основании модификаций обобщенного красно-черного упорядочения [14–18] либо упорядочения, ведущего к четно-нечетной редукции [19, 20].

Для решения больших разреженных СЛАУ с симметрической матрицей коэффициентов большой популярностью пользуется метод сопряженных градиентов (СГ) [21] и его различные модификации (см., например, [2, 22–24]), а в случае несимметрических матриц – бисопряженных градиентов (БСГ) [25], который является разновидностью биортогонализационного алгоритма Ланцоша [26] и его усовершенствования CGS [27], CGSTAB [28]. Эти алгоритмы оказались эффективными при анализе процессов переноса заряда в элементах сверхбольших интегральных схем (СБИС) [29, 30], моделировании нефтяных скважин [31], решении системы уравнений Навье–Стокса, описывающей движение вязкой несжимаемой жидкости [32], решении задач теплопереноса [33], моделировании явлений переноса плотной электронно-дырочной плазмы полупроводников [34] и в ряде других важных физических и инженерных приложений.

Целью настоящей работы является развитие многоуровневых методов предобуславливания СЛАУ, порождаемых алгебраизацией на прямоугольных (или топологически инвариантных прямоугольным) сетках и линеаризацией двухмерных краевых задач для ДУЧП эллиптического типа с оператором вида

$$-\nabla \cdot p \nabla u + q \cdot \nabla u = f \quad \text{в} \quad \Omega, \quad (1)$$

$$ru + su_n = g \quad \text{на} \quad \partial\Omega \quad (2)$$

в двумерной области Ω с кусочно-гладкой границей $\partial\Omega$. Алгебраизация задачи при этом осуществляется методом конечных разностей с привлечением пятиточечного шаблона. Матрица A системы

$$Ax = b, \quad A \in \mathbb{R}^{n \times n}, \quad x, b \in \mathbb{R}^n, \quad \det A \neq 0 \quad (3)$$

имеет пятидиагональную структуру и не является в общем случае симметрической ($A \neq A^T$).

Предобуславливающие процедуры строятся на основе так называемых k -линейных упорядочений узлов сетки. Эти упорядочения определяют привлекаемые грубые сетки соответствующего уровня. Разработанные предобуславливающие процедуры реализованы в ряде алгоритмов из класса методов псевдодвойственных направлений и исследованы на задаче двумерного моделирования процессов переноса заряда в кремниевых элементах СБИС с использованием дрейфово-диффузионной модели.

1. ЛИНЕЙНЫЕ УПОРЯДОЧЕНИЯ

В настоящем разделе рассматриваются способы построения последовательности грубых сеток, позволяющие на основании решения некоторым итерационным методом СЛАУ на грубой сетке самого низкого уровня сравнительно несложными преобразованиями восстанавливать решение задачи на основной сетке. Грубые сетки последующего уровня будут формироваться с использованием линейных упорядочений узлов сетки предыдущего уровня. В последующих разделах каждое из рассматриваемых упорядочений будет анализироваться с точки зрения эффективности решения СЛАУ на грубой сетке самого низкого уровня.

1.1. УПОРЯДОЧЕНИЯ НА ОСНОВНОЙ СЕТКЕ

Пусть имеется основная сетка ω размера $n = m_x \times m_y$, где m_x, m_y — число сеточных узлов вдоль осей OX, OY . Занумеруем сначала первую линию вдоль оси OX (слева направо), затем аналогичным образом последующие линии, двигаясь снизу вверх. Дискретизация на такой сетке интегро-интерполяционным методом [35] с использованием пятиточечного шаблона задачи (1), (2) приведет к системе (3), матрица коэффициентов которой является пятидиагональной $(n \times n)$ -матрицей с расстоянием m_x до дальней диагонали.

Построим грубую сетку $\tilde{\omega} \subset \omega$, исключив из основной сетки некоторое количество узлов. В соответствии с подразделением узлов на $\tilde{\omega}$ и $\omega \setminus \tilde{\omega}$ система (3) приобретет вид

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \quad (4)$$

при этом $A_{11} \in \mathbb{R}^{k \times k}$, $x_1 \in \mathbb{R}^k$, $b_1 \in \mathbb{R}^k$ и размерность грубой сетки равна $n-k$. Хо-

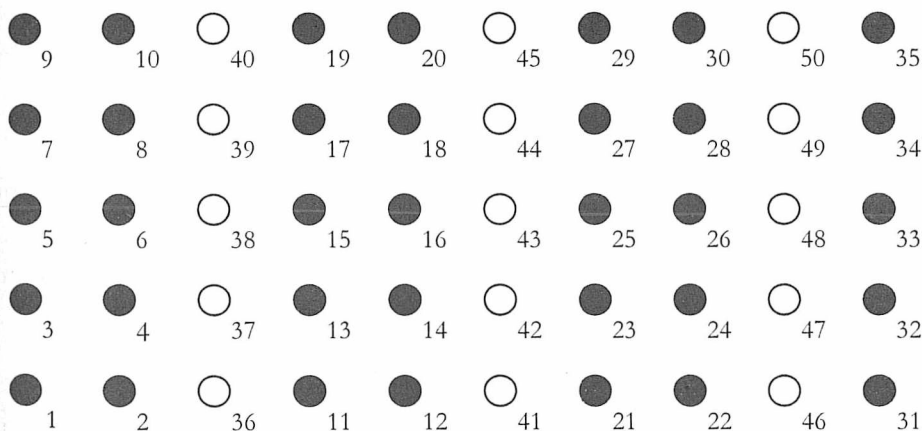


Рис. 1. Пример 3-линейного упорядочения на основной сетке 10×5

рошо известно [3], что исключение узлов одного цвета при красно-черном упорядочении порождает грубую сетку, размеры которой приблизительно вдвое меньше основной, при этом матрицы A_{11} и A_{22} являются диагональными. Изучим структуру матриц A_{ij} , $i, j = 1, 2$ системы (4) для r -линейных упорядочений.

Построим грубую сетку на основании r -линейного упорядочения узлов основной сетки, которое определим следующим образом. Прямоугольную сетку вдоль оси OX разобьем на участки, содержащие r узлов, за исключением, быть может, последнего участка, для которого число узлов может быть меньше r . Сеточные узлы линий, параллельных оси OY , раскрасим черным цветом для первых $r-1$ узлов каждого участка. Если число узлов последнего участка на оси OX меньше r , то все такие узлы раскрасим в черный цвет. Сеточные узлы линий вдоль оси OY , соответствующие узлам оси OX , кратным r , назовем белыми. Сначала слева направо снизу вверх занумеруем поблочно все черные узлы, затем аналогичным образом занумеруем белые узлы (рис. 1). Исключая узлы черного цвета, получим грубую сетку, размер которой приблизительно в r раз меньше размера основной сетки.

Матрица A_{11} при таком выборе грубой сетки будет блочно-диагональной:

$$A_{11} = \text{diag}\{F_1, F_2, \dots, F_m\}. \quad (5)$$

Каждый блок F_i , $i = \overline{1, m-1}$ имеет порядок $m_y(r-1)$. Порядок блока F_m равен $j m_y$, где $j \leq r-1$, $j = m_x - (m-1)r$. Матрица A_{11} , как и все ее блоки F_i , которые далее будем обозначать F (за исключением, быть может, последнего), имеет пятидиагональную структуру с шириной полосы $r-1$. Последний блок имеет такую же структуру, но ширина его полосы составляет j . Матрица A_{22} является трехдиагональной, а матрицы A_{12} и A_{21} — разреженные: A_{12} имеет не более двух ненулевых элементов в каждом столбце, A_{21} по структуре совпадает с A_{12}^T .

Если провести исключение неизвестных x_1 в системе (4), то получим редуцированную систему

$$Sx_2 = \tilde{b}, \quad (6)$$

где матрица $S = A_{22} - A_{21}A_{11}^{-1}A_{12}$ есть не что иное, как дополнение Шура матрицы A_{11} , $\tilde{b} = b_{22} - A_{21}A_{11}^{-1}b_1$.

Отметим, что матрица S , полученная в результате красно-черной редукции, будет девятидиагональной с такой же шириной полосы, как и у матрицы A .

1.2. УПОРЯДОЧЕНИЯ НА ГРУБОЙ СЕТКЕ

Рассмотрим возможность формирования r -линейного упорядочения на грубой сетке $\tilde{\omega}$, полученной после красно-черного упорядочения и исключения узлов одного цвета (рис. 2).

