УДК 519.632.4, 519.624.2

# ОБ ОЦЕНКЕ ПОГРЕШНОСТИ ПРИБЛИЖЕННОГО РЕШЕНИЯ ЭЛЛИПТИЧЕСКИХ УРАВНЕНИЙ С НЕКОЭРЦИТИВНОЙ БИЛИНЕЙНОЙ ФОРМОЙ

**А. Н. Боголюбов**[1], **А. А. Панин**[1]

Предложен алгоритм оценки погрешности приближенного решения эллиптического уравнения, основанный на методе Накао и пригодный также и в случае, когда билинейная форма задачи некоэрцитивна. Для уравнения Гельмгольца на основе метода Накао разработан метод, позволяющий получить более тонкую оценку. Приведены результаты тестовых расчетов оценок погрешности двумя методами.

**Ключевые слова:** эллиптические уравнения, проекционные методы, метод конечных элементов, оценка погрешности.

**1. Введение.** В последние годы широкое распространение получил проекционно-сеточный метод, называемый также методом конечных элементов (МКЭ). Для приближенных решений этим методом эллиптических краевых задач с положительно определенными операторами известна оценка погрешности $\|u-v\|_{H^1} \leqslant Ch\|f\|$, $\|u-v\| \leqslant Ch^2\|f\|$, где $u$ — точное решение, $v$ — приближенное решение, $f$ — правая часть, $C$ — некоторая константа и $\|\cdot\|$ — $L^2$-норма в рассматриваемой области [1]. Несколько другой результат предлагает лемма Сеа [2]: погрешность приближенного решения превосходит расстояние от точного решения до конечно-элементного подпространства не более чем в $C = M_0/A_0$ раз, а для задач с симметричной билинейной формой — не более чем в $C = \sqrt{M_0/A_0}$ раз: $\|u-v\|_V \leqslant C \inf\limits_{w\in V_h} \|u-w\|_V$, где $V$ — пространство, в котором рассматривается задача, а $M_0$ и $A_0$ — соответственно норма и константа эллиптичности билинейной формы задачи относительно нормы $\|\cdot\|_V$. Следует отметить также работы по апостериорному подходу к оценке погрешности (см., например, [3–5]).

Представляет интерес метод Накао и его коллег [6–8]. Достоинством метода является его применимость к задачам с некоэрцитивной билинейной формой (т.е. с незнакоопределенным оператором). В настоящей статье на его основе разработан метод, применимый к уравнению Гельмгольца и дающий для него более тонкую оценку.

**2. Дифференциальная задача и предварительные сведения.** Рассмотрим в выпуклой ограниченной области $\Omega \subset \mathbb{R}^n$ с кусочно-гладкой границей однородную краевую задачу Дирихле

$$\mathcal{L}u \equiv -\triangle u + b(x)\cdot\nabla u + c(x)u = f, \quad u\big|_{\partial\Omega} = 0, \tag{1}$$

где точкой обозначено скалярное произведение векторов. Будем считать, что $f \in L^2(\Omega)$, $b \in W^1_\infty(\Omega)$ и $c \in L^\infty(\Omega)$. Во многих интересных на практике случаях исследовать вопрос об обратимости оператора $\mathcal{L}$ можно в процессе нахождения численного решения (см. теоремы 1 и 1а).

Обобщенная постановка задачи состоит в поиске решения $u \in H^1_0(\Omega)$, удовлетворяющего условию

$$a(u,w) \equiv \int\limits_\Omega \Big(\nabla u\cdot\nabla w + \big(b(x)\cdot\nabla u\big)w + c(x)uw\Big)\,dx = \int\limits_\Omega f(x)w\,dx \quad \forall w \in H^1_0(\Omega), \tag{2}$$

где оператор $\nabla$ здесь и ниже действует из $H^1(\Omega)$ в $L^2(\Omega)$.

В качестве нормы в пространстве $H^1_0(\Omega)$ возьмем $\|\nabla u\|_{L^2}$, что возможно в силу известной эквивалентности норм. Если индекс $L^2$ опущен, то предполагается норма в $L^2(\Omega)$. В дальнейшем вместо $H^1_0(\Omega)$ часто будем писать $H^1_0$.

**3. Дискретизация задачи.** В пространстве $H^1_0(\Omega)$ рассмотрим семейство конечномерных подпространств $S^0_N$, натянутых на линейно независимые функции $\{\varphi_i\}_{i=1}^N$. Для всякого $v \in H^1_0(\Omega)$ определим

[1] Московский государственный университет им. М.В. Ломоносова, физический факультет, кафедра математики, Ленинские горы, 119992, Москва; А.Н. Боголюбов, профессор, e-mail: bogan7@yandex.ru; А.А. Панин, аспирант, e-mail: a-panin@yandex.ru

$H_0^1$-проекцию $P_N v \in S_N^0$ условием

$$\bigl(\nabla(v - P_N v), \nabla\varphi_{\text{discr}}\bigr)_{L^2} = 0 \quad \forall\varphi_{\text{discr}} \in S_N^0. \tag{3}$$

Существование такой проекции гарантируется теоремой Рисса.

Введем множество $X(\Omega)$, состоящее из тех элементов $v \in H^1(\Omega)$, для которых существует такая функция $f_v \in L^2(\Omega)$, что верно равенство

$$(\nabla v, \nabla w) = -(f_v, w) \quad \forall w \in H_0^1(\Omega). \tag{4}$$

Тогда на $X(\Omega) \cap H_0^1(\Omega)$ определим оператор $\triangle : X(\Omega) \cap H_0^1(\Omega) \to L^2(\Omega)$, сопоставляющий каждой функции $v \in X(\Omega) \cap H_0^1(\Omega)$ функцию $f_v \in L^2(\Omega)$ по формуле (4).

Для характеристики качества приближения функций $v \in X(\Omega) \cap H_0^1(\Omega)$ будем использовать константу $C(N)$, а именно потребуем, чтобы для всякого $v \in X(\Omega) \cap H_0^1(\Omega)$ выполнялось неравенство

$$\|v - P_N v\|_{H_0^1} \leqslant C(N)\|\triangle v\|_{L^2}. \tag{5}$$

Это основное и единственное требование, которому должны удовлетворять конечномерные подпространства $S_N^0$. Из (5) с помощью леммы Обэна–Нитше ([2], с. 139) заключаем, что верно неравенство (с той же самой константой $C(N)$)

$$\|v - P_N v\|_{L^2} \leqslant C(N)\|v - P_N v\|_{H_0^1}. \tag{6}$$

Указанные оценки верны, в частности, для некоторых пространств конечных элементов. Так, для кусочно-линейных элементов имеем $C(N) = h/\pi$, где $h$ — шаг сетки [9]; то же верно и для кусочно-билинейных элементов на квадратной сетке в прямоугольной области [9]. В случае триангуляции многоугольной области в $\mathbb{R}^2$ можно воспользоваться теоремой, доказанной Ф. Наттерером.

**Теорема** [10]. *Пусть в треугольнике $\Delta$ со сторонами $l_1$, $l_2$ и углом между ними $0 < \omega < \pi$ задана функция $u \in H^2$, а $P_h' u$ — ее линейная интерполянта, т.е. функция вида $Ax + By + C$, совпадающая с функцией $u$ в вершинах треугольника. Пусть $h^2 = (l_1^2 + l_2^2)/2$, $|\Delta| = (l_1 l_2 \sin\omega)/2$ — площадь треугольника $\Delta$ и $d = 2|\Delta|/h^2$. Тогда верно неравенство*

$$\bigl\|\nabla(u - P_h' u)\bigr\|_{L^2(\Delta)} \leqslant hc(\Delta)|u|_{H^2(\Delta)}, \quad c(\Delta) = \overline{c}\,\frac{1 + \sqrt{1 - d^2}}{\sqrt{1 - \sqrt{1 - d^2}}}, \quad 0{,}46 \leqslant \overline{c} \leqslant 0{,}81.$$

Для триангуляции $\mathcal{T}$ многоугольной области можно положить $hc(\Delta)$ равным $\max\limits_{\Delta\in\mathcal{T}} hc(\Delta)$. Если область является выпуклой, то решение из $H^1$ принадлежит пространству $H^2$ и верно неравенство $|u|_{H^2} \leqslant \|\triangle u\|$ [11]. В итоге получаем (5) и (6) с $C(N) = \max\limits_{\Delta\in\mathcal{T}} hc(\Delta)$, потому что данная оценка, верная для линейной интерполяции, тем более верна для проекции $P_N u$.

Определим $(N \times N)$-матрицы: $G = \{G_{ji}\} = \bigl\{(\nabla\varphi_i, \nabla\varphi_j)_{L^2} + (b \cdot \nabla\varphi_i, \varphi_j)_{L^2} + (c\varphi_i, \varphi_j)_{L^2}\bigr\} \equiv \{a(\varphi_i, \varphi_j)\}$ и $D = \{D_{ji}\} = \bigl\{(\nabla\varphi_i, \nabla\varphi_j)_{L^2}\bigr\}$. Матрица $D$ симметрична и положительно определена (следовательно, невырожденна). Если матрица $G$ также невырожденна, то назовем приближенным решением задачи (1) по методу Галеркина в подпространстве $S_N^0$ функцию $P_{\mathcal{L}} u$, определяемую условием

$$a(P_{\mathcal{L}} u, \varphi_{\text{discr}}) = (f, \varphi_{\text{discr}}) \quad \forall\varphi_{\text{discr}} \in S_N^0. \tag{7}$$

Существование такой функции гарантируется невырожденностью матрицы $G$, что очевидно, если записать функцию $P_{\mathcal{L}} u$ в виде $P_{\mathcal{L}} u = \sum\limits_{j=1}^{N} u_j^{(N)} \varphi_j$, а условие (7) — в виде $\sum\limits_{j=1}^{N} a(\varphi_j, \varphi_i) u_j^{(N)} = (f, \varphi_i)$, $i = 1, \ldots, N$.

Аналогичным образом для произвольной функции $v \in H_0^1$ можно определить ее (неортогональную, вообще говоря) проекцию $P_{\mathcal{L}} v$ на $S_N^0$ с помощью соотношения

$$a(P_{\mathcal{L}} v, \varphi_{\text{discr}}) = a(v, \varphi_{\text{discr}}) \quad \forall\varphi_{\text{discr}} \in S_N^0. \tag{8}$$

Существование проекции устанавливается аналогично случаю (7). Итак, при условии невырожденности матрицы $G$ мы определили проектор $P_{\mathcal{L}} : H_0^1 \to S_N^0$.

Рассмотрим спектральную матричную норму $\|\cdot\|_2$, подчиненную евклидовой векторной норме: $\|A\|_2 := \sup\limits_{\|x\|_2\neq 0} \dfrac{\|Ax\|_2}{\|x\|_2}$. Будем далее пользоваться обозначениями, введенными в [7]:

$$\begin{aligned} &C_{\operatorname{div} b} = \big\|\operatorname{div} b(x)\big\|_{L^\infty}, \quad C_b = \big\|\,\|b(x)\|_2\,\big\|_{L^\infty}, \quad C_c = \|c(x)\|_{L^\infty}, \\ &C_1 = C_p C_{\operatorname{div} b} + C_b, \quad C_2 = C_p C_c, \quad C_3 = C_b + C_p C_c, \quad C_4 = C_b + C(N) C_c. \end{aligned} \tag{9}$$

Здесь $C_p$ — константа в неравенстве Пуанкаре–Фридрихса $\|w\|_{L^2} \leqslant C_p \|w\|_{H^1_0}\ \forall w \in H^1_0(\Omega)$.

