ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 9 за 2023 г.

ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

Б.В. Критский

Применение методов Чебышева для решения эллиптических уравнений на воксельных сетках

Рекомендуемая форма библиографической ссылки: Критский Б.В. Применение методов Чебышева для решения эллиптических уравнений на воксельных сетках // Препринты ИПМ им. М.В.Келдыша. 2023. № 9. 24 с. https://doi.org/10.20948/prepr-2023-9
https://library.keldysh.ru/preprint.asp?id=2023-9

РОССИЙСКАЯ АКАДЕМИЯ НАУК
ОРДЕНА ЛЕНИНА
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М. В. КЕЛДЫША

Б.В. Критский

Применение методов Чебышева для решения эллиптических уравнений на воксельных сетках

Москва, 2023

Б.В. Критский, Применение методов Чебышева для решения эллиптических уравнений на воксельных сетках.

Аннотация В работе рассмотрены различные типы итерационных методов Чебышева для решения сеточных аппроксимаций эллиптических уравнений на воксельных сетках. Рассмотрены подходы к построению данных методов и различные их модификации. Произведено сравнение модификаций методов.

Ключевые слова: воксельные сетки, безматричные методы, итерационные методы Чебышева, эллиптические уравнения

B.V. Kritskiy, Application of Chebyshev methods for solving elliptic equations on voxel meshes.

Abstract The paper considers various types of Chebyshev iteration methods for solution of difference approximations of elliptic equations on voxel meshes. Approaches to the construction of these methods and their various modifications are considered. Comparison of modifications of methods is performed.

Key words and phrases: voxel meshes, matrix-free methods, Chebyshev iteration methods, elliptic equations

1 Введение

Целью данной работы является численное исследование сходимости и эффективности применения методов Чебышева для решения эллиптических уравнений с неоднородными коэффициентами на воксельных сетках. В работе рассматриваются уравнения вида

$$\nabla \cdot (k \nabla \boldsymbol{v}) = \boldsymbol{f} \tag{1}$$

с кусочно-постоянным полем коэффициента k в области Ω с однородными граничными условиями Дирихле. После аппроксимации задачи с помощью метода конечных разностей во всех внутренних узлах (i, j, k) сетки

$$\begin{aligned} &\frac{k_{i+1/2,j,k}u_{i+1,j,k} - (k_{i+1/2,j,k} + k_{i-1/2,j,k})u_{i,j,k} + k_{i-1/2,j,k}u_{i-1,j,k}}{h^2} + \\ &+ \frac{k_{i,j+1/2,k}u_{i,j+1,k} - (k_{i,j+1/2,k} + k_{i,j-1/2,k})u_{i,j,k} + k_{i,j-1/2,k}u_{i,j-1,k}}{h^2} + \\ &+ \frac{k_{i,j,k+1/2}u_{i,j,k+1} - (k_{i,j,k+1/2} + k_{i,j,k-1/2})u_{i,j,k} + k_{i,j,k-1/2}u_{i,j,k-1}}{h^2} = f_{i,j,k}, \end{aligned}$$

где $k_{i+1/2,j,k}$ — значение поля коэффициентов в точке $(x_{i+1/2}, y_j, z_k)$, $x_{i+1/2} = (x_i + x_{i+1})/2$ и т.п., она сводится к системе линейных уравнений

$$\boldsymbol{A}\boldsymbol{u} = \boldsymbol{b}. \tag{2}$$

Отметим, что при итерациях чебышевские методы требуют только матрично-векторного умножения и в них практически не нужно вычислять скалярные произведения. Соответственно, они могут быть эффективно распараллелены, поскольку имеют малое число точек синхронизации. Также они позволяют достичь высокой точности решения, поскольку ошибки округления эффективно подавляются в ходе итераций. Еще одним достоинством данных методов является то, что они хорошо подходят для безматричной реализации.

В первой части работы рассмотрены два подхода к построения чебышевских методов для решения систем линейных уравнений, приведены результаты анализа ошибок методов и их устойчивости, а также результаты проведенного численного исследования скорости работы методов. Во второй части приведены варианты модификации методов Чебышева и проведено их сравнение на модельной задаче.