Для конструирования r -линейного упорядочения на грубой сетке введем вспомогательную нумерацию узлов. Занумеруем сверху вниз узлы первой линии вдоль оси OY и продолжим эту нумерацию слева направо на узлах первой линии вдоль оси OX . Разобьем узлы в порядке нумерации на участки, содержащие r узлов, за исключением, быть может, последнего участка, для которого число узлов может быть меньше r . Сеточные узлы линий, расположенных по диагонали (рис. 2), раскрашиваются черным цветом для первых $r-1$ узлов каждого участка. Если число узлов последнего участка меньше r , то все такие узлы раскрашиваются в черный цвет. Сеточные узлы линий, соответствующие узлам с номерами кратными r , будем считать белыми. Сначала слева направо снизу вверх (см. рис. 2) нумеруются поблочно все черные узлы, затем аналогичным образом нумеруются белые узлы. Исключая узлы черного цвета, получим грубую сетку $\tilde{\omega}_1$, размер которой приблизительно в r раз меньше размера $\tilde{\omega}$. В соответствии с подразделением узлов на $\tilde{\omega}_1$ и $\tilde{\omega} \setminus \tilde{\omega}_1$ система приобретает вид

$$\begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \begin{bmatrix} x_{21} \\ x_{22} \end{bmatrix} = \begin{bmatrix} \tilde{b}_1 \\ \tilde{b}_2 \end{bmatrix}. \quad (7)$$

Матрица S_{11} такой системы, как и матрица A_{11} системы (4), будет блочно-диагональной, каждый блок которой, кроме, может быть, последнего, имеет порядок $m_y(r-1)$. Блоки матрицы S_{11} при $r=2$ будут трехдиагональными, при $r=3$ — семидиагональными, при $r \geq 4$ — девятидиагональными с расстоянием до дальней диагонали, равным $2r-3$. Матрицы S_{22} , S_{12} и S_{21} имеют такую же структуру, как и соответствующие подматрицы матрицы A в представлении (4).

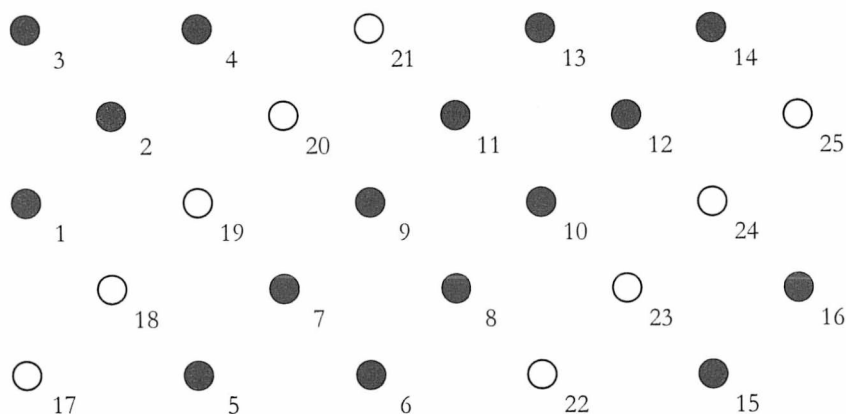


Рис. 2. Пример 3-линейного упорядочения на грубой сетке 5×5

2. ПРОЕКЦИОННЫЙ МЕТОД В ДВУХУРОВНЕВОМ ПРЕДОБУСЛОВЛИВАНИИ

2.1. СИММЕТРИЧНЫЙ СЛУЧАЙ

Из итерационных методов решения СЛАУ с симметричной положительно определенной $(n \times n)$ -матрицей наиболее эффективным является метод сопряженных градиентов в сочетании с подходящей предобуславливающей процедурой. Его основой является построение n векторов $p_0, p_1, \dots, p_{n-1}$, образующих базис в пространстве R^n и удовлетворяющих условию A -сопряженности

$$p_i^T A p_j = 0, \quad i \neq j. \quad (8)$$

Реализация этого метода может быть осуществлена по известной вычислительной схеме [36], на которую будем ссылаться как на НФСГ-метод.

Предположим теперь, что известны k ($k < n$) векторов

$$p_0, p_1, \dots, p_{k-1}, \quad (9)$$

удовлетворяющих условию (8). Их линейная оболочка образует подпространство $E_1 \subset R^n$, $\dim E_1 = k$. Нетрудно убедиться, что

$$Q = \sum_{i=1}^{k-1} \frac{p_i (A p_i)^T}{p_i^T A p_i} \quad (10)$$

является проектором на E_1 . Аналогично $R = I - Q$ (I — единичная матрица) есть проектор на некоторое подпространство E_2 . Согласно [37], $R^n = E_1 \oplus E_2$. Из (8) и (10) следуют соотношения

$$AQ = Q^T A = Q^T A Q, \quad AR = R^T A = R^T A R.$$

Легко доказывается следующая

Теорема 1. Если известно k векторов $p_0, p_1, \dots, p_{k-1}$, удовлетворяющих условию A -сопряженности, то решение системы (3) можно представить в виде

$$x = \sum_{i=0}^{k-1} \alpha_i p_i + z, \quad (11)$$

где $\alpha_i = \frac{p_i^T b}{p_i^T A p_i}$, вектор $z \in E_2$ и удовлетворяет системе

$$Az = R^T b. \quad (12)$$

Так как $z = R x$, вместо (12) рассмотрим систему

$$ARx = R^T b \quad (13)$$

с вырожденной матрицей $\mathbf{AR}$. Система (13) совместна в силу того, что $\text{rank } \mathbf{AR} = \text{rank } [\mathbf{AR}, \mathbf{R}^T \mathbf{b}] = \text{rank } \mathbf{R} = n - k$. Для решения системы (13) может быть применен метод сопряженных градиентов [22]. Причем векторы сопряженных градиентов, по которым строится решение системы (13), ортогональны гиперплоскости $\mathbf{AR}\mathbf{q} = 0$. Число таких векторов равно $n - k$. Обозначим их $\mathbf{p}_k, \mathbf{p}_{k+1}, \dots, \mathbf{p}_{n-1}$.

Вследствие вырожденности матрицы $\mathbf{AR}$ система (13) имеет множество решений. Их можно представить в виде

$$\mathbf{x} = \mathbf{y} + \sum_{i=k}^{n-1} \alpha_i \mathbf{p}_i,$$

где α_i — коэффициенты разложения вектора $\mathbf{z}$ по векторам $\mathbf{p}_k, \mathbf{p}_{k+1}, \dots, \mathbf{p}_{n-1}$, а $\mathbf{y}$ — произвольный вектор из E_1 . Положив

$$\mathbf{y} = \mathbf{x}_0 = \sum_{i=0}^{k-1} \alpha_i \mathbf{p}_i, \quad (14)$$

получим, согласно (11), решение системы (3). Таким образом, если известно k векторов $\mathbf{p}_0, \mathbf{p}_1, \dots, \mathbf{p}_{k-1}$, удовлетворяющих условию $\mathcal{A}$ -сопряженности, можно построить, используя (10), предобуславливающую матрицу $\mathbf{R} = \mathbf{I} - \mathbf{Q}$.

Если в схеме метода сопряженных градиентов (НФСГ) решение системы (3) искать с использованием предобуславливателя $\mathbf{R}$, то получим вычислительную схему ПСГ-метода.

Алгоритм 1 (ПСГ)

- 1) Построение $\mathbf{x}_0$ в соответствии с (14).
- 2) Вычисление начального направления: $\mathbf{r}_0 = \mathbf{b} - \mathbf{Ax}_0$, $\tilde{\mathbf{p}}_0 = \mathbf{Rr}_0$.
- 3) Итерационное уточнение ($i = 0, 1, 2, \dots$):

$$3.1) \alpha_i = \frac{\mathbf{r}_i^T \mathbf{Rr}_i}{\tilde{\mathbf{p}}_i^T \mathbf{A} \tilde{\mathbf{p}}_i},$$

$$3.2) \mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \tilde{\mathbf{p}}_i,$$

$$3.3) \mathbf{r}_{i+1} = \mathbf{r}_i - \alpha_i \mathbf{A} \tilde{\mathbf{p}}_i,$$

$$3.4) \beta_i = \frac{\mathbf{r}_{i+1}^T \mathbf{Rr}_{i+1}}{\mathbf{r}_i^T \mathbf{Rr}_i},$$

$$3.5) \mathbf{p}_{i+1} = \mathbf{Rr}_{i+1} + \beta_i \tilde{\mathbf{p}}_i.$$

Схема ПСГ-метода имеет очевидное преимущество перед схемой СГ-метода: в отсутствие ошибок округления первая из них обеспечивает сходимость к решению в общем случае через $(n - k)$ итераций, вторая — через n итераций.

Скорость сходимости метода сопряженных градиентов зависит от числа обусловленности $\chi = \lambda_{\max} / \lambda_{\min}$ матрицы системы, где $\lambda_{\max}$ и $\lambda_{\min}$ — ее максимальное и минимальное собственные значения. Известно [38], что удачный выбор положительно определенной симметрической матрицы $\mathbf{H}$ позволяет уменьшить число обусловленности матрицы $\mathbf{H}^{-1}\mathbf{A}$ [39]. Поэтому вместо (3) решают систему $\mathbf{H}^{-1}\mathbf{A}\mathbf{x} = \mathbf{H}^{-1}\mathbf{b}$ по схеме НФСГ-метода.

Если заданы k $\mathbf{A}$ -сопряженных векторов, то, построив предобуславливатель $\mathbf{R}$ и найдя начальное приближение $\mathbf{x}_0$ по формуле (14), решение системы (3) можно найти, вычислив компоненту $\mathbf{z}$ (см. формулу (11)) по схеме НФПСГ-метода. Эта схема получается из схемы НФСГ-метода так же, как схема ПСГ-метода из схемы СГ-метода.