Введем еще константы (аргумент $N$ часто будет лишь подразумеваться)

$$M(N) = \|L^{\mathrm{T}} G^{-1} L\|_2, \ \ \kappa(N) = C(N)\big[C(N)M(N)(C_1+C_2)C_3 + C_4\big], \ \ \sigma(N) = \frac{1 + C_p M(N) C_3}{1 - \kappa(N)}, \tag{10}$$

где $D = LL^{\mathrm{T}}$ — разложение Холецкого матрицы $D$.

**4. Метод Накао.** В настоящем разделе дано компактное, но полное изложение доказательства метода, предложенного Накао, по статьям [6–8]. На его основе в данной работе сформулирован явный алгоритм вычисления приближенного решения и построения оценок его погрешности.

**Теорема 1.** *Если матрица $G$ обратима и $\kappa < 1$, то оператор $\mathcal{L}$ задачи (1) обратим.*

Для доказательства теоремы потребуется установить некоторые вспомогательные утверждения.

**4.1. Операторы $A$ и $[I-A]_N^{-1}$.** Введем оператор $A : H^1_0 \to H^1_0 \cap X(\Omega)$, положив для каждого $u \in H^1_0$ элемент $Au \in H^1_0$ равным обобщенному решению из $H^1_0$ задачи $\triangle(Au) = b\cdot\nabla u + cu$, $(Au)\big|_{\partial\Omega} = 0$:

$$\big(\nabla(Au), \nabla w\big) = -(b\cdot\nabla u + cu, w) \quad \forall w \in H^1_0. \tag{11}$$

Иногда будет использоваться запись вида $Au = \triangle^{-1}(b\cdot\nabla u + cu)$, где $\triangle^{-1} : L^2(\Omega) \to H^1_0(\Omega) \cap X(\Omega)$ понимается как оператор, сопоставляющий функции $f \in L^2(\Omega)$ решение $\psi \in H^1_0(\Omega)\cap X(\Omega)$ задачи Дирихле для оператора Лапласа в обобщенной постановке:

$$(\nabla\psi, \nabla w) = -(f, w) \quad \forall w \in H^1_0, \quad \psi \in H^1_0. \tag{12}$$

Такое определение корректно, потому что задача (12) имеет единственное решение [11]. Тогда задачу (2) можно переформулировать в виде

$$u = Au + v. \tag{13}$$

Действительно, пусть $v \in H^1_0(\Omega)$ — решение однородной задачи Дирихле для оператора Лапласа: $-\triangle v = f$, т.е. $(\nabla v, \nabla w) = (f, w)$ при любом $w \in H^1_0$. Тогда (2) равносильно задаче $\big(\nabla(v-u), \nabla w\big) = (b\cdot\nabla u + cu, w)\ \forall w \in H^1_0$, или, в соответствии с (11), $u - v = Au$. Оператор $A$ вполне непрерывен ([11], с. 96–98). Значит, к нему применима альтернатива Фредгольма, и для однозначной разрешимости уравнения $u = Au + v$ (а значит, и задачи (1)) достаточно доказать, что уравнение

$$u = Au \tag{14}$$

имеет только тривиальное решение.

Рассмотрим теперь оператор $(I - P_N A)\big|_{S^0_N}$ и найдем его матрицу в пространстве $S^0_N$ в базисе $\{\varphi_i\}_{i=1}^N$:

$$\begin{aligned} \big(\nabla(I-P_NA)\varphi_i, \nabla\varphi_j\big) &= (\nabla\varphi_i, \nabla\varphi_j) - (\nabla P_N A\varphi_i, \nabla\varphi_j) = \big\{\text{формула (3)}\big\} = \\ &= (\nabla\varphi_i, \nabla\varphi_j) - (\nabla A\varphi_i, \nabla\varphi_j) = \big\{\text{формула (11)}\big\} = \\ &= (\nabla\varphi_i, \nabla\varphi_j) + (b\cdot\nabla\varphi_i, \varphi_j) + (c\varphi_i, \varphi_j) = G_{ji}. \end{aligned}$$

Поскольку $\{\varphi_i\}_{i=1}^N$ не является ортогональным базисом в смысле скалярного произведения $(\cdot,\cdot)_{H^1_0}$, то матрица оператора $(I - P_N A)\big|_{S^0_N}$ имеет вид $D^{-1}G$, где $D$ — матрица Грама базисных векторов. Итак, если матрица $G$ невырожденна, то существует оператор $\Big((I - P_N A)\big|_{S^0_N}\Big)^{-1}$. Поскольку на $S^0_N$ проектор $P_N$ является тождественным оператором, то $\Big((I - P_N A)\big|_{S^0_N}\Big)^{-1} = \Big(P_N(I - A)\big|_{S^0_N}\Big)^{-1}$, чем мы в дальнейшем и будем пользоваться. Для сокращения записи обозначим введенный оператор символом $[I-A]_N^{-1}$. Квадратные скобки имеют особый смысл только в этом обозначении.

**4.2. Леммы.** Для удобства изложения сформулируем отдельно три леммы, которые будут в дальнейшем неоднократно использоваться. Обозначим предварительно $P_\perp := I - P_N$.

**Лемма 1.** *Пусть матрица $G$ обратима и $w \in H_0^1$. Тогда $\|P_N A w_\perp\|_{H_0^1} \leqslant C(N)(C_1 + C_2)\|w_\perp\|_{H_0^1}$, где в условии этой и всех последующих лемм $w_\perp \equiv P_\perp w$ и $w_N \equiv P_N w$.*

**Доказательство.** Ясно, что $P_N$ — ортопроектор в смысле скалярного произведения $(\cdot,\cdot)_{H_0^1}$. Действительно, $P_N^2 = P_N$ по определению. Далее, $P_N^* = P_N$, так как для всех $u, v \in H_0^1$ верно $(\nabla P_N u, \nabla v) = \overline{(\nabla v, \nabla P_N u)} = \overline{(\nabla P_N v, \nabla P_N u)} = (\nabla P_N u, \nabla P_N v) = (\nabla u, \nabla P_N v)$. Поскольку $P_\perp = I - P_N$ тоже ортопроектор, то при любом $v \in H_0^1$ имеем $\|P_N v\|_{H_0^1} \leqslant \|v\|_{H_0^1}$ и $\|P_\perp v\|_{H_0^1} \leqslant \|v\|_{H_0^1}$. Следовательно, $\|P_N A w_\perp\|_{H_0^1} \leqslant \|A w_\perp\|_{H_0^1}$. Вспомним еще, что $A w_\perp \equiv \triangle^{-1}(b \cdot \nabla w_\perp + c w_\perp)$ в смысле формулы (11), и оценим слагаемые в правой части отдельно, определив функции $\psi_1$ и $\psi_2$ как обобщенные решения однородных задач Дирихле для оператора Лапласа в области $\Omega$ с правыми частями соответственно $b \cdot \nabla w_\perp$ и $c w_\perp$. Получим

$$\begin{aligned}\|\psi_1\|_{H_0^1}^2 &\equiv (\nabla\psi_1, \nabla\psi_1) = (-b \cdot \nabla w_\perp, \psi_1) = (-\nabla w_\perp, b\psi_1) = \\ &= \big(w_\perp, \operatorname{div}(b\psi_1)\big) \leqslant \|w_\perp\|\,\|\operatorname{div}(b\psi_1)\| \leqslant C(N)\|w_\perp\|_{H_0^1}\|\psi_1 \operatorname{div} b + b \cdot \nabla\psi_1\| \leqslant \\ &\leqslant C(N)\|w_\perp\|_{H_0^1}\Big(\|\psi_1\|\,\|\operatorname{div} b\|_{L^\infty} + \|\psi_1\|_{H_0^1}\|\,\|b\|_2\,\|_{L^\infty}\Big) \leqslant C(N)\|w_\perp\|_{H_0^1}\|\psi_1\|_{H_0^1}(C_{\operatorname{div} b}C_p + C_b),\end{aligned} \tag{15}$$

где использована оценка $\|w_\perp\| \leqslant C(N)\|w_\perp\|_{H_0^1}$ (следствие оценки (6)) и неравенство Пуанкаре–Фридрихса. Далее, для $\psi_2$ имеем

$$\|\psi_2\|_{H_0^1}^2 = (\psi_2, -c w_\perp) \leqslant \|\psi_2\|C_c\|w_\perp\| \leqslant C_p\|\psi_2\|_{H_0^1}C_c C(N)\|w_\perp\|_{H_0^1}. \tag{16}$$

Теперь из (15), (16) и неравенства треугольника, примененного к $\psi_1 + \psi_2$, получим

$$\|A w_\perp\|_{H_0^1} \leqslant \|\psi_1 + \psi_2\|_{H_0^1} \leqslant C(N)\|w_\perp\|_{H_0^1}(C_1 + C_2),$$

где $C_1 \equiv C_{\operatorname{div} b}C_p + C_b$ и $C_2 \equiv C_c C_p$. Лемма доказана.

**Лемма 2.** *Пусть матрица $G$ обратима. Тогда при любом $w \in H_0^1$ верна оценка*

$$\big\|(I - P_N)Aw\big\|_{H_0^1} \leqslant C(N)\big(C_3\|w_N\|_{H_0^1} + C_4\|w_\perp\|_{H_0^1}\big).$$

**Доказательство.** Заметим прежде всего, что в силу предположения (5) выполнено

$$\big\|(I - P_N)A(w_N + w_\perp)\big\|_{H_0^1} \leqslant C(N)\big\|\triangle A(w_N + w_\perp)\big\| \equiv C(N)\big\|b \cdot \nabla(w_N + w_\perp) + c(w_N + w_\perp)\big\|,$$

где $A(w_N + w_\perp) \in X(\Omega)$ в силу определения оператора $A$. Далее оценим по отдельности слагаемые с $w_N$ и с $w_\perp$:

$$\begin{aligned}&\|b \cdot \nabla w_N + c w_N\| \leqslant C_b\|w_N\|_{H_0^1} + C_c\|w_N\| \leqslant C_b\|w_N\|_{H_0^1} + C_c C_p\|w_N\|_{H_0^1} = (C_b + C_c C_p)\|w_N\|_{H_0^1}, \\ &\|b \cdot \nabla w_\perp + c w_\perp\| \leqslant C_b\|w_\perp\|_{H_0^1} + C_c\|w_\perp\| \leqslant \big(C_b + C_c C(N)\big)\|w_\perp\|_{H_0^1}.\end{aligned}$$

Следовательно, имеем

$$\begin{aligned}\big\|(I - P_N)A(w_N + w_\perp)\big\|_{H_0^1} &\leqslant C(N)\Big[(C_b + C_c C_p)\|w_N\|_{H_0^1} + \big(C_b + C(N)C_c\big)\|w_\perp\|_{H_0^1}\Big] = \\ &= C(N)\big(C_3\|w_N\|_{H_0^1} + C_4\|w_\perp\|_{H_0^1}\big),\end{aligned}$$

где $C_3 \equiv C_b + C_c C_p$ и $C_4 \equiv C_b + C_c C(N)$. Лемма доказана.