2 Итерационный метод Чебышева

В данном разделе приведем два варианта метода Чебышева для произвольных невырожденных матриц $\boldsymbol{A}$ с действительными коэффициентами. В пер-

вом случае будем считать спектр матрицы действительным, во втором — комплексным.

Случай действительного спектра. Первый способ построения, применимый в случае действительного положительного спектра у $\boldsymbol{A}$, можно найти в работе [1]. Далее опишем логику построения данного метода.

Зададим форму итерационного процесса как

$$\boldsymbol{u}_{k+1} = \boldsymbol{u}_k + \tau_{k+1}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{u}_k), \quad k = 0, 1, .., p-1,$$

где τ_k — параметры метода, p — количество итераций, $\boldsymbol{u}_k$ — приближение к решению на шаге k.

Рассмотрим теперь, как ведет себя ошибка $\boldsymbol{e}_p = \boldsymbol{u}_p - \boldsymbol{u}_*$ ($\boldsymbol{u}_*$ — точное решение (2)) в процессе итераций. Несложно получить

$$\boldsymbol{e}_p = F_p(\boldsymbol{A})\boldsymbol{e}_0, \tag{3}$$

где

$$F_p(\boldsymbol{A}) = \prod_{k=0}^{p-1}(\boldsymbol{I} - \tau_k \boldsymbol{A}).$$

В силу коммутативности $F_p(\boldsymbol{A})$ и $\boldsymbol{A}$ невязка $\boldsymbol{r}_p = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{u}_p$ ведет себя как ошибка, согласно (3), то есть $\boldsymbol{r}_p = F_p(\boldsymbol{A})\boldsymbol{r}_0$. Используя оператор F_p, результат итераций можно переписать следующим образом:

$$\boldsymbol{u}_p = F_p(\boldsymbol{A})\boldsymbol{u}_0 + [\boldsymbol{I} - F_p(\boldsymbol{A})]\,\boldsymbol{A}^{-1}\boldsymbol{b}.$$

Выберем значения параметров τ_k так, чтобы ошибка была минимальна после заданного числа шагов. Рассмотрим оценку [2]

$$\|F_p(\boldsymbol{A})\| \leqslant \|F_p(\boldsymbol{t})\| = \max_{\lambda_{\min} \leqslant t \leqslant \lambda_{\max}} \left| \prod_{k=0}^{p-1}(1 - \tau_k t) \right|, \tag{4}$$

где $\|F_p(\boldsymbol{A})\|$ обозначает норму оператора $F_p(\boldsymbol{A})$ в $\mathbb{R}^3$, $\lambda_{\min}$ и $\lambda_{\max}$ — минимальное и максимальное собственные значения матрицы $\boldsymbol{A}$, в правой части неравенства стоит полином $F_p(t)$. Теперь задачу о минимизации ошибки можно переформулировать в виде задачи о нахождении полинома указанного выше вида $F_p(t)$, принимающего минимальное среди всех полиномов максимальное по модулю значение на отрезке $[\lambda_{\min}, \lambda_{\max}] \subset \mathbb{R}^+$ («минимально» уклоняющегося от нуля). Данный полином будет иметь вид [1]

$$F_p(t) = T_p\left(\frac{1 - t/\lambda_{\text{ave}}}{\rho_0}\right) \Big/ T_p(1/\rho_0), \tag{5}$$

где

$$\eta = \frac{\lambda_{\min}}{\lambda_{\max}}, \quad \lambda_{\text{ave}} = \frac{\lambda_{\min} + \lambda_{\max}}{2}, \quad \rho_0 = \frac{1+\eta}{1-\eta}, \quad \rho_1 = \frac{1+\sqrt{\eta}}{1-\sqrt{\eta}},$$

а T_p — полином Чебышева степени p, который определяется как

$$T_p(\lambda) = \begin{cases} \cos(p \arccos t), & |t| \in [-1,1], \\ \cosh(p \operatorname{arccosh} t), & |t| \notin [-1,1]. \end{cases}$$

Корни μ_k полинома Чебышева имеют вид

$$\mu_k = \cos\frac{\pi(2k+1)}{2p}, \quad k = 0, .., p-1.$$

Отсюда для итерационных параметров имеем $\tau_k = 1/[\lambda_{\text{ave}}(1 - \rho_0\mu_k)]$.