Алгоритм 2 (НФПСГ)

- 1) Построение матрицы $\mathbf{H}$.
- 2) Построение $\mathbf{x}_0$ в соответствии с (14).
- 3) Вычисление начального направления: $\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$, $\tilde{\mathbf{p}}_0 = \mathbf{R}\mathbf{H}^{-1}\mathbf{r}_0$.
- 4) Итерационное уточнение ($i = 0, 1, 2, \dots$):

$$4.1) \alpha_i = \frac{\mathbf{r}_i^T \mathbf{R} \mathbf{H}^{-1} \mathbf{r}_i}{\tilde{\mathbf{p}}_i^T \mathbf{A} \tilde{\mathbf{p}}_i},$$

$$4.2) \mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \tilde{\mathbf{p}}_i,$$

$$4.3) \mathbf{r}_{i+1} = \mathbf{r}_i - \alpha_i \mathbf{A} \tilde{\mathbf{p}}_i,$$

$$4.4) \beta_i = \frac{\mathbf{r}_{i+1}^T \mathbf{R} \mathbf{H}^{-1} \mathbf{r}_{i+1}}{\mathbf{r}_i^T \mathbf{R} \mathbf{H}^{-1} \mathbf{r}_i},$$

$$4.5) \mathbf{p}_{i+1} = \mathbf{R} \mathbf{H}^{-1} \mathbf{r}_{i+1} + \beta_i \tilde{\mathbf{p}}_i.$$

Отметим, что при реализации схемы НФПСГ-метода фактически решается система

$$\mathbf{H}^{-1} \mathbf{A} \mathbf{R} \mathbf{x} = \mathbf{H}^{-1} \mathbf{R}^T \mathbf{b} \quad (15)$$

от начального приближения $\mathbf{x}_0$. Используя экстремальные свойства характеристических чисел регулярного пучка квадратичных форм [40], нетрудно убедиться, что ненулевые собственные значения матрицы $\mathbf{H}^{-1} \mathbf{A} \mathbf{R}$ лежат на промежутке $[\lambda_{\min}(\mathbf{H}^{-1} \mathbf{A}), \lambda_{\max}(\mathbf{H}^{-1} \mathbf{A})]$, а соответствующие им собственные векторы принадлежат подпространству E_2 . Так как решение системы (15) ищется в E_2 , то число

обусловленности матрицы $\mathbf{H}^{-1}\mathbf{A}\mathbf{R}$ не превосходит числа обусловленности матрицы $\mathbf{H}^{-1}\mathbf{A}$. Поэтому сходимость НФПСГ-метода не хуже сходимости метода НФСГ.

Для построения $\mathcal{A}$ -сопряженных векторов $\mathbf{p}_0, \mathbf{p}_1, \dots, \mathbf{p}_{k-1}$ распространим известный метод ортогонализации [40]. Возьмем некоторый набор линейно независимых векторов $\mathbf{q}_0, \mathbf{q}_1, \dots, \mathbf{q}_k$ и проведем над ними процедуру $\mathcal{A}$ -ортогонализации Грамма – Шмидта. Выясним структуру матриц $\mathbf{Q}$ и $\mathbf{R}$ при специальном выборе векторов $\mathbf{q}_0, \dots, \mathbf{q}_k$. Возьмем k натуральных чисел $i_1, i_2, \dots, i_k$ таких, что $1 \leq i_1 \leq i_2 \leq \dots \leq i_k \leq n$, и введем следующие числовые множества: $N_n = \{1, 2, \dots, n\}$, $N_n^1 = \{i_1, i_2, \dots, i_k\}$, $N_n^2 = N_n / N_n^1$. Пусть $\mathbf{e}_i$ – единичный вектор, у которого в i -й позиции стоит единица, а все остальные компоненты – нули. Положим $\mathbf{q}_0 = \mathbf{e}_{i_1}, \mathbf{q}_1 = \mathbf{e}_{i_2}, \dots, \mathbf{q}_{k-1} = \mathbf{e}_{i_k}$ и на их основании построим $\mathcal{A}$ -сопряженные векторы $\mathbf{p}_0, \mathbf{p}_1, \dots, \mathbf{p}_{k-1}$. Сформируем затем проекторы $\mathbf{Q}$ и $\mathbf{R}$, что дает возможность определить подпространства E_1 и E_2 .

Теорема 2. Все столбцы матрицы $\mathbf{R}$ с номерами из N_n^1 и все строки матрицы $\mathbf{Q}$ с номерами из N_n^2 нулевые.

Доказательство теоремы можно найти в [23].

Вышеприведенные рассуждения справедливы для любой системы (3) с положительно определенной матрицей $\mathbf{A}$. Обратимся теперь к случаю, когда система (3) является системой сеточных уравнений. Для построения оператора $\mathbf{R}$ воспользуемся r -линейным упорядочением узлов основной сетки. Каждому черному узлу поставим в соответствие вектор $\mathbf{e}_i$, где i – номер черного узла в первоначальной нумерации. На их основании можно построить набор $\mathbf{p}_0, \mathbf{p}_1, \dots, \mathbf{p}_{k-1}$ $\mathcal{A}$ -сопряженных векторов. Так как каждому исключаемому узлу соответствует вектор $\mathbf{e}_i$, то, согласно теореме 2, предобуславливатель $\mathbf{R}$ исключает из системы (3) приблизительно $m_y(m_x - [m_x/r])$ уравнений.

Практическое построение оператора $\mathbf{R}$ состоит в следующем. Умножение системы (3) на матрицу $\mathbf{R}$ равносильно исключению из системы уравнений, соответствующих узлам $\omega \setminus \tilde{\omega}$ основной сетки. Поэтому

$$\mathbf{Q} = \begin{bmatrix} \mathbf{A}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{A}_{11}^{-1}\mathbf{A}_{12} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \text{ и } \mathbf{R} = \begin{bmatrix} \mathbf{0} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}.$$

Следовательно, чтобы построить проектор $\mathbf{R}$, необходимо найти $\mathbf{A}_{11}^{-1}$ и $\mathbf{A}_{12}$, сохраняя при этом первоначальную нумерацию строк. Такое сохранение необходимо для построения предобуславливателя $\mathbf{H}$ матрицы $\mathbf{A}$.

Вычислительные затраты предложенных итерационных схем, реализующих идеи метода сопряженных градиентов, включают издержки на построение матрицы $\mathbf{R}$ и связаны в основном с трудоемкостью обращения матрицы $\mathbf{A}_{11}$. Так как проектор $\mathbf{R}$ используется на каждой итерации в ПСГ и НФПСГ-методах, то его следует

представить в форме, допускающей эффективную реализацию произведения $\mathbf{Rq}$, где $\mathbf{q}$ – произвольный вектор. Наиболее эффективным в этом случае является использование полной факторизации матрицы $\mathbf{A}_{11}$. Затраты на ее построение и хранение зависят от расстояния до дальней диагонали, которое, в свою очередь, определяется l -линейным упорядочением узлов основной сетки. Так, при двухлинейном упорядочении матрица $\mathbf{A}_{11}$ является трехдиагональной, ее полная факторизация не порождает новых ненулевых элементов. При трехлинейном упорядочении матрица $\mathbf{A}_{11}$ имеет пятидиагональную структуру с расстоянием до дальней диагонали, равным 2. Поскольку ее ближняя диагональ содержит приблизительно $1/3$ нулевых элементов, то полная факторизация такой матрицы порождает $n/3$ новых ненулевых элементов. С другой стороны, чем шире полоса ненулевых элементов у матрицы $\mathbf{A}_{11}$, тем меньшие затраты на выполнение других арифметических операций в алгоритмах ПСГ и НФСГ. Причина этого кроется в том, что размерность матрично-векторных произведений, скалярных произведений и операций типа $\mathbf{s} = \mathbf{ax} + \mathbf{y}$ ($\mathbf{saxpy}$) составляет $n - k$ вместо n . Поэтому затраты на итерациях алгоритмов ПСГ и НФСГ для некоторых l -линейных упорядочений (см. табл. 1) оказываются заметно меньшими, чем в традиционных алгоритмах СГ и НФСГ. В табл. 1 приведены затраты на итерации СГ-, ПСГ-, НФСГ- и НФСГ-методов при решении СЛАУ с пятидиагональной матрицей и использовании экономичной трехдиагональной неполной LU -факторизации [41]. Здесь и далее через t_y и t_{cl} обозначено время выполнения арифметических операций умножения (деления) и сложения. Отметим также, что для реализации алгоритмов, перечисленных в табл. 1, не требуется дополнительная оперативная память.