**Лемма 3.** *Пусть матрица $G$ обратима. Тогда верна оценка $\big\|[I - A]_N^{-1}\big\|_{H_0^1} \leqslant M$, где оператор $[I - A]_N^{-1}$ определен в разделе 4.1.*

**Доказательство.** Поскольку (см. раздел 4.1) $D^{-1}G$ — матрица оператора $(I - P_N A)\big|_{S_N^0}$, то $G^{-1}D$ — матрица обратного к нему оператора $[I - A]_N^{-1}$. Поэтому, если для функций $\psi_N$ и $v_N$ из $S_N^0$ верно $\psi_N = [I - A]_N^{-1} v_N$, где $\psi_N \equiv \sum\limits_{i=1}^N \Psi_i\varphi_i$ и $v_N \equiv \sum\limits_{i=1}^N V_i\varphi_i$, то $\Psi = G^{-1}DV$, где $\Psi = (\Psi_1, \ldots, \Psi_N)^{\mathrm{T}}$ и $V = (V_1, \ldots, V_N)^{\mathrm{T}}$. Отсюда

$$\begin{aligned}\|\psi_N\|_{H_0^1}^2 &= \Psi^{\mathrm{T}} D\Psi = \Psi^{\mathrm{T}} D G^{-1} D V = (L^{\mathrm{T}}\Psi)^{\mathrm{T}}(L^{\mathrm{T}}G^{-1}L)(L^{\mathrm{T}}V) \leqslant \|L^{\mathrm{T}}\Psi\|_2\,\|L^{\mathrm{T}}G^{-1}L\|_2\,\|L^{\mathrm{T}}V\|_2 = \\ &= \big\{\forall w \in \mathbb{R}^N\ \|L^{\mathrm{T}}w\|_2^2 = (L^{\mathrm{T}}w, L^{\mathrm{T}}w) = w^{\mathrm{T}}LL^{\mathrm{T}}w = w^{\mathrm{T}}Dw\big\} = \|\psi_N\|_{H_0^1}\|L^{\mathrm{T}}G^{-1}L\|_2\|v_N\|_{H_0^1}.\end{aligned} \tag{17}$$

Здесь $D = LL^{\mathrm{T}}$ — разложение Холецкого матрицы $D$. Если мы теперь обозначим $\|L^{\mathrm{T}}G^{-1}L\|_2$ символом $M$, то из (17) будем иметь $\|\psi_N\|_{H_0^1} \leqslant M\|v_N\|_{H_0^1}$. Лемма доказана.

**4.3. Доказательство теоремы 1.** Представим уравнение (14) в виде равносильной ему системы $P_N u = P_N Au$, $(I - P_N)u = (I - P_N)Au$. Введем еще операторы

$$Q_N u \equiv P_N u - [I - A]_N^{-1} P_N (I - A)u, \quad Tu \equiv Q_N u + (I - P_N)Au. \tag{18}$$

Из $u = Au$ и первого из равенств (18) следует $Q_N u = P_N u$, а тогда из второго получим $u = Tu$, поэтому для доказательства однозначной разрешимости уравнения $u = Au$ достаточно показать однозначную разрешимость уравнения $u = Tu$. Сделаем это следующим образом.

Пусть $\zeta$ и $\gamma$ — произвольные положительные числа. Определим множества

$$U_N := \{u_N \in S_N^0 \mid \|u_N\|_{H_0^1} \leqslant \gamma\}, \quad U_\perp := \{u_\perp \in S_N^\perp \mid \|u_\perp\|_{H_0^1} \leqslant \zeta\}, \tag{19}$$

где $S_N^\perp := H_0^1 \ominus S_N^0$, и множество $U = U_N + U_\perp = \{u_N + u_\perp \mid u_N \in U_N, u_\perp \in U_\perp\}$. Предположим, что нашлось такое ненулевое $u^*$, что $u^* = Tu^*$. Тогда в силу линейности оператора $T$ для любого $\lambda \in \mathbb{R}$ верно $\lambda u^* = T(\lambda u^*)$. В частности, при действии оператора $T$ останутся на месте элементы $\partial U \cap \{\lambda u^* \mid \lambda \in \mathbb{R}\}$. Поскольку $\partial U = \{u \in U \mid \|u_N\|_{H_0^1} = \gamma$ или $\|u_\perp\|_{H_0^1} = \zeta\}$ и по определению (18) $T = Q_N + (I - P_N)A$, то ни один элемент множества $\partial U$ не останется неподвижен, если потребовать выполнения неравенств

$$\|Q_N U\|_{H_0^1} \equiv \sup_{u \in U} \|Q_N u\|_{H_0^1} < \gamma, \tag{20}$$

$$\big\|(I - P_N)AU\big\|_{H_0^1} \equiv \sup_{u \in U} \big\|(I - P_N)Au\big\|_{H_0^1} < \zeta \tag{21}$$

при некоторых положительных $\gamma$ и $\zeta$.

Оценим $\|Q_N U\|_{H_0^1}$ и $\big\|(I-P_N)AU\big\|_{H_0^1}$. Для любого $u = u_N + u_\perp \in U_N + U_\perp$ положим $\psi_N := Q_N(u_N + u_\perp)$ и получим

$$\begin{aligned}\psi_N &= u_N - [I - A]_N^{-1} P_N (I - A)(u_N + u_\perp) = u_N - [I - A]_N^{-1} P_N (I - A)u_N - [I - A]_N^{-1} P_N (I - A)u_\perp = \\ &= u_N - u_N - 0 + [I - A]_N^{-1} P_N Au_\perp = [I - A]_N^{-1} P_N Au_\perp \equiv [I - A]_N^{-1} v_N,\end{aligned} \tag{22}$$

где $v_N := P_N Au_\perp$. В силу леммы 3 имеем

$$\|\psi_N\|_{H_0^1} \leqslant M\|v_N\|_{H_0^1} = M\|P_N Au_\perp\|_{H_0^1} \leqslant \{\text{лемма 1}\} \leqslant MC(N)(C_1 + C_2)\|u_\perp\|_{H_0^1} \leqslant MC(N)(C_1 + C_2)\zeta,$$

где последнее неравенство следует из предположений (19). Таким образом,

$$\|Q_N U\|_{H_0^1} \leqslant MC(N)(C_1 + C_2)\|u_\perp\|_{H_0^1} \leqslant MC(N)(C_1 + C_2)\zeta. \tag{23}$$

Далее, в силу леммы 2 и предположений (19) имеем

$$\big\|(I - P_N)AU\big\|_{H_0^1} \leqslant C(N) \sup_{u \in U} \big(C_3\|u_N\|_{H_0^1} + C_4\|u_\perp\|_{H_0^1}\big) \leqslant C(N)(C_3\gamma + C_4\zeta). \tag{24}$$

Как мы уже сказали, выполнение неравенств (20) и (21) гарантирует однозначную разрешимость уравнения $u = Tu$, а с ним и уравнения $u = Au$. В силу полной непрерывности оператора $A$ отсюда следует обратимость $I - A$, а следовательно, и оператора исходной задачи (1). С другой стороны, (20) и (21) будут выполнены, как это следует из (23) и (24), если верны следующие соотношения: $MC(N)(C_1 + C_2)\zeta < \gamma$, $C(N)(C_3\gamma + C_4\zeta) < \zeta$. Для произвольного $\varepsilon > 0$ положим $\gamma := MC(N)(C_1 + C_2)\zeta + \varepsilon$, тогда выполнено первое из неравенств. Подставив $\gamma = MC(N)(C_1 + C_2)$ во второе, получим достаточное условие в виде

$$\begin{gathered}C(N)\Big(C_3\big(MC(N)(C_1 + C_2)\zeta + \varepsilon\big) + C_4\zeta\Big) < \zeta, \quad \text{или} \\ \Big(1 - C(N)\big(C_3 MC(N)(C_1 + C_2) + C_4\big)\Big)\zeta > C(N)C_3\varepsilon.\end{gathered}$$

Поскольку значение $\varepsilon$ можно выбрать сколь угодно малым, то для выполнения этого последнего неравенства достаточно потребовать $C(N)\big(C_3 MC(N)(C_1 + C_2) + C_4\big) < 1$. Итак, мы нашли достаточное условие на константу $\kappa$, которое обеспечивает однозначную разрешимость задачи $u = Au$, а следовательно, и однозначную разрешимость исходной задачи (1). Теорема доказана.

**Теорема 2.** *Пусть матрица $G$ обратима. Если $\kappa < 1$ и $u \in H_0^1(\Omega)$ — решение задачи (1), то $u \in X(\Omega)$ и имеет место оценка $\|u - P_N u\|_{H_0^1} \leqslant C(N)\sigma\|f\|_{L^2}$.*

**Доказательство.** Заметим, что принадлежность $u$ пространству $X(\Omega)$ согласно нашему определению очевидна, поскольку можно положить $f_v = -f + b \cdot \nabla u + cu$. Как и в доказательстве теоремы 1, рассмотрим уравнение в форме (13), где (см. раздел 4.1) $v = -\triangle^{-1} f$. Далее, в условиях теоремы 1 решение уравнения $u = Au + v$ существует. Разложим $u$ и $v$ на составляющие по взаимно ортогональным подпространствам $H_0^1(\Omega) = S_N^0 \oplus S_N^\perp$ с помощью проекторов $P_N$ и $I - P_N$: $u_N := P_N u$, $u_\perp := (I - P_N)u$ и т.д. Непосредственно проверим следующие равенства:

$$u_N = [I - A]_N^{-1}(P_N A u_\perp + P_N v), \quad u_\perp = (I - P_N)A(u_N + u_\perp) + (I - P_N)v.$$

Действительно, поскольку $(I - A)u = v$ или, что то же самое, $Au = u - v$, то имеем

$$(I - P_N)A(u_N + u_\perp) + (I - P_N)v = (I - P_N)Au + (I - P_N)v = (I - P_N)(u - v + v) = (I - P_N)u = u_\perp,$$

$$\begin{aligned}
[I - A]_N^{-1}(P_N A u_\perp + P_N v) &= [I - A]_N^{-1}\big(P_N A u_\perp + P_N (I - A)u\big) = \\
&= [I - A]_N^{-1}\big(P_N A u_\perp + P_N (I - A)(u_N + u_\perp)\big) = [I - A]_N^{-1} P_N (A u_\perp + u_N + u_\perp - A u_N - A u_\perp) = \\
&= [I - A]_N^{-1} P_N (u_N - A u_N + u_\perp) = [I - A]_N^{-1} P_N (I - A) u_N + [I - A]_N^{-1} P_N u_\perp = \\
&= [I - A]_N^{-1} P_N (I - A) u_N + 0 = u_N.
\end{aligned}$$

Вспомним (лемма 3), что $\big\|[I - A]_N^{-1}\big\|_{H_0^1} \leqslant M$, поэтому можем записать $\|u_N\|_{H_0^1} \leqslant M\|P_N A u_\perp + P_N v\|_{H_0^1}$. Далее, по лемме 1 имеем оценку $\|P_N A u_\perp\|_{H_0^1} \leqslant C(N)(C_1 + C_2)\|u_\perp\|_{H_0^1}$; следовательно,

$$\|u_N\|_{H_0^1} \leqslant MC(N)(C_1 + C_2)\|u_\perp\|_{H_0^1} + M\|P_N v\|_{H_0^1}. \tag{25}$$

Норма функции $u_\perp$ оценивается так (с применением леммы 2):

$$\begin{aligned}
\|u_\perp\|_{H_0^1} &\leqslant \big\|(I - P_N)A(u_N + u_\perp)\big\|_{H_0^1} + \big\|(I - P_N)v\big\|_{H_0^1} \leqslant \\
&\leqslant C(N)\big(C_3\|u_N\|_{H_0^1} + C_4\|u_\perp\|_{H_0^1}\big) + \big\|(I - P_N)v\big\|_{H_0^1}.
\end{aligned} \tag{26}$$

Подставим оценку нормы $\|u_N\|_{H_0^1}$ из (25) в правую часть (26). Решив полученное неравенство относительно $\|u_\perp\|_{H_0^1}$, будем иметь

$$\|u_\perp\|_{H_0^1} \leqslant \frac{C(N)C_3 M}{1 - \kappa}\|P_N v\|_{H_0^1} + \frac{1}{1 - \kappa}\big\|(I - P_N)v\big\|_{H_0^1}.$$

Для оценки первого слагаемого заметим сначала, что норма ортопроектора не превосходит единицы (а $P_N$ — ортопроектор именно в смысле скалярного произведения $(\cdot,\cdot)_{H_0^1}$) и что для $v = -\triangle^{-1} f$ верно неравенство $\|v\|_{H_0^1} \leqslant C_p\|f\|$, так как

$$\|v\|_{H_0^1}^2 \equiv (\nabla v, \nabla v) = (v, f) \leqslant \|v\|\,\|f\| \leqslant C_p\|v\|_{H_0^1}\|f\|. \tag{27}$$

Для второго же слагаемого непосредственно из (5) получим $\big\|(I - P_N)v\big\|_{H_0^1} \leqslant C(N)\|\triangle v\| \equiv C(N)\|f\|$. Вспомним теперь, что $\sigma = \big(1 + C_p M(N)C_3\big)/\big(1 - \kappa(N)\big)$, и получим утверждение теоремы. Теорема доказана.

**4.4. Оценки погрешности приближенного решения.** Введем теперь константы

$$\alpha = \sqrt{1 + \big(M(N)C(N)(C_1 + C_2)\big)^2} \quad \text{и} \quad \beta = 1 + C_p M(N)(C_1 + C_2).$$

**Теорема 3.** *Пусть матрица $G$ обратима. Тогда для произвольного $v \in H_0^1(\Omega)$ верны оценки*

$$\begin{aligned}
&\|v - P_{\mathcal{L}} v\|_{H_0^1} \leqslant \alpha\|v - P_N v\|_{H_0^1}, \\
&\|v - P_{\mathcal{L}} v\|_{L^2} \leqslant C(N)\beta\|v - P_N v\|_{H_0^1} \leqslant C(N)\beta\|v - P_{\mathcal{L}} v\|_{H_0^1}.
\end{aligned} \tag{28}$$

**Доказательство.** Заметим, что $P_N$ и $P_{\mathcal{L}}$ суть проекторы на одно и то же подпространство $S_N^0$, причем первый из них — ортогональный. Поэтому для любого $v \in H_0^1(\Omega)$ верны тождества

$$\begin{aligned}\|v - P_{\mathcal{L}}v\|_{H_0^1}^2 &= \big\|(v - P_N v) + (P_N v - P_{\mathcal{L}}v)\big\|_{H_0^1}^2 = \\ &= \|v - P_N v\|_{H_0^1}^2 + \|P_{\mathcal{L}}v - P_N v\|_{H_0^1}^2 + (v - P_N v, P_N v - P_{\mathcal{L}}v)_{H_0^1} = \\ &= \|v - P_N v\|_{H_0^1}^2 + \|P_{\mathcal{L}}v - P_N v\|_{H_0^1}^2\end{aligned} \tag{29}$$

(откуда сразу же следует второе неравенство второй строки (28)) и $P_{\mathcal{L}}(P_N v) = P_N v$.

Положим теперь $g = v - P_N v$ и заметим, что $P_{\mathcal{L}}v - P_N v = P_{\mathcal{L}}(v - P_N v)$. Для оценки этой функции заметим сначала, что при любом $w \in H_0^1$ верно

$$P_N(I - A)P_{\mathcal{L}}w = P_N(I - A)w. \tag{30}$$

Действительно, при произвольном $j = 1, \ldots, N$ имеем

$$\begin{aligned}\big(\nabla P_N(I - A)P_{\mathcal{L}}w, \nabla\varphi_j\big) &= \big\{\text{формула (3)}\big\} = \big(\nabla(I - A)P_{\mathcal{L}}w, \nabla\varphi_j\big) = \\ &= (\nabla P_{\mathcal{L}}w, \nabla\varphi_j) + (b \cdot \nabla P_{\mathcal{L}}w + cP_{\mathcal{L}}w, \varphi_j) \equiv a(P_{\mathcal{L}}w, \varphi_j) = \big\{\text{формула (8)}\big\} = a(w, \varphi_j) = \\ &= (\nabla w, \nabla\varphi_j) + (b \cdot \nabla w + cw, \varphi_j) \equiv (\nabla(I - A)w, \nabla\varphi_j) = \big\{\text{формула (3)}\big\} = \big(\nabla P_N(I - A)w, \nabla\varphi_j\big).\end{aligned} \tag{31}$$

Поскольку обе части (30) принадлежат $S_N^0$, а в силу (31) их разность ортогональна всем базисным векторам этого подпространства, то равенство (30) справедливо. Если применить к обеим его частям оператор $[I - A]^{-1} \equiv (P_N(I - A)|_{S_N^0})^{-1}$, получим равенство $P_{\mathcal{L}}w = [I - A]_N^{-1}P_N(I - A)w$.

Положим в нем $w = g \equiv P_\perp v$. Имеем

$$P_{\mathcal{L}}g = [I - A]_N^{-1}P_N(I - A)g = \{P_N g = 0\} = [I - A]_N^{-1}P_N(-Ag) = [I - A]_N^{-1}P_N\psi, \tag{32}$$

где $\psi \equiv -Ag \equiv -\triangle^{-1}(b \cdot \nabla + c)g$, т.е. $\psi$ является обобщенным решением из $H_0^1$ однородной задачи Дирихле: $(\nabla\psi, \nabla\varphi) = ((b \cdot \nabla + c)g, \varphi) \quad \forall\varphi \in H_0^1$. Формула (32) вместе с леммой 3 позволяют оценить второе слагаемое в (29) следующим образом:

$$\begin{aligned}\|P_{\mathcal{L}}v - P_N v\|_{H_0^1} = \|P_{\mathcal{L}}g\|_{H_0^1} &\leqslant M(N)\|P_N\psi\|_{H_0^1} = M(N)\|P_N Ag\|_{H_0^1} \leqslant \\ &\leqslant M(N)C(N)(C_1 + C_2)\|g\|_{H_0^1} \equiv M(N)C(N)(C_1 + C_2)\|v - P_N v\|_{H_0^1}.\end{aligned} \tag{33}$$

Последнее неравенство составляет утверждение леммы 1, если вспомнить, что $g \in S_N^\perp$.

Таким образом, первое из неравенств (28) доказано соотношениями (33) и (29).

Из (33) с помощью неравенства Пуанкаре–Фридрихса получим:

$$\|P_{\mathcal{L}}v - P_N v\|_{L^2} \leqslant C_p M(N)C(N)(C_1 + C_2)\|v - P_N v\|_{H_0^1}.$$

Отсюда следует первое неравенство из второй строки (28):

$$\|v - P_{\mathcal{L}}v\|_{L^2} \leqslant \|v - P_N v\|_{L^2} + \|P_{\mathcal{L}}v - P_N v\|_{L^2} \leqslant C(N)\|v - P_N v\|_{H_0^1} + C(N)C_p M(N)(C_1 + C_2)\|v - P_N v\|_{H_0^1}.$$

Теорема доказана.

Далее, пусть $S_N$ — конечномерные подпространства $H^1(\Omega)$, причем $S_N^0 \subset S_N$. Формально они могут совпадать с $S_N^0$, но для получения хороших оценок, по-видимому, необходимо, чтобы они были шире $S_N^0$ и были предельно плотны в $H^1(\Omega)$, точно так же, как для хорошего приближения решения нужно, чтобы $S_N^0$ были предельно плотны в $H_0^1(\Omega)$. Определим оператор $P$ проектирования из $L^2(\Omega)$ в $S_N$, который каждому $v \in L^2(\Omega)$ сопоставляет $Pv \in S_N$ по правилу

$$(v - Pv, \widetilde{\varphi}_{\text{discr}})_{L_2} = 0 \quad \forall\widetilde{\varphi}_{\text{discr}} \in S_N. \tag{34}$$

Теперь определим осредненный градиент приближенного решения $(\overline{\nabla}v) \equiv (P\nabla_x v, P\nabla_y v, P\nabla_z v)$, каждая компонента которого принадлежит $S_N$ (эта формула выписана, очевидно, для $n = 3$).

Для этого дополним набор базисных функций $\{\varphi_i\}_{i=1}^N$ до базиса в $S_N$, что для метода конечных элементов на практике будет означать добавление базисных функций, соответствующих граничным узлам. Во избежание путаницы с номерами обозначим полученную систему функций через $\{\widetilde{\varphi}_j\}_{j=1}^M$. Пусть

$\{v_i\}_{i=1}^N \equiv V$ — коэффициенты в разложении приближенного решения $v$ по базису $\{\varphi_i\}_{i=1}^N$ пространства $S_N^0$. Тогда приближенное решение имеет вид $v = \sum\limits_{i=1}^N v_i\varphi_i(x)$, а его градиент — $\nabla v = \sum\limits_{i=1}^N v_i\nabla\varphi_i$. Задача состоит в получении проекции каждой компоненты $\nabla v$ на линейную оболочку функций $\{\widetilde{\varphi}_j\}_{j=1}^M$. Для этого запишем, в полном соответствии с (34), уравнения $\big((\overline{\nabla}v)_k, \widetilde{\varphi}_j\big)_{L^2} = \big((\nabla v)_k, \widetilde{\varphi}_j\big)_{L^2}$, $j = 1,\ldots,M$, $k = 1,\ldots,n$, или $\sum\limits_{l=1}^M \overline{v}_l^k(\widetilde{\varphi}_l, \widetilde{\varphi}_j)_{L^2} = \sum\limits_{i=1}^N v_i\left(\dfrac{\partial\varphi_i}{\partial x_k}, \widetilde{\varphi}_j\right)_{L^2}$. Здесь символами $\overline{v}_l^k$ обозначены пока неизвестные коэффициенты в разложении проекции $\dfrac{\partial v}{\partial x_k}$ по новому набору базисных функций $\{\widetilde{\varphi}_l\}_{l=1}^M$. Если теперь определить $(M\times M)$-матрицу Грама базисных функций $\widetilde{D} = \{\widetilde{D}_{jl}\}_{j,l=1}^M = \big\{(\widetilde{\varphi}_l, \widetilde{\varphi}_j)_{L^2}\big\}$ и $(M\times N)$-матрицы $W^k = \{w_{ji}^k\}_{j=1,\ldots,M;\, i=1,\ldots,N} \equiv \left\{\left(\dfrac{\partial\varphi_i}{\partial x_k}, \widetilde{\varphi}_j\right)_{L^2}\right\}$, $k = 1,\ldots,n$, то предыдущие соотношения можно записать в виде $\sum\limits_{l=1}^M \widetilde{D}_{jl}\overline{v}_l^k = \sum\limits_{i=1}^N w_{ji}^k v_i$, $j = 1,\ldots,M$, $k = 1,\ldots,n$, или, в матричных обозначениях, $\widetilde{D}\overline{V}^k = W^kV$, $k = 1,\ldots,n$. Разрешив эти $n$ линейных систем относительно $\{\overline{v}_l^k\}_{l=1}^M \equiv \overline{V}^k$, мы сможем построить осредненный градиент $\overline{\nabla}v$.