Таблица 1

Оценка эффективности исследуемых алгоритмов

Метод	ПСГ	НФСГ
Упорядочение	Цена итерации	Цена итерации
–	$9 t_{cl} + 10 t_y$	$13 t_{cl} + 16 t_y$
Красно-черное	$6 t_{cl} + 8 t_y$	$10 t_{cl} + 14 t_y$
Двухлинейное	$6 t_{cl} + 8 t_y$	$10 t_{cl} + 14 t_y$
Трехлинейное	$17/3 t_{cl} + 26/3 t_y$	$29/3 t_{cl} + 44/3 t_y$
Четырехлинейное	$5,5 t_{cl} + 9,75 t_y$	$10,5 t_{cl} + 15,75 t_y$

2.2. НЕСИММЕТРИЧНЫЙ СЛУЧАЙ

Рассмотрим систему (3) в случае, когда матрица $\mathbf{A}$ является несимметричной. Эффективным методом ее решения является БСГ-метод [25], который предполагает построение взаимных базисов $\mathbf{p}_0, \mathbf{p}_1, \dots, \mathbf{p}_{n-1}$ и $\hat{\mathbf{p}}_0, \hat{\mathbf{p}}_1, \dots, \hat{\mathbf{p}}_{n-1}$ в $\mathbf{R}^n$ по методу Ланцоша [25], удовлетворяющих условию бисопряженности:

$$\hat{\mathbf{p}}_i^T \mathbf{A} \mathbf{p}_j = \mathbf{p}_i^T \mathbf{A}^T \hat{\mathbf{p}}_j = 0, \quad i \neq j, \quad \hat{\mathbf{p}}_i^T \mathbf{A} \mathbf{p}_i = \mathbf{p}_i^T \mathbf{A}^T \hat{\mathbf{p}}_i \neq 0. \quad (17)$$

Решение системы (3) можно представить, как и в случае симметричной матрицы, в виде линейной комбинации векторов $\mathbf{p}_0, \mathbf{p}_1, \dots, \mathbf{p}_{n-1}$.

Допустим, что известны векторы

$$\mathbf{p}_0, \mathbf{p}_1, \dots, \mathbf{p}_{k-1} \text{ и } \hat{\mathbf{p}}_0, \hat{\mathbf{p}}_1, \dots, \hat{\mathbf{p}}_{k-1}, \quad (18)$$

где $k < n$, удовлетворяющие условию (17). Обозначим их линейные оболочки через E и F соответственно. Рассуждая аналогично п. 2.1, построим проектор

$$\mathbf{R} = \mathbf{I} - \sum_{i=0}^{k-1} \frac{\mathbf{p}_i (\mathbf{A}^T \hat{\mathbf{p}}_i)^T}{\hat{\mathbf{p}}_i^T \mathbf{A} \mathbf{p}_i}$$

на $F^\perp$ вдоль E , где $F^\perp$ — A -ортогональное дополнение к F и проектор

$$\hat{\mathbf{R}} = \mathbf{I} - \sum_{i=0}^{k-1} \frac{\hat{\mathbf{p}}_i (\mathbf{A} \mathbf{p}_i)^T}{\hat{\mathbf{p}}_i^T \mathbf{A} \mathbf{p}_i}$$

на $E^\perp$ вдоль F .

Решение системы (3) представим в виде

$$\mathbf{x} = \mathbf{y} + \mathbf{z} = \sum_{i=0}^{k-1} \alpha_i \mathbf{p}_i + \mathbf{z},$$

где $\alpha_i = \frac{\hat{\mathbf{p}}_i^T \mathbf{b}}{\hat{\mathbf{p}}_i^T \mathbf{A} \mathbf{p}_i}$, а $\mathbf{z} \in E^\perp$. В этом случае достаточно определить лишь компоненту

$\mathbf{z}$, которая является решением системы

$$\mathbf{A} \mathbf{z} = \hat{\mathbf{R}}^T \mathbf{b}.$$

Можно показать, что $\mathbf{z} = \mathbf{R} \mathbf{x}$, и поэтому вместо (3) достаточно решить СЛАУ

$$\mathbf{A} \mathbf{R} \mathbf{x} = \hat{\mathbf{R}}^T \mathbf{b}. \quad (19)$$

Для того чтобы $\mathbf{x}$ из (19) являлось решением системы (3), необходимо определить дополнительное условие. Таким условием будет начальное приближение, вычисленное в подпространстве E :

$$\mathbf{y} = \sum_{i=0}^{k-1} \alpha_i \mathbf{p}_i. \quad (20)$$

Известно, что БСГ-метод наряду с (3) решает также сопряженную систему $\mathbf{A}^T \hat{\mathbf{x}} = \mathbf{b}$, начальное приближение в подпространстве F для которой имеет вид

$$\hat{\mathbf{y}} = \sum_{i=0}^{k-1} \hat{\alpha}_i \hat{\mathbf{p}}_i, \quad (21)$$

где $\hat{\alpha}_i = \frac{\mathbf{p}_i^T \mathbf{b}}{\hat{\mathbf{p}}_i^T \mathbf{A} \mathbf{p}_i}$.

Соотношения (20) и (21) определяют начальное приближение для БСГ-метода. Необходимо в процессе итерационного уточнения обеспечить A -ортогональность вновь построенных векторов $\mathbf{p}$ к гиперплоскости $\mathbf{ARw} = 0$ и A^T -ортогональность векторов $\hat{\mathbf{p}}$ к гиперплоскости $\mathbf{A}^T \hat{\mathbf{R}}\mathbf{w} = 0$. Требуемая ортогонализация осуществляется в рамках предлагаемого ниже проекционного БСГ-алгоритма (ПБСГ).

Алгоритм 3 (ПБСГ)

1. Построение предобусловливателя $\mathbf{H}$.
2. Вычисление начальных приближений $\mathbf{x}_0 = \mathbf{y}$ и $\hat{\mathbf{x}}_0 = \hat{\mathbf{y}}$ из (20), (21).
3. Инициализация: $\mathbf{r}_0 = \mathbf{b} - \mathbf{Ax}_0$, $\hat{\mathbf{r}}_0 = \mathbf{b} - \mathbf{A}^T \hat{\mathbf{x}}_0$, $\mathbf{p}_0 = \hat{\mathbf{p}}_0 = \mathbf{0}$, $\rho_0 = 1$.
4. Итерационное уточнение ($i=1,2,3,\dots$):

4.1) $\mathbf{H}\tilde{\mathbf{y}} = \mathbf{r}_{i-1}$;	4.7) $\hat{\mathbf{p}}_i = \hat{\mathbf{R}}\mathbf{z} + \beta_i \hat{\mathbf{p}}_{i-1}$;
4.2) $\mathbf{y} = \mathbf{R}\tilde{\mathbf{y}}$;	4.8) $\mathbf{v} = \mathbf{Ap}_i$;
4.3) $\rho_i = \hat{\mathbf{r}}_{i-1}^T \mathbf{y}$;	4.9) $\alpha_i = \rho_i / (\hat{\mathbf{p}}_i^T \mathbf{v})$;
4.4) $\beta_i = \rho_i / \rho_{i-1}$;	4.10) $\mathbf{x}_i = \mathbf{x}_{i-1} + \alpha_i \mathbf{p}_i$;
4.5) $\mathbf{p}_i = \mathbf{y} + \beta_i \mathbf{p}_{i-1}$;	4.11) $\mathbf{r}_i = \mathbf{r}_{i-1} - \alpha_i \mathbf{v}$;
4.6) $\mathbf{H}^T \mathbf{z} = \hat{\mathbf{r}}_{i-1}$;	4.12) $\hat{\mathbf{r}}_i = \hat{\mathbf{r}}_{i-1} - \alpha_i \mathbf{A}^T \hat{\mathbf{p}}_i$.

Если удастся построить достаточно большое число бисопряженных векторов (18), удовлетворяющих условию (17), и на их основе осуществить вычисление проекционных операторов $\mathbf{R}$, $\hat{\mathbf{R}}$, то затраты ПБСГ-алгоритма на итерации могут быть существенно меньшими, чем у традиционного БСГ-алгоритма [25]. Причина этого состоит в том, что размерность матрично-векторных произведений, скалярных произведений и операций типа $\mathbf{s} = \mathbf{ax} + \mathbf{y}$ (сахру) составляет $n - k$ вместо n (экономия на этих операциях должна быть больше, чем затраты на двух дополнительных операциях проектирования).

В алгоритме БСГ последовательности $\mathbf{r}_i$ и $\hat{\mathbf{r}}_i$ сходятся к нулю. Для получения решения исходной системы (3) используется свойство сходимости к нулю только последовательности $\mathbf{r}_i$. В алгоритме CGS [27] последовательность $\hat{\mathbf{r}}_i$ вообще не строится, а все усилия сконцентрированы на сходимости к нулю $\mathbf{r}_i$.

Применяя предложенную методику проекционного предобусловливания, построим проекционный алгоритм метода CGS.