Введем еще функции $R = \nabla v - \overline{\nabla}v$, $S = f + \operatorname{div}(\overline{\nabla}v) - b\cdot\nabla v - cv$ и сформулируем теорему об оценках.

**Теорема 4.** *Пусть матрица $G$ обратима. При $\kappa < 1$ для решения и задачи $\mathcal{L}u = f$ и приближенного решения $P_{\mathcal{L}}u \equiv v$ имеем оценки*

$$\|u - P_{\mathcal{L}}u\|_{H_0^1} \leqslant C(N)\alpha\sigma\|f\|_{L^2}, \quad \|u - P_{\mathcal{L}}u\|_{L^2} \leqslant \big(C(N)\big)^2\beta\sigma\|f\|_{L^2}. \tag{35}$$

*Верна также оценка (использующая приближенное решение)*

$$\|u - P_{\mathcal{L}}u\|_{H_0^1} \leqslant \|R\|_{L^2} + C(N)\beta\|S\|_{L^2} + \big(C(N)\big)^2\beta\sigma\big(C_b + C(N)C_c\beta\big)\|f\|_{L^2}. \tag{36}$$

**Доказательство.** Первые две оценки следуют из теорем 2 и 3. Положим теперь $e \equiv u - v$. Тогда имеем

$$\begin{aligned}\|u-v\|_{H_0^1}^2 &= (\nabla e, \nabla u) - (\nabla e, \nabla v) = (e, f) - (e, b\cdot\nabla u + cu) - (\nabla e, \nabla v) = \\ &= (e, f - b\cdot\nabla v - cv) - (e, b\cdot\nabla e + ce) - (\nabla e, \nabla v).\end{aligned}$$

Здесь во втором равенстве используется обобщенная постановка (2) исходной задачи, а в качестве $w$ взята $e \in H_0^1$. Поскольку $\big(\nabla w, (\overline{\nabla}v)\big) = \big(w, -\operatorname{div}(\overline{\nabla}v)\big)$ при любом $w \in H_0^1$, то, полагая $w = e$, получим:

$$\begin{aligned}\|u-v\|_{H_0^1}^2 &= \big(e, f - b\cdot\nabla v - cv + \operatorname{div}(\overline{\nabla}v)\big) - (e, b\cdot\nabla e + ce) - (\nabla e, \nabla v - \overline{\nabla}v) = \\ &= (e, S) - (e, b\cdot\nabla e + ce) - (\nabla e, R) \leqslant \|e\|_{L^2}\|S\|_{L^2} + \|e\|_{L^2}\|b\cdot\nabla e + ce\|_{L^2} + \|e\|_{H_0^1}\|R\|_{L^2}.\end{aligned}$$

С учетом оценки $\|b\cdot\nabla e + ce\|_{L^2} \leqslant \big\|\|b\|_2\big\|_{L^\infty}\|e\|_{H_0^1} + \|c\|_{L^\infty}\|e\|_{L^2}$ и (из теоремы 3) $\|e\|_{L^2} \leqslant C(N)\beta\|e\|_{H_0^1}$, группируя все множители, имеем третью оценку:

$$\begin{aligned}\|e\|_{H_0^1}^2 &\leqslant \|e\|_{L^2}\big(\|S\|_{L^2} + \big\|\|b\|_2\big\|_{L^\infty}\|e\|_{H_0^1} + \|c\|_{L^\infty}\|e\|_{L^2}\big) + \|e\|_{H_0^1}\|R\|_{L^2} \leqslant \\ &\leqslant C(N)\beta\|e\|_{H_0^1}\|S\|_{L^2} + \|e\|_{L^2}C_b\|e\|_{H_0^1} + \|e\|_{L^2}C_cC(N)\beta\|e\|_{H_0^1} + \|e\|_{H_0^1}\|R\|_{L^2}, \quad \text{или}\end{aligned}$$

$$\begin{aligned}\|e\|_{H_0^1} &\leqslant C(N)\beta\|S\|_{L^2} + \|e\|_{L^2}\big(C_b + C_cC(N)\beta\big) + \|R\|_{L^2} \leqslant \{\text{теоремы 2, 3}\} \leqslant \\ &\leqslant C(N)\beta\|S\|_{L^2} + \big(C_b + C_cC(N)\beta\big)\big(C(N)\big)^2\beta\sigma\|f\|_{L^2} + \|R\|_{L^2}.\end{aligned}$$

Теорема доказана.

**4.5. Алгоритм вычисления приближенного решения и оценки погрешности.** В этом разделе на основе вышеизложенной теории предложен алгоритм вычисления приближенного решения задачи (1) и построения оценок погрешности (35) и (36).

1. Первый этап алгоритма.

а) Оценить сверху или определить аналитически константу $C_p$, входящую в неравенство Пуанкаре–Фридрихса $\|w\| \leqslant C_p\|w\|_{H_0^1} \quad \forall w \in H_0^1(\Omega)$. Хорошо известно [4], что $C_p = \lambda_1^{-1/2}$, где $\lambda_1$ — наименьшее

собственное значение задачи Дирихле для оператора Лапласа в $\Omega$. Для параллелепипеда (прямоугольника, отрезка) оно равно $\sum\limits_{i=1}^{n}\left(\dfrac{\pi}{l_i}\right)^2$, где $l_i$ — стороны параллелепипеда. Произвольную же область можно вписать в какую-либо область $\widetilde{\Omega}$, первое собственное значение задачи Дирихле для оператора Лапласа в которой известно (например, параллелепипед), и взять $C_p(\widetilde{\Omega})$ в качестве оценки сверху для $C_p(\Omega)$ [4].

б) Исходя из выбранной дискретизации, определить константу $C(N)$. В частности, если $\Omega$ — прямоугольник и используются кусочно-билинейные конечные элементы на квадратной сетке шага $h$, то $C(N)=\dfrac{h}{\pi}$ [9]. Если же область $\Omega$ представляет собой выпуклый многоугольник, то для линейных конечных элементов на треугольной сетке соответствующая оценка получена Наттерером (см. раздел 3).

в) Исходя из выбранной дискретизации, сформировать матрицы $D=\{D_{ji}\}=\big\{(\nabla\varphi_i,\nabla\varphi_j)_{L^2}\big\}$ и $G=\big\{G_{ji}\big\}=\{(\nabla\varphi_i,\nabla\varphi_j)_{L^2}+(b\cdot\nabla\varphi_i,\varphi_j)_{L^2}+(c\varphi_i,\varphi_j)_{L^2}\big\}$.

г) Вычислить вектор-столбец правой части: $F=\{F_i\}^{\mathrm{T}}=\{\int_\Omega f(x)\varphi_i\,dx\}^{\mathrm{T}}$.

д) Получить вектор $V$ коэффициентов разложения приближенного решения по функциям $\{\varphi_i\}_{i=1}^N$ (степени свободы) как решение линейной системы $GV=F$. Теперь можно вычислить само приближенное решение: $v=\sum\limits_{i=1}^{N}V_i\varphi_i$.

е) Зная коэффициенты уравнения (1), аналитически найти или численно оценить сверху константы $C_{\operatorname{div} b}=\|\operatorname{div} b(x)\|_{L^\infty}$, $C_b=\big\|\,\|b(x)\|_2\,\big\|_{L^\infty}$, $C_c=\|c(x)\|_{L^\infty}$, $C_1=C_pC_{\operatorname{div} b}+C_b$, $C_2=C_pC_c$, $C_3=C_b+C_pC_c$ и $C_4=C_b+C(N)C_c$.

ж) Вычислить константы $M(N)=\|L^{\mathrm{T}}G^{-1}L\|_2$, $\kappa(N)=C(N)\big[C(N)M(N)(C_1+C_2)C_3+C_4\big]$, $\sigma=\dfrac{1+C_pM(N)C_3}{1-\kappa(N)}$.

Вычислить $\alpha=\sqrt{1+\big(M(N)C(N)(C_1+C_2)\big)^2}$ и $\beta=1+C_pM(N)(C_1+C_2)$.

з) Проверить, что $\kappa$ оказалось меньше 1. Напомним, что выкладки имеют смысл лишь в этом случае, ибо иначе (теорема 1) не гарантируется разрешимость исходной дифференциальной задачи и (теорема 2) происходит деление неравенства на неположительное число $1-\kappa$.

и) Вычислить оценки погрешности $\|u-P_{\mathcal{L}}u\|_{H_0^1}\leqslant C(N)\alpha\sigma\|f\|_{L^2}$ и $\|u-P_{\mathcal{L}}u\|_{L^2}\leqslant\big(C(N)\big)^2\beta\sigma\|f\|_{L^2}$.

2. Второй этап алгоритма.

а) Зная вектор $V$ координат (степеней свободы) приближенного решения $v$, вычислить градиент $\nabla v$ приближенного решения.

б) Вычислить матрицу Грама $\widetilde{D}$ функций $\{\widetilde{\varphi}_j\}_{j=1}^M$ и матрицы $W^k$ по формулам $\widetilde{D}_{jl}=(\widetilde{\varphi}_l,\widetilde{\varphi}_j)_{L^2}$ и $(W^k)_{ji}=\left(\widetilde{\varphi}_j,\dfrac{\partial\varphi_i}{\partial x_k}\right)_{L^2}$.

в) Решив для каждого $k=1,\ldots,n$ линейные системы $\widetilde{D}\overline{V}^{\,k}=W^kV$, найти коэффициенты $\overline{V}^{\,k}$ в разложении $k$-й компоненты осредненного градиента $\overline{\nabla}v$.

г) Зная функции $\{\widetilde{\varphi}_j\}_{j=1}^M$ и их градиенты, вычислить осредненный градиент

$$\overline{\nabla}v=\left(\sum_{j=1}^{M}(\overline{V}^{\,1})_j\widetilde{\varphi}_j,\ldots,\sum_{j=1}^{M}(\overline{V}^{\,n})_j\widetilde{\varphi}_j\right)\quad\text{и}\quad\operatorname{div}(\overline{\nabla}v)=\sum_{i=1}^{n}\sum_{j=1}^{M}(\overline{V}^{\,i})_j\,\frac{\partial\widetilde{\varphi}_j}{\partial x_i}\,.$$

д) Вычислить $\|R\|=\big\|\nabla v-\overline{\nabla}v\big\|=\sqrt{\sum\limits_{i=1}^{n}\big\|R_i\big\|^2}$ и $\|S\|=\big\|f+\operatorname{div}(\overline{\nabla}v)-b\cdot\nabla v-cv\big\|$. При этом $\|R\|$ можно найти как функционал от коэффициентов $V$ и $\overline{V}^{\,k}$, заранее аналитически найдя необходимые интегралы от произведений функций $\dfrac{\partial\varphi_i}{\partial x_k}$ и $\widetilde{\varphi}_j$, а $\|S\|$, вообще говоря, придется рассчитывать численно, кроме случая простых $f$, $b$ и $c$.

е) Вычислить оценку $\|u-P_{\mathcal{L}}u\|_{H_0^1}\leqslant\|R\|_{L^2}+C(N)\beta\|S\|_{L^2}+\big(C(N)\big)^2\beta\sigma\big(C_b+C(N)C_c\beta\big)\|f\|_{L^2}$.

**5. Модификация предложенной Накао оценки для уравнения Гельмгольца.** Рассмотрим уравнение Гельмгольца $-\triangle u-k^2(x)=f$, где $k^2$, вообще говоря, не является постоянной величиной, но ограничено сверху и снизу положительными постоянными. В этом случае теория и алгоритмы, изложенные выше, применимы. Однако можно построить более точную оценку, что и является основной целью данной работы.

**5.1. Основная идея.** В леммах 1 и 2 фактически использовалась оценка $\|P_N A P_\perp\|_{H_0^1} \leqslant C(N) C_c C_p$ (в применении к рассматриваемому случаю). Однако если $c \equiv -k^2$ не является постоянной величиной, то можно вычислить более точную оценку. Для этого прежде всего заметим, что у нас $c_1 \leqslant c \leqslant c_0 < 0$ и $b \equiv 0$, а следовательно, оператор $A$, заданный формулой (11), является ограниченным, самосопряженным и положительно определенным. Введем еще, для удобства обозначений, оператор $A_1$, совпадающий с оператором $A$ при $k^2 \equiv 1$, т.е. заданный формулой $(\nabla u, A_1 \nabla w) = (u, w)$. Напомним, что всюду, где знак скалярного произведения или нормы употребляется без индексов, они понимаются в смысле $L^2$. Тогда неравенство (6) может быть записано в виде $(P_\perp v, A_1 P_\perp v)_{H_0^1} \leqslant \big(C(N)\big)^2 (P_\perp v, P_\perp v)_{H_0^1}$.

Аналогично, из (6) с учетом тождества $k^2(x) \equiv -c(x)$ и (9) имеем

$$(P_\perp v, A P_\perp v)_{H_0^1} = \big(k^2(x) P_\perp v, P_\perp v\big) \leqslant C_c (P_\perp v, P_\perp v) \leqslant C_c \big(C(N)\big)^2 (P_\perp v, P_\perp v)_{H_0^1},$$

отсюда в силу самосопряженности операторов $P_\perp$ и $P_\perp A P_\perp$ получим

$$\|P_\perp A P_\perp\|_{H_0^1} \leqslant C_c \big(C(N)\big)^2. \tag{37}$$

Теперь заметим, что

$$\|P_N A P_\perp\|_{H_0^1} = \sup_{\|u\|_{H_0^1}=1, \|v\|_{H_0^1}=1,\ u,v \in H_0^1} \big|(u, P_N A P_\perp v)_{H_0^1}\big|. \tag{38}$$

Используя (37), оценим скалярное произведение, стоящее в правой части (38). Поскольку $A$ — положительно определенный самосопряженный оператор, то из него можно извлечь квадратный корень ([12], с. 219) и следующие преобразования справедливы:

$$\begin{aligned} \big|(u, P_N A P_\perp v)_{H_0^1}\big| &= \big|(P_N u, A P_\perp u)_{H_0^1}\big| = \Big|\big(\sqrt{A} P_N u, \sqrt{A} P_\perp v\big)_{H_0^1}\Big| \leqslant \big\|\sqrt{A} P_N u\big\|_{H_0^1} \big\|\sqrt{A} P_\perp v\big\|_{H_0^1} = \\ &= \sqrt{(u, P_N A P_N u)_{H_0^1}}\, \sqrt{(v, P_\perp A P_\perp v)_{H_0^1}} \leqslant \sqrt{\|P_N A P_N\|_{H_0^1}}\, \|u\|_{H_0^1} \sqrt{\|P_\perp A P_\perp\|_{H_0^1}}\, \|v\|_{H_0^1} \leqslant \\ &\leqslant \sqrt{\|P_N A P_N\|_{H_0^1}} \|u\|_{H_0^1} \sqrt{C_c}\, C(N) \|v\|_{H_0^1}. \end{aligned} \tag{39}$$

Норма $\|P_N A P_N\|_{H_0^1}$ не превосходит нормы этого же оператора в $S_N^0$, так как

$$\|P_N A P_N\|_{H_0^1} \equiv \|P_N A P_N^2\|_{H_0^1} \leqslant \|P_N A P_N|_{S_N^0}\|_{H_0^1}\, \|P_N\|_{H_0^1} = \|P_N A P_N|_{S_N^0}\|_{H_0^1}.$$

Поскольку $(\varphi_i, P_N A P_N \varphi_j)_{H_0^1} = H_{ji} := D_{ji} - G_{ji}$, где $\{\varphi_i\}_{i=1}^N$ — базис в $S_N^0$, а $D$ является матрицей Грама этого базиса, то $\|P_N A P_N|_{S_N^0}\|_{H_0^1}$ — наибольшее собственное значение задачи $H\nu = \lambda D \nu$.

**Замечание 1.** Поскольку в рассматриваемом случае матрица $G$ является симметричной, то норму оператора $[I - A]_N^{-1}$ нет необходимости оценивать по формуле $\big\|[I - A]_N^{-1}\big\|_{H_0^1} \leqslant M \equiv \|L^{\mathrm{T}} G^{-1} L\|_2$. Действительно, если (см. доказательство леммы 3) $\Psi = G^{-1} D V$, то $\big\|[I - A]_N^{-1}\big\|_{H_0^1}$ есть величина, обратная модулю наименьшего по абсолютной величине собственного значения $(I - P_N A)\big|_{S_N^0}$, поскольку эти операторы взаимно обратны. Однако это собственное значение, аналогично рассмотренному случаю оператора $P_N A P_N\big|_{S_N^0}$, равно наименьшему по модулю собственному значению задачи $G\nu = \lambda D \nu$, т.е. можно положить $M := \big\|[I - A]_N^{-1}\big\|_{H_0^1} = \left(\min\limits_{\lambda\text{— соб. знач. } G\nu = \lambda D\nu} |\lambda|\right)^{-1}$.

**Замечание 2.** Уточнить оценку $(P_\perp v, A P_\perp v)_{H_0^1} \leqslant C_c \big(C(N)\big)^2 \|v\|_{H_0^1}^2$, вообще говоря, нельзя. Рассмотрим такой пример. Пусть $\Omega$ представляет собой отрезок $[0;1]$, а $S_N^0$ — пространство непрерывных функций, аффинных на каждом отрезке $[i/N;(i+1)/N]$, $i = 1, \ldots, N-1$, и равных нулю на концах отрезка. Тогда [9] $P_N v$ совпадает с $v$ в точках $i/N$, $i = 0, \ldots, N$ (т.е. в узлах сетки). Из результатов [9] следует, что $S_N^0$ удовлетворяет условиям (5) и (6) с $C(N) = 1/(N\pi)$, где $1/N$ — шаг сетки. Положим $k^2 = 1$ на $[0;1/2]$ и $k^2 = 4$ на $(1/2;1]$ и рассмотрим функцию $v$, равную нулю на $[0;0{,}9]$ и $\sin 10\pi x$ на $(0{,}9;1]$. Это кусочно-гладкая функция, принадлежащая $H_0^1[0;1]$. Выберем $N = 10$. Как сказано выше, для рассматриваемых $\Omega$ и $S_N^0$ функции $P_N v$ и $v$ совпадают в узлах сетки, т.е. $(P_N v)(i/N) = v(i/N) = 0$ при всех $i = 0, \ldots, 10$. В силу того, что $P_N v$ — аффинная на отрезках

$[i/N;(i+1)/N]$ функция, из этого равенства следует, что $P_N v \equiv 0$ на $[0;1]$, откуда $P_\perp v = v$. Следовательно, $(P_\perp v, AP_\perp v)_{H_0^1} = (v, Av)_{H_0^1}$. С учетом определения оператора $A$ (см. раздел 4.1), а также обращения $v$ в ноль на $[0;0{,}9]$ имеем $(v,Av)_{H_0^1} = \big(k^2(x)v, v\big) = 4\int\limits_{0,9}^{1} \sin^2 10\pi x\, dx = 4/20 = 1/5$. В то же время, $(v,v)_{H_0^1} \equiv (v',v') = \int\limits_{0,9}^{1} (10\pi)^2 \cos^2 10\pi x\, dx = 5\pi^2$. Следовательно,

$$(P_\perp v, AP_\perp v)_{H_0^1} / \|v\|^2_{H_0^1} = (1/5)/(5\pi^2) = 0{,}04/\pi^2.$$

Поскольку $C_c \equiv \big\|k^2(x)\big\|_{L^\infty} = 4$ и $C(N) = 1/(N\pi) = 0{,}1/\pi$, имеем $C_c\big(C(N)\big)^2 = 0{,}04/\pi^2$, т.е. оценка $(P_\perp v, AP_\perp v)_{H_0^1} \leqslant C_c\big(C(N)\big)\|v\|^2_{H_0^1}$ точна.

**Замечание 3.** Необходимо отметить, что если вместо $k^2(x)$ в уравнении присутствует член $c(x)$ переменного знака, то оператор $A$ не является знакоопределенным и поэтому выкладки (39) не могут быть проведены. Действительно, даже в конечномерном пространстве легко привести пример самосопряженного незнакоопределенного оператора $B$ и ортопроекторов $P_1$ и $P_2 = I - P_1$, для которых неравенство $\|P_1BP_2\| \leqslant \sqrt{\|P_1BP_1\|\,\|P_2BP_2\|}$ неверно: $B = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, $P_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$, $P_2 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$. Тогда $P_1BP_1 = P_2BP_2 = 0$, но $P_1BP_2 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, поэтому $\|P_1BP_2\| > \|P_1BP_1\|\,\|P_2BP_2\|$.

**5.2. Вывод модифицированных оценок.** Пусть рассматриваемая задача выглядит так:

$$\mathcal{L}_H u \equiv -\triangle u - k^2(x)u = f, \quad u\big|_{\partial\Omega} = 0, \tag{40}$$

где $k^2 \equiv -c \in [k_1^2; k_2^2]$, $k_1^2 > 0$ и $k_2^2 < \infty$. Здесь индекс $H$ подчеркивает, что рассматривается уравнение Гельмгольца. Ниже, если применяется какое-либо из использованных ранее обозначений, оно имеет тот же смысл, что и прежде, но применительно к оператору $\mathcal{L}_H$. Однако мы введем и новые константы (здесь и далее аргумент $N$ подразумевается, если не написан): $L(N) = \sqrt{\big\|P_N AP_N|_{S_N^0}\big\|_{H_0^1} C_c}$, $\theta(N) = \big(C(N)\big)^2 (L^2M + C_c)$, $\tau(N) = \dfrac{1 + C_p ML}{1 - \theta(N)}$. Теперь мы готовы сформулировать необходимые утверждения.

**Теорема 1а.** *Если матрица $G$ обратима и $\theta < 1$, то оператор $\mathcal{L}_H$ обратим.*

**Доказательство.** Прежде всего описанным выше образом вводим операторы $A$ и $[I - A]_N^{-1}$. Понадобятся еще две леммы.

**Лемма 1а.** *Пусть матрица $G$ обратима. Для любого $w \in H_0^1$ верна оценка*

$$\|P_N AP_\perp w\|_{H_0^1} \leqslant LC(N)\|P_\perp w\|_{H_0^1}.$$

**Доказательство.** В разделе 5.1 получено следующее неравенство: $\|P_N AP_\perp\|_{H_0^1} \leqslant LC(N)$. Из него следует, что $\|P_N AP_\perp w\|_{H_0^1} = \|P_N AP_\perp^2 w\|_{H_0^1} \leqslant LC(N)\|P_\perp w\|_{H_0^1}$. Лемма доказана.