Алгоритм 4 (ПГСГ)

1. Построение предобусловливателя $\mathbf{H}$.
2. Вычисление начальных приближений $\mathbf{x}_0 = \mathbf{y}$ и $\hat{\mathbf{x}}_0 = \hat{\mathbf{y}}$ из (20), (21).
3. Инициализация: $\mathbf{r}_0 = \mathbf{b} - \mathbf{Ax}_0$, $\hat{\mathbf{r}} = \mathbf{b} - \mathbf{A}^T \mathbf{x}_0$, $\mathbf{p}_0 = \mathbf{q}_0 = \mathbf{0}$, $\rho_0 = 1$.
4. Итерационное уточнение ($i=1,2,3,\dots$):

$$4.1) \rho_i = \hat{\mathbf{r}}^T \mathbf{r}_{i-1};$$

$$4.2) \beta_i = \rho_i / \rho_{i-1};$$

$$4.3) \mathbf{u} = \mathbf{r}_{i-1} + \beta_i \mathbf{q}_{i-1};$$

$$4.4) \mathbf{p}_i = \mathbf{u} + \beta_i (\mathbf{q}_{i-1} + \beta_i \mathbf{p}_{i-1});$$

$$4.5) \mathbf{H}\tilde{\mathbf{y}} = \mathbf{p}_i;$$

$$4.6) \mathbf{y} = \mathbf{R}\tilde{\mathbf{y}};$$

$$4.7) \mathbf{v} = \mathbf{A}\mathbf{y};$$

$$4.8) \alpha_i = \rho_i / (\hat{\mathbf{r}}^T \mathbf{v});$$

$$4.9) \mathbf{q}_i = \mathbf{u} - \alpha_i \mathbf{v};$$

$$4.10) \mathbf{H}\tilde{\mathbf{z}} = \mathbf{u} + \mathbf{q}_i;$$

$$4.11) \mathbf{z} = \mathbf{R}\tilde{\mathbf{z}};$$

$$4.12) \mathbf{x}_i = \mathbf{x}_{i-1} + \alpha_i \mathbf{z};$$

$$4.13) \mathbf{r}_i = \mathbf{r}_{i-1} - \alpha_i \mathbf{A}\mathbf{z}.$$

В данном алгоритме операции вычисления скалярного произведения, матрично-векторного умножения, скахру имеют размерность $n - k$. Абсолютный выигрыш для алгоритма CGS в результате применения проекционного предобусловливания оказывается на одну операцию скахру большим, чем для метода БСГ. В то же время количество операций проектирования осталось прежним.

Результаты работ ряда авторов свидетельствуют о преимуществе CGS-метода перед методом БСГ (см., к примеру, [25, 30, 31]). Однако оба этих метода имеют негативную особенность, состоящую в отсутствии минимизирующих свойств. Вследствие этого наблюдается немонотонная сходимость к нулю последовательности невязок. Алгоритм CGSTAB [28] свободен от этого недостатка благодаря полиномиальной форме метода. CGSTAB минимизирует невязку в норме $\|\mathbf{r}_i\|_2$. Это свойство подтверждается проведенным в [42] динамическим контролем.

Применяя процедуру проекционного предобусловливания, построим проекционный CGSTAB-алгоритм.

Алгоритм 5 (PCGSTAB)

1. Построение предобусловливателя $\mathbf{H}=\mathbf{LU}$.

2. Вычисление начальных приближений $\mathbf{x}_0 = \mathbf{y}$ и $\hat{\mathbf{x}}_0 = \hat{\mathbf{y}}$ из (20), (21).

3. Инициализация: $\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$, $\hat{\mathbf{r}}_0 = \mathbf{b} - \mathbf{A}^T \hat{\mathbf{x}}_0$, $\mathbf{p}_0 = \mathbf{v}_0 = \mathbf{0}$, $\rho_0 = \alpha = \omega_0 = 1$.

4. Итерационное уточнение ($i=1,2,3,\dots$):

$$4.1) \rho_i = \hat{\mathbf{r}}^T \mathbf{r}_{i-1};$$

$$4.2) \beta_i = (\rho_i / \rho_{i-1})(\alpha / \omega_{i-1});$$

$$4.3) \mathbf{p}_i = \mathbf{r}_{i-1} + \beta_i (\mathbf{p}_{i-1} - \omega_{i-1} \mathbf{v}_{i-1});$$

$$4.4) \mathbf{H}\tilde{\mathbf{y}} = \mathbf{p}_i;$$

$$4.5) \mathbf{y} = \mathbf{R}\tilde{\mathbf{y}};$$

$$4.6) \mathbf{v}_i = \mathbf{A}\mathbf{y};$$

$$4.7) \alpha = \rho_i / (\hat{\mathbf{r}}^T \mathbf{v}_i);$$

$$4.8) \mathbf{q} = \mathbf{r}_{i-1} - \alpha \mathbf{v}_i;$$

$$4.9) \mathbf{H}\tilde{\mathbf{z}} = \mathbf{q};$$

$$4.10) \mathbf{z} = \mathbf{R}\tilde{\mathbf{z}};$$

$$4.11) \mathbf{t} = \mathbf{A}\mathbf{z};$$

$$4.12) \omega_i = ((\mathbf{L}^{-1} \mathbf{t})^T (\mathbf{L}^{-1} \mathbf{q})) / ((\mathbf{L}^{-1} \mathbf{t})^T (\mathbf{L}^{-1} \mathbf{t}));$$

$$4.13) \mathbf{x}_i = \mathbf{x}_{i-1} + \alpha \mathbf{y} + \omega_i \mathbf{z};$$

$$4.14) \mathbf{r}_i = \mathbf{q} - \omega_i \mathbf{t}.$$

Заметим, что п. 4.12 предполагает решение одной вспомогательной системы с нижней треугольной матрицей, в то время как вектор $L^{-1}q$ является промежуточным результатом расчетов в п. 4.9. Нетрудно убедиться, что применение проекционного преобразования в алгоритме CGSTAB является еще более эффективным, чем в двух предыдущих методах, поскольку в данном случае размерность $n - k$ имеют четыре скалярных произведения, два матрично-векторных умножения и четыре операции над векторами типа $sa \cdot ru$.

Оценим эффективность проекционных методов на СЛАУ с пятидиагональной матрицей. Как и в случае симметрических матриц, грубую сетку будем строить на основе r -линейных упорядочений. В табл. 2 приведены затраты на итерации БСГ-, CGS-, CGSTAB-методов в их традиционном исполнении и при использовании проекционного предобуславливания на основе 2- и 3-линейных упорядочений. Видно, что построенные в этом разделе алгоритмы обладают меньшими затратами на итерации, чем традиционные версии [25, 27, 28]. В качестве H в этих методах используется неполное LU -разложение матрицы A вида $ILU(0)$ [41].

Таблица 2

Оценка эффективности исследуемых алгоритмов

Метод	БСГ	CGS	CGSTAB
—	$29t_H + 23t_{CA}$	$30t_H + 25t_{CA}$	$35t_H + 28t_{CA}$
Двухлинейное	$25t_H + 18t_{CA}$	$25,5t_H + 19t_{CA}$	$29,5t_H + 21,5t_{CA}$
Трехлинейное	$78/3t_H + 53/3t_{CA}$	$80/3t_H + 55/3t_{CA}$	$91/3t_H + 62/3t_{CA}$

3. МЕТОД РЕДУКЦИИ В ДВУХЛИНЕЙНОМ ПРЕДОБУСЛОВАИВАНИИ

Пусть проведена редукция СЛАУ (4) в соответствии с r -линейным упорядочением сеточных узлов и получена система (6). Будем рассматривать предобуславливающие матрицы для дополнения Шура S матрицы A_{11} . Для этого аппроксимируем A_{11}^{-1} с помощью B таким образом, чтобы матрица $A_{22} - A_{21}BA_{12}$ имела простую структуру. Разрабатываемые далее двухсеточные методы основаны на возможности привлечения таких упорядочений, которые обеспечивают эффективное решение линейной системы с матрицей A_{11} прямыми методами. В этом случае результирующий алгоритм решения системы (3) принимает следующий вид.

Алгоритм 6

- 1) Решить систему $A_{11}v_1 = b_1$.
- 2) $v_2 \leftarrow b_2 - A_{21}v_1$.
- 3) Решить систему $Sx_2 = v_2$.
- 4) $y \leftarrow A_{12}x_2$.
- 5) Решить систему $A_{11}z = y$.
- 6) $x_1 \leftarrow v_1 - z$.

Система с матрицей $\mathbf{S}$ на шаге 3) алгоритма решается одним из итерационных методов, в то время как системы с матрицей $\mathbf{A}_{11}$ на шагах 1), 5) и в итерационном цикле шага 3) – прямым методом. Практические реализации алгоритма 7 отличаются друг от друга способом представления системы в виде (4), т. е. упорядочениями узлов сетки пространственной дискретизации.

Известны случаи применения рассмотренного приема в сочетании с красным и простейшим двухлинейным упорядочениями [14, 18], которые приводят к матрице $\mathbf{S}$ порядка примерно вдвое меньшего, чем порядок исходной матрицы $\mathbf{A}$. Скорость сходимости итерационных алгоритмов зависит от распределения собственных значений матрицы системы. Нами изучены спектральные свойства матрицы $\mathbf{S}$ при двух-, трех- и четырехлинейном упорядочении узлов сетки. В [43] даны оценки спектрального радиуса редуцированной СЛАУ, а в [44] получены оценки спектрального радиуса итерационной матрицы для решения редуцированной системы, аппроксимирующей конвективно-диффузионное уравнение с постоянными коэффициентами на прямоугольных сетках, в зависимости от способа аппроксимации первых производных.

Сосредоточим внимание на использовании предобусловленных с привлечением неполной факторизации СГ- и БСГ-методов для решения системы вида $\mathbf{S}\mathbf{x}=\mathbf{b}$. Приведем предобусловленный алгоритм БСГ-метода [25].

Алгоритм 7 (БСГ)

1. Построение предобусловливателя $\mathbf{H}$ для $\mathbf{S}$.

2. Инициализация: $\mathbf{r}_0 = \mathbf{b} - \mathbf{S}\mathbf{x}_0$, $\hat{\mathbf{r}}_0 = \mathbf{r}_0$, $\mathbf{p}_0 = \hat{\mathbf{p}}_0 = \mathbf{0}$, $\rho_0 = 1$.

3. Итерационное уточнение ($i=1,2,3,\dots$):

$$3.1) \mathbf{H}\mathbf{y} = \mathbf{r}_{i-1};$$

$$3.7) \mathbf{v} = \mathbf{S}\mathbf{p}_i;$$

$$3.2) \rho_i = \hat{\mathbf{r}}_{i-1}^T \mathbf{y};$$

$$3.8) \alpha_i = \rho_i / (\hat{\mathbf{p}}_i^T \mathbf{v});$$

$$3.3) \beta_i = \rho_i / \rho_{i-1};$$

$$3.9) \mathbf{x}_i = \mathbf{x}_{i+1} + \alpha_i \mathbf{p}_i;$$

$$3.4) \mathbf{p}_i = \mathbf{y} + \beta_i \mathbf{p}_{i-1};$$

$$3.10) \mathbf{r}_i = \mathbf{r}_{i-1} - \alpha_i \mathbf{v};$$

$$3.5) \mathbf{H}^T \mathbf{z} = \hat{\mathbf{r}}_{i-1};$$

$$3.11) \hat{\mathbf{r}}_i = \hat{\mathbf{r}}_{i-1} - \alpha_i \mathbf{S}^T \hat{\mathbf{p}}_i.$$

$$3.6) \hat{\mathbf{p}}_i = \mathbf{z} + \beta_i \hat{\mathbf{p}}_{i-1};$$

Для организации умножения матриц $\mathbf{S}$ и $\mathbf{S}^T$ на вектор $\mathbf{p}$ нет необходимости вычислять $\mathbf{S}$ явно. Из структуры матрицы $\mathbf{S}$ следует, что наиболее сложной вычислительной процедурой является умножение матрицы $\mathbf{A}_{11}^{-1}$ на некоторый вектор $\mathbf{q}$. Результат умножения можно рассматривать как решение прямым методом системы $\mathbf{A}_{11}\mathbf{z}=\mathbf{q}$. Матрица $\mathbf{A}_{11}$ – ленточная, с шириной полосы $2r-1$. При малых r такая процедура не будет обременительной по числу операций и не потребует дополнительной памяти.

Наиболее важным для ускорения сходимости БСГ-метода (как и других предобусловленных методов) является построение предобусловливателя для матрицы $\mathbf{S}$. Рассмотрим этот вопрос более подробно.

Будем предполагать, что $\mathbf{A}$ является M -матрицей с диагональным преобладанием. Матрица $\mathbf{A}_{11}$ – блочно-диагональная. Рассмотрим произвольный блок $\mathbf{F}$ (см. замечание к (5)), который в свою очередь также имеет блочную структуру и состоит из m_y^2 блоков размеров $l \times l$, где $l = r - 1$. Запишем i -ю строку блока $\mathbf{F}$:

$$(0, \dots, 0, g_{i-l} \delta_{i-l, j_1}, 0, \dots, 0, f_{i-l} \delta_{i-l, j_2}, a_i \delta_{i, j}, d_i \delta_{i+1, j_3}, 0, \dots, 0, c_i \delta_{i+l, j_4}, 0, \dots, 0),$$

где δ_{ij} – символ Кронекера, второй индекс которого указывает на номер столбца матрицы.

Для построения аппроксимации матрицы $\mathbf{F}$ воспользуемся ее неполной факторизацией в виде $\mathbf{LVU}$, где матрицы $\mathbf{L}$ и $\mathbf{U}$ – нижняя и верхняя треугольные, матрица $\mathbf{V}$ – диагональная. Известно [36], что выражение $\mathbf{LVU}$ является хорошим приближением матрицы $\mathbf{F}$, если

$$\mathbf{V} = \text{diag}\{\tilde{a}_1, \tilde{a}_2, \dots, \tilde{a}_s\}, \quad \mathbf{L} = \mathbf{G} + \mathbf{H} + \mathbf{I}, \quad \mathbf{U} = \mathbf{I} + \mathbf{D} + \mathbf{C}.$$

Здесь $s = lm_y$, $\mathbf{I}$ – единичная матрица, $\mathbf{G} = \|g_{i-l} \delta_{i-l, j} / \hat{a}_{i-l}\|$, $\mathbf{H} = \|\hat{f}_{i-l} \delta_{i-l, j} / \hat{a}_{i-l}\|$,

$$\mathbf{D} = \|\hat{d}_i \delta_{i+1, j} / \hat{a}_i\|, \quad \mathbf{C} = \|c_i \delta_{i+l, j} / \hat{a}_i\|, \quad \hat{a}_i = a_i - \hat{f}_{i-l} \hat{d}_{i-l} / \hat{a}_{i-l} - c_{i-l} g_{i-l} / \hat{a}_{i-l},$$

$\hat{d}_i = d_i - \hat{f}_{i-l} c_{i-l} / \hat{a}_{i-l} \omega(l)$, $\hat{f}_i = f_i - g_{i-l} \hat{d}_{i-l} / \hat{a}_{i-l} \omega(l)$. В последних формулах $\omega(l) = 1$ при $l = 2$ и $\omega(l) = 0$ при $l > 2$.

Изучим некоторые свойства матрицы $\mathbf{L}^{-1}$, в частности, найдем оценку ее элементов.

Обозначим через q максимум модуля недиагонального элемента матрицы $\mathbf{L}$ и построим матрицу $\mathbf{Q}$, которая получается из матрицы $\mathbf{L}$ заменой внедиагональных элементов на $-q$. Тогда можно доказать следующую лемму.

Лемма. Имеет место неравенство

$$0 \leq \mathbf{L}^{-1} \leq \mathbf{Q}^{-1}. \quad (22)$$

Матрица $\mathbf{Q}$ имеет блочный нижний треугольный вид:

$$\mathbf{Q}_{ik} = \begin{cases} \mathbf{I} - q\mathbf{I}_1, & \text{если } i = k, \\ q\mathbf{I}, & \text{если } i = k + 1, \\ \mathbf{0}, & \text{если } i > k + 1 \text{ или } i < k, \end{cases}$$

где матрица $\mathbf{I}_1$ имеет первую поддиагональ, состоящую из единиц, а все остальные ее элементы равны нулю. Нетрудно убедиться, что матрица $\mathbf{Q}^{-1}$ будет иметь также блочный нижний треугольный вид, блоки $(\mathbf{Q}^{-1})_{ik}$ которого имеют размеры $l \times l$, где $k = \overline{1, i}$, $i = \overline{1, m_y}$. Структура этих блоков – нижняя треугольная. Для того чтобы получить представление для $(\mathbf{Q}^{-1})_{ik}$, введем числа Z_s^p . Для всех натуральных

ных p положим $Z_0^p = 1$. Для $s > 0$ числа Z_s^p определяются при помощи рекуррентных соотношений $Z_s^p = \sum_{k=1}^p Z_{s-1}^k$. Тогда справедлива следующая

Теорема 3. Ненулевые блоки матрицы Q^{-1} имеют вид

$$(Q^{-1})_{ik} = q^{k-i} \left(I + \sum_{p=1}^{r-2} Z_{i-k}^p q^p I_1^p \right), \quad k = \overline{1, i}, \quad i = \overline{1, m_y}.$$

Следствие. Если $r = 3$, то ненулевые блоки матрицы Q^{-1} представимы в виде

$$(Q^{-1})_{ik} = q^{k-i} (I + (i-k+1)qI_1), \quad k = \overline{1, i}, \quad i = \overline{1, m_y}.$$

Так как структура матрицы U совпадает со структурой матрицы L^T , то для блоков матрицы U^{-1} также имеет место подобная оценка, вытекающая из неравенства (22) и теоремы 3. Этим обстоятельством будем пользоваться при построении предобусловливателя H для матрицы S . Оставим в матрицах L^{-1} и U^{-1} наиболее существенные диагонали, их расположение будет зависеть от q , r и m_y . В частности, при $r = 3$ для L^{-1} из леммы и следствия имеем, что наиболее существенными являются поддиагонали, ближайшие к главным диагоналям, если $iq^{i-1} < 1$, $i = \overline{2, m_y}$.

Тогда на основании [36] и этого замечания в матрицах L^{-1} и U^{-1} можно оставить только три диагонали. Такое представление для аппроксимации матрицы A_{11} будет использовано в реализации соответствующего алгоритма.