**Лемма 2а.** *Пусть матрица $G$ обратима. Для произвольного $w \in H_0^1$ верна оценка*

$$\|P_\perp Aw\|_{H_0^1} \leqslant C(N)\big(C(N)C_c\|w_\perp\|_{H_0^1} + L\|w_N\|_{H_0^1}\big).$$

**Доказательство.** Имеем $\|P_\perp AP_\perp w\|_{H_0^1} = \|P_\perp AP_\perp^2 w\|_{H_0^1} \leqslant \|P_\perp AP_\perp\|_{H_0^1}\|w_\perp\|_{H_0^1} \leqslant C_c\big(C(N)\big)^2\|w_\perp\|_{H_0^1}$, а $\|P_\perp AP_N w\|_{H_0^1} = \big\|P_\perp A(P_N)^2 w\big\|_{H_0^1} \leqslant \|P_\perp AP_N\|_{H_0^1}\|w_N\|_{H_0^1}$. Пользуясь равенством спектральных норм взаимно сопряженных операторов, имеем по доказанному: $\|P_\perp AP_N\|_{H_0^1} = \|P_N AP_\perp\|_{H_0^1} \leqslant LC(N)$. С помощью неравенства треугольника получаем утверждение леммы:

$$\big\|P_\perp A(P_\perp + P_N)w\big\|_{H_0^1} \leqslant C_c\big(C(N)\big)^2\|w_\perp\|_{H_0^1} + LC(N)\|w_N\|_{H_0^1}.$$

Лемма доказана.

Используя обозначения из доказательства теоремы 1 и проведя рассуждения аналогично этому доказательству вплоть до формулы (22), в силу замечания 1 будем иметь оценку $\|\psi_N\|_{H_0^1} \leqslant M\|v_N\|_{H_0^1}$,

где $v_N = P_N A u_\perp$. Далее, с учетом леммы 1а получим $\|\psi_N\|_{H_0^1} \leqslant M\|P_N A u_\perp\|_{H_0^1} \leqslant MC(N)L\zeta$ и, таким образом, $\|Q_N U\|_{H_0^1} \leqslant MC(N)L\zeta$. Далее, в силу леммы 2а имеем

$$\big\|(I-P_N)AU\big\|_{H_0^1} \leqslant C(N)\big(L\|w_N\|_{H_0^1} + C_c C(N)\|u_\perp\|_{H_0^1}\big) \leqslant C(N)\big(L\gamma + C_c C(N)\zeta\big).$$

Тем самым достаточно потребовать (см. (20), (21)), чтобы $MC(N)L\zeta < \gamma$, $C(N)\big(L\gamma + C_c C(N)\zeta\big) < \zeta$.

Полагая $\gamma = MC(N)L\zeta + \varepsilon$ для произвольного $\varepsilon > 0$, имеем достаточное условие разрешимости в виде

$$C(N)\Big[L\big(MC(N)L\zeta + \varepsilon\big) + C_c C(N)\zeta\Big] < \zeta. \tag{41}$$

Таким образом, достаточно потребовать, чтобы выполнялось неравенство

$$\zeta\big[C(N)LMC(N)L + C(N)C_c C(N)\big] < \zeta,$$

поскольку тогда за счет малости выбранного $\varepsilon$ можно добиться, чтобы выполнялось и (41). Итак, мы доказали, что при условии $\theta \equiv \big(C(N)\big)^2(C_c + L^2 M) < 1$ можно гарантировать обратимость оператора $A$, а с ним и оператора исходной задачи $\mathcal{L}$. Отметим, что для рассматриваемого случая $\theta$, вообще говоря, меньше $\kappa$. Теорема доказана.

**Теорема 2а.** *Пусть матрица $G$ обратима. Если $\theta < 1$ и $u$ — решение задачи* (40), *то $u \in X(\Omega)$ и* $\|u - P_N u\|_{H_0^1} \leqslant C(N)\tau\|f\|$.

**Доказательство.** Аналогично доказательству теоремы 2 получим

$$\|u_N\|_{H_0^1} \leqslant M\|P_N A u_\perp + P_N v\|_{H_0^1}, \quad u_\perp = (I-P_N)A(u_N + u_\perp) + (I-P_N)v. \tag{42}$$

Поскольку, в силу леммы 1а, $\|P_N A u_\perp\|_{H_0^1} \leqslant LC(N)\|u_\perp\|_{H_0^1}$, то

$$\|u_N\|_{H_0^1} \leqslant MLC(N)\|u_\perp\|_{H_0^1} + M\|P_N v\|_{H_0^1} \leqslant MLC(N)\|u_\perp\|_{H_0^1} + M\|v\|_{H_0^1}. \tag{43}$$

Оценивая первое слагаемое во второй из формул (42) по лемме 2а, получим:

$$\|u_\perp\|_{H_0^1} \leqslant C(N)\big(L\|u_N\|_{H_0^1} + C_c C(N)\|u_\perp\|_{H_0^1}\big) + \big\|(I-P_N)v\big\|_{H_0^1}.$$

Подставив сюда оценку для $\|u_N\|_{H_0^1}$, найденную в (43), имеем:

$$\|u_\perp\|_{H_0^1} \leqslant C(N)\Big\{LM\big(LC(N)\|u_\perp\|_{H_0^1} + \|v\|_{H_0^1}\big) + C_c C(N)\|u_\perp\|_{H_0^1}\Big\} + \|(I-P_N)v\|_{H_0^1}.$$

Приводим подобные слагаемые и учитываем, что в силу (5) $\big\|(I-P_N)v\big\|_{H_0^1} \leqslant C(N)\|f\|$:

$$\|u_\perp\|_{H_0^1}\Big(1 - C(N)\big(L^2 MC(N) + C_c C(N)\big)\Big) \leqslant C(N)LM\|v\|_{H_0^1} + C(N)\|f\|.$$

Поскольку еще $\|v\|_{H_0^1} \leqslant C_p\|f\|$ (см. формулу (27)), то окончательно

$$\|u_\perp\|_{H_0^1}\Big(1 - \big(C(N)\big)^2(L^2 M + C_c)\Big) \leqslant C(N)LMC_p\|f\| + C(N)\|f\|,$$

откуда, деля на $1 - \theta \equiv 1 - \big(C(N)\big)^2(L^2 M + C_c)$, имеем $\|u_\perp\|_{H_0^1} \leqslant \dfrac{1}{1-\theta}\,C(N)(1 + LMC_p)\|f\|$. Теорема доказана.

**Теорема 3а.** *Пусть матрица $G$ обратима. При произвольном $v \in H_0^1$, если $\theta < 1$, верны оценки*

$$\|v - P_{\mathcal{L}} v\|_{H_0^1} \leqslant \widetilde{\alpha}\|v - P_N v\|_{H_0^1}, \quad \|v - P_{\mathcal{L}} v\| \leqslant C(N)\widetilde{\beta}\|v - P_N v\|_{H_0^1} \leqslant C(N)\widetilde{\beta}\|v - P_{\mathcal{L}} v\|_{H_0^1},$$

*где* $\widetilde{\alpha} = \sqrt{1 + \big(MC(N)L\big)^2}$ и $\widetilde{\beta} = 1 + C_p ML$.

**Доказательство.** Получив в (33) оценку $\|P_{\mathcal{L}} v - P_N v\|_{H_0^1} \leqslant M\|P_N\psi\|_{H_0^1}$, где $\psi = -Ag = \triangle^{-1}\big(k^2(x)g\big)$, в силу леммы 1а имеем далее (вспомним, что $g \in S_N^\perp$):

$$M\|P_N\psi\|_{H_0^1} = M\|P_N Ag\|_{H_0^1} = M\|P_N AP_\perp g\| \leqslant MC(N)L\|g\|_{H_0^1} \equiv MC(N)L\|v - P_N v\|_{H_0^1}. \tag{44}$$

Оценку для $\|v - P_{\mathcal{L}}v\|$ получаем, применив к (44) неравенство Пуанкаре–Фридрихса. Теорема доказана.

**Теорема 4а.** *Пусть матрица $G$ обратима. Если $\theta < 1$, то для точного $u$ и приближенного $P_{\mathcal{L}}u$ решений задачи* (40) *верны оценки*

$$\begin{aligned} &\|u - P_{\mathcal{L}}u\|_{H_0^1} \leqslant C(N)\widetilde{\alpha}\tau\|f\|, \quad \|u - P_{\mathcal{L}}u\| \leqslant \big(C(N)\big)^2\widetilde{\beta}\tau\|f\|, \\ &\|u - P_{\mathcal{L}}u\|_{H_0^1} \leqslant C(N)\widetilde{\beta}\|S\| + \|R\| + \big(C(N)\big)^3\widetilde{\beta}^2C_c\tau\|f\|. \end{aligned} \tag{45}$$

**Доказательство.** Первые две оценки очевидно следуют из двух предыдущих теорем, а последняя получается аналогично последней оценке теоремы 4, но с использованием лемм 1а и 2а вместо лемм 1 и 2. Действительно, получив неравенство $\|e\|_{H_0^1}^2 \leqslant \|e\|\,\|S\| + \|e\|\,\|ce\| + \|e\|_{H_0^1}\|R\|$, имеем

$$\|e\|_{H_0^1}^2 \leqslant \|e\|\big(\|S\| + C_c\|e\|\big) + \|e\|_{H_0^1}\|R\| \leqslant \|e\|_{H_0^1}C(N)\widetilde{\beta}\big(\|S\| + C_c\|e\|\big) + \|e\|_{H_0^1}\|R\|,$$

где мы оценили $\|e\|$ с помощью теоремы 3а. Заметим, что в силу теорем 2а и 3а выполнена оценка $C_c\|e\| \leqslant C_c\big(C(N)\big)^2\widetilde{\beta}\tau\|f\|$, откуда окончательно имеем $\|e\|_{H_0^1} \leqslant C(N)\widetilde{\beta}\|S\| + \|R\| + \big(C(N)\big)^3\widetilde{\beta}^2C_c\tau\|f\|$. Теорема доказана.

**5.3. Алгоритм вычисления приближенного решения и модифицированных оценок погрешности для уравнения Гельмгольца.**

1. Первый этап алгоритма.

а) Оценить сверху или определить аналитически константу Пуанкаре–Фридрихса $C_p$.

б) Исходя из выбранной дискретизации, определить константу $C(N)$.

в) Сформировать матрицы

$$G = \{G_{ji}\} = \Big\{(\nabla\varphi_i, \nabla\varphi_j)_{L^2} - \big(k^2(x)\varphi_i, \varphi_j\big)_{L^2}\Big\}, \quad D = \{D_{ji}\} = \big\{(\nabla\varphi_i, \nabla\varphi_j)_{L^2}\big\}.$$

г) Вычислить вектор-столбец правой части $F = \{F_i\}^{\mathrm{T}} = \big\{\int_\Omega f(x)\varphi_i\,dx\big\}^{\mathrm{T}}$.

д) Получить вектор $V$ коэффициентов разложения приближенного решения по функциям $\{\varphi_i\}_{i=1}^N$ (степеней свободы) как решение линейной системы $GV = F$. Вычислить приближенное решение: $v = \sum\limits_{i=1}^{N} V_i\varphi_i$.

е) Найти константу $C_c = \big\|c(x)\big\|_{L^\infty} \equiv \sup\limits_{x\in\Omega} k^2(x)$.

ж) Вычислить $\lambda_1$ — наименьшее по модулю собственное значение задачи $G\nu = \lambda D\nu$ — и положить $M(N) = |\lambda_1|^{-1}$.