Таким образом, доказанная теорема дает возможность поиска наиболее существенных диагоналей при построении приближения A_{11}^{-1} .

4. ТРЕХУРОВНЕВОЕ ПРЕДОБУСЛОВЛИВАНИЕ

Пусть проведена редукция системы (3) на основе красно-черного упорядочения. Как отмечалось в разделе 1.2, матрица S системы (6) будет девятидиагональной с расстоянием m_x до дальней диагонали.

В данном разделе остановимся на построении НФПСГ-алгоритмов для решения СЛАУ вида (6). Построим метод, в основе которого лежит решение редуцированной системы. При этом на этапе итерационного уточнения будем использовать предобусловливатель $\tilde{R}$ для матрицы S .

Алгоритм 8 (РНФПСГ)

- 1) Расчет x_0 в соответствии с (14).
- 2) Формирование S , $\tilde{b}$, R и $\tilde{R}$.
- 3) Построение предобусловливателя H для S .
- 4) Расчет начального приближения $\tilde{y}$ в соответствии с (14), примененной к системе (6).

5) Вычисление начального направления: $\mathbf{r}_0 = \tilde{\mathbf{b}} - \mathbf{S}\mathbf{y}$, $\tilde{\mathbf{p}}_0 = \tilde{\mathbf{R}}\mathbf{H}^{-1}\mathbf{r}_0$.

6) Итерационное уточнение ($i=0,1,2,\dots$):

$$6.1) \alpha_i = \frac{\mathbf{r}_i^T \tilde{\mathbf{R}}\mathbf{H}^{-1}\mathbf{r}_i}{\tilde{\mathbf{p}}_i^T \mathbf{S}\tilde{\mathbf{p}}_i};$$

$$6.2) \tilde{\mathbf{x}}_{i+1} = \tilde{\mathbf{x}}_i + \alpha_i \tilde{\mathbf{p}}_i;$$

$$6.3) \mathbf{r}_{i+1} = \mathbf{r}_i - \alpha_i \mathbf{S}\tilde{\mathbf{p}}_i;$$

$$6.4) \beta_i = \frac{\mathbf{r}_{i+1}^T \tilde{\mathbf{R}}\mathbf{H}^{-1}\mathbf{r}_{i+1}}{\mathbf{r}_i^T \tilde{\mathbf{R}}\mathbf{H}^{-1}\mathbf{r}_i};$$

$$6.5) \mathbf{p}_{i+1} = \tilde{\mathbf{R}}\mathbf{H}^{-1}\mathbf{r}_{i+1} + \beta_i \tilde{\mathbf{p}}_i.$$

7) Расчет $\mathbf{z} = \mathbf{R}\mathbf{T}\tilde{\mathbf{x}}$.

8) Получение решения системы (3) $\mathbf{x} = \mathbf{x}_0 + \mathbf{z}$.

Матрица $\mathbf{R}$ редуцирует систему (3) к виду (6), матрица $\mathbf{T}$ трансформирует решение системы (6) на основную сетку [13]. Проектор $\tilde{\mathbf{R}}$ строится на основе r -линейного упорядочения на грубой сетке. Рассуждая, как в разделе 2.1, можно представить его в виде

$$\tilde{\mathbf{R}} = \begin{bmatrix} \mathbf{0} & \mathbf{S}_{11}^{-1}\mathbf{S}_{12} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}.$$

Нетрудно установить, что в процессе решения все компоненты векторов $\mathbf{r}$ и $\mathbf{S}\mathbf{p}$, соответствующие исключаемым узлам на грубой сетке, являются нулевыми. Поэтому для хранения массивов освобождается дополнительная память, которая может быть использована для хранения матрицы $\tilde{\mathbf{R}}$. Заметим также, что при расчете скалярных произведений в описанном алгоритме достаточно перемножить лишь $1/r$ компонент соответствующих векторов, поскольку остальные элементы (по крайней мере у одного из сомножителей) будут равны нулю.

В случае использования двухлинейного упорядочения на грубой сетке для построения предобуславливателя $\tilde{\mathbf{R}}$ суммарные затраты на итерации составляют $10,75t_y + 9t_{cl}$.

Можно построить трехуровневую процедуру, в которой предобуславливатель $\mathbf{H}$ формируется к дополнению Шура $\tilde{\mathbf{S}} = \mathbf{S}_{22} - \mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}$ матрицы $\mathbf{S}_{11}$. Повторяя рассуждения п. 3, можно получить алгоритм, аналогичный алгоритму 7.

Алгоритм 9

1) Решить систему $\mathbf{A}_{11}\mathbf{v}_1 = \mathbf{b}_1$.

$$2) \mathbf{v}_2 = \begin{bmatrix} \mathbf{v}_{21} \\ \mathbf{v}_{22} \end{bmatrix} \leftarrow \mathbf{b}_2 - \mathbf{A}_{21}\mathbf{v}_1.$$

$$3) \text{ Решить систему } \mathbf{S}_{11}\mathbf{u}_1 = \mathbf{v}_{21}.$$

$$4) \mathbf{y} \leftarrow \mathbf{v}_{22} - \mathbf{S}_{21}\mathbf{u}_1.$$

$$5) \text{ Решить систему } \tilde{\mathbf{S}}\mathbf{x}_{22} = \mathbf{y}.$$

$$6) \mathbf{z}_2 \leftarrow \mathbf{S}_{12}\mathbf{x}_{22}.$$

$$7) \text{ Решить систему } \mathbf{S}_{11}\mathbf{z}_1 = \mathbf{z}_2.$$

$$8) \mathbf{x}_{21} \leftarrow \mathbf{u}_1 - \mathbf{z}_1.$$

$$9) \text{ Сформировать } \mathbf{x}_2 = \begin{bmatrix} \mathbf{x}_{21} \\ \mathbf{x}_{22} \end{bmatrix}_1.$$

$$10) \mathbf{w} \leftarrow \mathbf{A}_{12}\mathbf{x}_2.$$

$$11) \text{ Решить систему } \mathbf{A}_{11}\mathbf{u}_2 = \mathbf{w}.$$

$$12) \mathbf{x}_1 \leftarrow \mathbf{v}_1 - \mathbf{u}_2.$$

Как и в алгоритме 7, СЛАУ из пп. 1, 3, 7 и 11 решаются прямым методом, а из п. 7 – одним из предобусловленных итерационных методов, рассмотренных в разделах 2.1 или 2.2. Проблема построения предобусловливателя $\mathbf{H}$ решается так же, как и в разделе 3. Отметим, что в случае двухлинейного упорядочения на грубой сетке, матрица $\mathbf{S}_{11}$ трехдиагональная. Построение ее приближенного обращения хорошо изучено [3].

5. ЧИСЛЕННОЕ ИССЛЕДОВАНИЕ

Рассмотрим задачу численного моделирования процессов переноса заряда в элементах кремниевых СБИС. Дрейфово-диффузионная модель предполагает решение следующего набора ДУЧП [40]:

$$\nabla \cdot (\varepsilon \nabla \varphi) = -\rho, \quad (23)$$

$$\frac{\partial n}{\partial t} = \frac{1}{q} \nabla j_n - R, \quad (24)$$

$$\frac{\partial p}{\partial t} = -\frac{1}{q} \nabla j_p - R, \quad (25)$$

где $j_n = -q\mu_n(n\nabla\varphi - kT\nabla n)$, $j_p = -q\mu_p(p\nabla\varphi + kT\nabla p)$,

$$\rho = \begin{cases} q(p - n - C) & - \text{ в объеме полупроводника,} \\ Q_{ss} & - \text{ на границе Si/SiO}_2. \end{cases}$$

Обозначения: ε – относительная диэлектрическая проницаемость; φ – электростатический потенциал; ρ – заряд; n, p – концентрации электронов и дырок; t – время;

j_n, j_p – плотности электронного и дырочного токов; R – превышение скорости рекомбинации носителей заряда над скоростью генерации; q – заряд электрона; μ_n, μ_p – подвижности электронов и дырок; k – постоянная Больцмана; T – абсолютная температура; C – результирующая концентрация легирующей примеси; Q_{ss} – связанный заряд на поверхности Si/SiO. Индексы: кэ – коллектор-эмиттер; бэ – база-эмиттер.

В рамках данной модели учитываются сужение ширины запрещенной зоны, рекомбинационные механизмы Шокли–Рида и Оже, генерация вследствие лавинного умножения в сильном поле, фотовозбуждения и туннелирования. Модель подвижности учитывает рассеяния на ионизированных примесях, акустических фононах, электронно-дырочное рассеяние, поверхностное рассеяние, а также эффект насыщения дрейфовой скорости [41, 45].