з) Вычислить константы

$$L(N) = \sqrt{\big\|P_NAP_N|_{S_N^0}\big\|_{H_0^1}C_c}\,, \quad \theta(N) = \big(C(N)\big)^2(L^2M + C_c), \quad \tau(N) = \frac{1 + C_pML}{1 - \theta(N)}\,.$$

Остановимся на вычислении операторной нормы $\big\|P_NAP_N|_{S_N^0}\big\|_{H_0^1}$. Это норма симметричного положительно определенного оператора в конечномерном пространстве, и ее можно найти как наибольшее собственное значение задачи $H\nu = \lambda D\nu$, где $H = D - G$.

и) Проверить, что $\theta < 1$.

к) Вычислить константы $\widetilde{\alpha} = \sqrt{1 + \big(MC(N)L\big)^2}$ и $\widetilde{\beta} = 1 + C_pML$.

л) Вычислить оценки погрешности по формулам

$$\|u - P_{\mathcal{L}}u\|_{H_0^1} \leqslant C(N)\widetilde{\alpha}\tau\|f\|_{L^2}, \quad \|u - P_{\mathcal{L}}u\|_{L^2} \leqslant \big(C(N)\big)^2\widetilde{\beta}\tau\|f\|_{L^2}.$$

2. Второй этап алгоритма.

а) Зная вектор $V$ координат (степеней свободы) приближенного решения $v$, вычислить градиент $\nabla v$ приближенного решения.

б) Вычислить матрицу Грама $\widetilde{D}$ функций $\{\widetilde{\varphi}_j\}_{j=1}^M$ и матрицы $W^k$ по формулам $\widetilde{D}_{jl} = (\widetilde{\varphi}_l, \widetilde{\varphi}_j)_{L^2}$ и $(W^k)_{ji} = \left(\widetilde{\varphi}_j, \dfrac{\partial\varphi_i}{\partial x_k}\right)_{L^2}$.

в) Решив линейные системы $\widetilde{D}\overline{V}^k = W^kV$ для каждого $k = 1, \ldots, n$, найти коэффициенты $\overline{V}^k$ в разложении $k$-й компоненты осредненного градиента $\overline{\nabla}v$.

г) Зная функции $\{\widetilde{\varphi}_j\}_{j=1}^M$ и их градиенты, вычислить осредненный градиент

$$\overline{\nabla} v = \left( \sum_{j=1}^{M} (\overline{V}^{\,1})_j \widetilde{\varphi}_j, \ldots, \sum_{j=1}^{M} (\overline{V}^{\,n})_j \widetilde{\varphi}_j \right) \quad \text{и} \quad \operatorname{div}(\overline{\nabla} v) = \sum_{i=1}^{n} \sum_{j=1}^{M} (\overline{V}^{\,i})_j \frac{\partial \widetilde{\varphi}_j}{\partial x_i}.$$

д) Вычислить $\|R\| = \|\nabla v - \overline{\nabla} v\| = \sqrt{\sum_{i=1}^{n} \|R_i\|^2}$ и $\|S\| = \|f + \operatorname{div}(\overline{\nabla} v) + k^2(x) v\|$.

е) Вычислить оценку $\|u - P_{\mathcal{L}} u\|_{H_0^1} \leqslant C(N)\widetilde{\beta}\|S\| + \|R\| + \big(C(N)\big)^3 \widetilde{\beta}^2 C_c \tau \|f\|$.

**5.4. Тестовые расчеты.** В наших численных экспериментах предложенная модификация позволила улучшить все три оценки (т.е. уменьшить переоценку погрешности) примерно в 1,2 ÷ 1,5 раза. Теоретически — при других $k^2(x)$ — это отношение может быть меньше или больше, но заведомо предлагаемая здесь оценка тоньше, чем исходная. В качестве примера приведем результаты расчетов для одномерного уравнения Гельмгольца $u'' + k^2(x)u = -f$ на отрезке $[0;1]$. Мы брали $k$ равным $k_0$ на промежутке $[0;1/2)$ и $2k_0$ на $[1/2;1]$. В первой серии расчетов рассматривался случай $u = x(1-x)/2$; соответственно, правая часть имеет вид $f = 1 - k^2(x)x(1-x)/2$. Во второй серии расчетов рассматривался случай

$$u = \begin{cases} u_0 = \dfrac{\cos\dfrac{k_0}{2} - 1}{k_0^2 \sin\dfrac{k_0}{2}} \sin k_0 x - \dfrac{1}{k_0^2} \cos k_0 x + \dfrac{1}{k_0^2}, & x \in [0;1/2), \\ u_1 = A\left( \dfrac{\cos\dfrac{2k_0}{2} - 1}{(2k_0)^2 \sin\dfrac{2k_0}{2}} \sin 2k_0(x - 1/2) - \dfrac{1}{(2k_0)^2} \cos 2k_0(x - 1/2) + \dfrac{1}{(2k_0)^2} \right), & x \in [1/2;1], \end{cases}$$

где множитель $A$ вводился для выполнения условия сопряжения для производных и был равен

$$\frac{u_0'(1/2)}{u_1'(1/2)} = \frac{\dfrac{\cos\dfrac{k_0}{2} - 1}{k_0 \sin\dfrac{k_0}{2}} \cos\frac{k_0}{2} + \dfrac{1}{k_0} \sin\dfrac{k_0}{2}}{\dfrac{1}{\sin\dfrac{2k_0}{2}} \dfrac{1}{2k_0} \left( \cos\dfrac{2k_0}{2} - 1 \right)}.$$

Соответственно, для правой части имеем $f = -1$ на $[0;1/2)$ и $f = -A$ на $[1/2;1]$.

Серия 1

| $k_0 = 3$ | $N = 50$ | $N = 100$ | $N = 500$ |
|---|---|---|---|
| $L^2$ | $1{,}5165 \times 10^2$ | $1{,}5069 \times 10^2$ | $1{,}5038 \times 10^2$ |
| | $1{,}0786 \times 10^2$ | $1{,}0741 \times 10^2$ | $1{,}0729 \times 10^2$ |
| $(H_0^1)_1$ | $1{,}5330 \times 10^1$ | $1{,}5113 \times 10^1$ | $1{,}5044 \times 10^1$ |
| | $1{,}2865 \times 10^1$ | $1{,}2745 \times 10^1$ | $1{,}2706 \times 10^1$ |
| $(H_0^1)_2$ | $1{,}7523 \times 10^1$ | $1{,}6506 \times 10^1$ | $1{,}6180 \times 10^1$ |
| | $1{,}4618 \times 10^1$ | $1{,}4013 \times 10^1$ | $1{,}3819 \times 10^1$ |

Серия 2а

| $k_0 = 3$ | $N = 50$ | $N = 100$ | $N = 500$ |
|---|---|---|---|
| $L^2$ | $2{,}0257 \times 10^1$ | $2{,}0109 \times 10^1$ | $2{,}0062 \times 10^1$ |
| | $1{,}4401 \times 10^1$ | $1{,}4334 \times 10^1$ | $1{,}4313 \times 10^1$ |
| $(H_0^1)_1$ | $4{,}8647 \times 10^0$ | $4{,}7961 \times 10^0$ | $4{,}7743 \times 10^0$ |
| | $4{,}0827 \times 10^0$ | $4{,}0445 \times 10^0$ | $4{,}0325 \times 10^0$ |
| $(H_0^1)_2$ | $5{,}8113 \times 10^0$ | $5{,}4849 \times 10^0$ | $5{,}3760 \times 10^0$ |
| | $4{,}9562 \times 10^0$ | $4{,}7615 \times 10^0$ | $4{,}6952 \times 10^0$ |

Серия 2б

| $k_0 = 4$ | $N = 50$ | $N = 100$ | $N = 500$ |
|---|---|---|---|
| $L^2$ | $2{,}4792 \times 10^2$ | $2{,}2958 \times 10^2$ | $2{,}2427 \times 10^2$ |
| | $1{,}6326 \times 10^2$ | $1{,}5548 \times 10^2$ | $1{,}5315 \times 10^2$ |
| $(H_0^1)_1$ | $4{,}0356 \times 10^1$ | $3{,}1917 \times 10^1$ | $2{,}9180 \times 10^1$ |
| | $3{,}0098 \times 10^1$ | $2{,}5572 \times 10^1$ | $2{,}4085 \times 10^1$ |
| $(H_0^1)_2$ | $1{,}7579 \times 10^2$ | $7{,}4466 \times 10^1$ | $4{,}4196 \times 10^1$ |
| | $1{,}0764 \times 10^2$ | $5{,}3237 \times 10^1$ | $3{,}6373 \times 10^1$ |

Таблицы устроены следующим образом. Все графы заполнены мерой переоценки ошибки, т.е. отношением оценки к фактической ошибке. Чем она *меньше*, тем лучше оценка. В каждой строке таблицы сверху стоит мера переоценки ошибки для оценки, вычисленной по общему методу Накао, снизу — для

предлагаемой модифицированной оценки. Число $N$ — количество отрезков разбиения. В левом столбце указан тип оценки: $L^2$ означает ошибку в норме $L^2$; $(H_0^1)_1$ означает ошибку в норме $H_0^1$, где оценка вычисляется по первой из формул (35) (по первой из формул (45) для модифицированных оценок); $(H_0^1)_2$ — по формуле (36) (соответственно по последней из формул (45) для модифицированных оценок), т.е. с использованием найденного приближенного решения.

СПИСОК ЛИТЕРАТУРЫ


1. *Марчук Г.И., Агошков В.И.* Введение в проекционно-сеточные методы. М.: Наука, 1981.
2. *Сьярле Ф.* Метод конечных элементов для эллиптических задач. М.: Мир, 1980.
3. *Репин С.И.* Двусторонние оценки отклонения от точного решения для равномерно эллиптических уравнений // Труды Санкт-Петербургского математического общества. Т. 9. Новосибирск: Научная книга, 2001. 148–179.
4. *Репин С.И., Фролов М.Е.* Об апостериорных оценках точности приближенных решений краевых задач для уравнений эллиптического типа // Журн. вычисл. матем. и матем. физики. 2002. **42**, № 12. 1774–1787.
5. Functional a posteriori error estimates for PDE's (`http://www.pdmi.ras.ru/~repin/ApoPDE.pdf`).
6. *Nakao M.T., Hashimoto K., Watanabe Y.* A numerical method to verify the invertibility of linear elliptic operators with applications to nonlinear problems // Computing. 2005. **75**, N 1. 1–14.
7. *Nakao M.T., Hashimoto K.* Constructive error estimates of finite element approximations for non-coercive elliptic problems and its applications (http://hdl.handle.net/2324/3405).
8. *Nakao M.T.* Numerical verification methods for solutions of ordinary and partial differential equations // Numer. Funct. Anal. and Optimiz. 2001. **22**, N 3. 321–356.
9. *Nakao M.T., Yamamoto N., Kimura S.* On the best constant in the error bound for the $H_0^1$-projection into piecewise polynomial spaces // J. Approx. Theory. 1998. **93**, N 3. 491–500.
10. *Natterer F.* Berechenbare Fehlerschranken für die Methode der Finiten Elemente // International Series of Numerical Mathematics. Vol. 28. Basel: Birkhäuser Verlag, 1975. 109–121.
11. *Ладыженская О.А.* Краевые задачи математической физики. М.: Наука, 1973.
12. *Рид М., Саймон Б.* Методы современной математической физики. Т. 1. Функциональный анализ. М.: Мир, 1977.


Поступила в редакцию
10.11.2008