Одним из наиболее распространенных подходов к решению нелинейной системы (23)–(25) является последовательное итерирование до получения самосогласованного решения [41, 45]. При этом на каждой внешней итерации необходимо последовательно решить три СЛАУ. При использовании непрерывных прямоугольных сеток для пространственной дискретизации задачи и пятиточечных разностных схем матрицы коэффициентов указанных СЛАУ имеют пятидиагональную структуру. Матрица алгебраизованного уравнения Пуассона является симметричной положительно определенной. Вследствие сильного диагонального преобладания для решения соответствующей линейной системы, как правило, достаточно небольшого числа итераций НФСГ-метода. При использовании концентраций носителей заряда в качестве переменных матрицы линейных систем алгебраизованных уравнений непрерывности (24), (25) являются несимметрическими. Подобные СЛАУ представляют серьезные трудности для численного решения [30]. Для этого здесь будут применяться все описанные в настоящей работе алгоритмы типа БСГ. Можно сделать в уравнениях (24), (25) такую замену переменных, что СЛАУ будут иметь симметричную матрицу, и, следовательно, для их решения можно применять СГ-алгоритмы. Существенным недостатком такого подхода является то, что в ряде важных для приложений случаев элементы матрицы могут оказаться очень большими и выйти за пределы разрядной сетки ЭВМ.

В качестве тестового примера рассмотрим субмикронный биполярный транзистор с глубиной эмиттерного перехода 0,1 мкм. Для получения самосогласованного решения был использован последовательный квазигуммелевский алгоритм КГ1 [41] (критерием достижения заданной точности служило условие $\delta < 10^{-4}$, где δ – максимальная по абсолютной величине компонента относительного изменения электростатического потенциала). По мере уменьшения δ точность решения СЛАУ возрастала в соответствии с [23].

Опишем условия проведения вычислительных экспериментов. Все расчеты проводились на IBM Pentium 166.

Основные вычислительные характеристики – количество итераций на решение и затраченное при этом время, измеренные на задаче (23)–(25), даны в табл. 3. В секции 1 приведены результаты моделирования транзистора при $V_{бэ} = 0,5$ В, $V_{кэ} = 1$ В. Сетка пространственной дискретизации содержала в этом эксперименте $24 \times 25 = 600$ узлов. Линейные системы решались одним из СГ-методов (ϵ – относительная погрешность

решения СЛАУ). В данной секции N – суммарное число итераций, понадобившееся для решения трех СЛАУ на всех квазиуммелевских итерациях, t – общее время, Δt – оценка преимущества рассматриваемых алгоритмов над традиционным НФСГ-алгоритмом. В секции 2 приведено также суммарное число итераций N для каждого из исследуемых методов, общее время t решения всех СЛАУ, время τ решения полной задачи моделирования субмикронного транзистора в режиме $V_{бз} = 0,85$ В, $V_{кз} = 5$. Сетка пространственной дискретизации содержала в этом случае $51 \times 63 = 3213$ узлов. В секции 3 проведено сравнение СГ-алгоритмов с двухуровневыми и трехуровневыми предобусловливателями. Сетка пространственной дискретизации содержала в этой группе экспериментов $41 \times 63 = 2457$ узлов. Моделирование проводилось в режиме $V_{бз} = 0,5$ В, $V_{кз} = 1$ В. В столбцах с третьего по шестой приводится число итерации N и процессорное время t , требуемое для решения СЛАУ алгебраизованных уравнений (24) и (25) на первой квазиуммелевской итерации. В последних двух столбцах дано число итерации N , потребовавшееся для решения СЛАУ полной задачи (23)–(25) на всех квазиуммелевских итерациях, и общее время t на их выполнение.

Таблица 3

Результаты вычислительных экспериментов

Секция 1									
Упорядочения	$\varepsilon=10^{-6}$			$\varepsilon=10^{-9}$					
	N	t, с	Δt , %	N	t, с	Δt , %			
–	194	1,60		294	2,43				
Красно-черное	190	1,38	14	265	1,87	23			
Двухлинейное	132	0,98	39	202	1,43	41			
Трехлинейное	117	0,83	48	172	1,19	51			
Четырехлинейное	115	0,85	49	173	1,23	49			
Секция 2									
Метод	BCI*			CGS			CGSTAB		
Упорядочения	N	t, с	τ , с	N	t, с	τ , с	N	t, с	τ , с
–	6281	147,83	174,64	4947	114,71	145,48	4300	127,43	158,5
Двухлинейное	5065	100,04	130,82	4195	82,91	113,74	3876	93,09	123,6
Трехлинейное	4746	88,95	119,77	3891	75,08	105,84	3604	83,45	114,0
Секция 3									
Предобус- ловливание	Упорядочения	(24)		(25)		(23)–(25)			
		N	t, с	N	t, с	N	t, с		
Двухуровневое	Редукция	60	0,50	41	0,33	571	4,67		
Трехуровневое	Двухлинейное	57	0,39	41	0,28	551	3,62		
ДвухуровневоеНФ для A	Трехлинейное	66	0,55	42	0,33	615	4,88		
ДвухуровневоеНФ для S	Трехлинейное	55	0,44	38	0,28	529	4,31		

Анализ табл. 3 позволяет сделать следующие выводы:

1. При решении СЛАУ с симметричной матрицей наиболее эффективным оказалось трехлинейное упорядочение. Четырехлинейное упорядочение несколько сократило общее число итераций, однако возросшие затраты на итерации не позволили повысить эффективность метода.

2. При решении СЛАУ с несимметричной матрицей CGS-метод потребовал меньших временных затрат на решение задачи, чем два других метода (БСГ и CGSTAB). Этот вывод сохраняет силу и при использовании в них r -линейных упорядочений. Как и следовало ожидать, трехлинейное упорядочение и в этом случае оказалось наиболее эффективным.

3. Трехуровневая предобуславливающая процедура с двухлинейным упорядочением потребовала наименьших временных затрат при решении всех трех СЛАУ дрейфово-диффузионной модели.

4. Предложенные многоуровневые методы решения СЛАУ исследованы на задаче численного моделирования микроэлектронных структур. Поскольку СЛАУ с пятидиагональной матрицей возникают в различных областях инженерных и физических приложений, то предложенные методы имеют более широкий круг применений.

ЛИТЕРАТУРА

1. Форсайт Д., Молер К. Численное решение систем линейных алгебраических уравнений. М., 1960.
2. Хейгеман А., Янг Д. Прикладные итерационные методы. М., 1968.
3. Axelsson O., Gustafsson I. Iterative Solution Methods. Cambridge, 1996.
4. Greenbaum A. // Numer. Math. 1979. V. 33. P. 181.
5. Булеев Н. И. // Мат. сб. 1960. Т. 51, № 2. С. 227.
6. Elman H. C., Golub G. H. // Math. and Comp. 1991. V. 54, № 190. P. 671.
7. Mandel J. // Numer. Math. 1991. V. 58, № 1. P. 79.
8. Nicolaidis R. A. // SIAM Numer. Anal. 1987. V. 24, № 2. P. 355.
9. Федоренко Р. П. // Журн. выч. мат. и мат. физ. 1961. Т. 1, № 5. С. 922.
10. Hackbusch W. Multi-Grid Methods and Applications. Berlin, 1985.
11. Golub G. H., van Loan C. F. Matrix Computations. Baltimore, 1996.
12. Mulyarchik S. G., Bielawski S. S., Popov A. V. // J. Comput. Phys. 1994. V. 110. № 2. P. 201.
13. Bielawski S. S., Mulyarchik S. G., Popov A. V. // J. Comput. Appl. Math. 1996. V. 70, № 2. P. 189.
14. Brand C. W. // Numer. Math. 1992. V. 61, № 3. P. 433.
15. Saad Y., Zhang I. // SIAM J. Matrix Anal. and Appl. 1999. V. 21, № 7. P. 279.
16. Axelsson O., Gustafsson I. // J. Comput. Phys. 1980. V. 35, № 2. P. 284.
17. Axelsson O., Eijkhout V. // SIAM J. Sci. Statist. Comput. 1991. V. 12. P. 1373.
18. Elman H. C., Golub G. H. // SIAM J. Sci. Statist. Comput. 1992. V. 13, № 1. P. 339.
19. Rodrigue G., Wolitzer I. // Math. Comp. 1984. V. 42, № 166. P. 549.
20. Elman H. C. // SIAM J. Sci. Statist. Comput. 1989. V. 10, № 3. P. 581.
21. Hestenes M.R., Stiefel E. // NBS J. Res. 1952. V. 49, № 6. P. 409.
22. Пшеничный Б. П., Данилин Ю. М. Численные методы в экстремальных задачах. М., 1975.
23. Белявский С. С., Мулярчик С. Г., Попов А. В. // Дифференц. уравнения. 1993. Т. 29, № 9. С. 1575.
24. Белявский С. С., Мулярчик С. Г., Попов А. В. // Дифференц. уравнения. 1994. Т. 30, № 12. С. 2106.
25. Fletcher R. // Leck. Notes Math. 1976. № 506. P. 73.
26. Lanczos C. // J. Res. Bur. Stand. V. 49, № 195. P. 33.
27. Sonneveld P. // SIAM J. SCI. State. Compute. 1989. V. 10, № 1. P. 36.
28. Van der Vorst H. A. // SIAM J. Sci. Stat. Comput. 1992. V. 13, № 2. P. 631.