Случай комплексного спектра. Перейдем к рассмотрению метода в случае наличия комплексных собственных значений. Рассмотрим итерационный процесс в виде

$$\boldsymbol{u}_{k+1} = \boldsymbol{u}_k + \sum_{i=0}^{k} \gamma_{k+1,i}\boldsymbol{r}_i, \quad k = 0, 1, .., p-1.$$

Тогда, как и в первом случае, ошибка метода удовлетворяет (3) и требуется найти такой $F_p(A)$ [3], что:

1. $F_p(0) = 1$;
2. $F_p(z)$ «минимально» уклоняется от нуля на спектре $\boldsymbol{A}$;
3. Если $\boldsymbol{A}$ имеет нелинейные простые делители (т.е. характеристический многочлен содержит кратные или комплексные корни), то производные $F_p(z)$ также должны быть малыми, как и сам $F_p(z)$.

По сути задача состоит в нахождении чебышевского полинома согласно следующему определению.

Определение 2.1. *Для некоторой области E комплексной плоскости $\mathbb{C}$, содержащей спектр $\boldsymbol{A}$, полином $\hat{T}_p(z)$ степени p, удовлетворяющий свойству 1 и свойству 2 в смысле*

$$\max_{z\in E} |\hat{T}_p(z)| \leqslant \min_{Q_p} \max_{z\in E} |Q_p(z)|,$$

где минимум берется по всем полиномам степени p со старшим коэффициентом 1*, называют полиномом Чебышева.*

Для произвольной области E комплексной плоскости полиномы Чебышева можно пытаться строить из определения, но в случае, когда область E представляет из себя эллипс, можно воспользоваться результатом [4]. Отметим, что если E представляет из себя дугу эллипса, то в [5] получен точный результат. Согласно [4] для эллипсов $E = F(d, c)$ с центром в d и фокусами в $d \pm c$ (тут c и d могут быть комплексными), не содержащих 0, полином F_p определяется с помощью замены переменных как

$$F_p(z) = T_p\left(\frac{d-z}{c}\right) \Big/ T_p\left(\frac{d}{c}\right), \tag{6}$$

где

$$T_p(z) = \cosh(p \cdot \operatorname{acosh}(z)).$$

Полином (6) не будет является полиномом Чебышева в смысле определения (2.1). Тем не менее он будет обладать свойством оптимальности в пределе [3]:

$$m(F_p(z)) \leqslant M(Q_p) \leqslant M(F_p(z)), \quad p \to \infty,$$

где

$$m(Q_p) = \min_{z \in E} |Q_p(z)|, \quad M(Q_p) = \max_{z \in E} |Q_p(z)|.$$

Более того,

$$\lim_{p\to\infty} M(Q_p)^{1/p} = \lim_{p\to\infty} M(F_p)^{1/p}.$$

Теперь, если потребовать, чтобы полином ошибок F_p определялся (6) на каждой итерации, то имеем рекуррентное соотношение

$$T_{p+1}\left(\frac{d-\lambda}{c}\right) = 2\frac{d}{c}T_p\left(\frac{d-\lambda}{c}\right) - 2\frac{\lambda}{c}T_p\left(\frac{d-\lambda}{c}\right) - T_{p-1}\left(\frac{d-\lambda}{c}\right). \tag{7}$$

Соответственно домножив T_p на $\boldsymbol{e}_0$ и имея в виду $\boldsymbol{e}_p = F_p(A)\boldsymbol{e}_0$, получим:

$$\boldsymbol{e}_{p+1} = \left[2\frac{d}{c}T_p\left(\frac{d}{c}\right)\boldsymbol{e}_p - \frac{2}{c}T_p\left(\frac{d}{c}\right)\boldsymbol{A}\boldsymbol{e}_p - T_{p-1}\left(\frac{d}{c}\right)\boldsymbol{e}_{p-1}\right] \Big/ T_{p+1}\left(\frac{d}{c}\right).$$

Аналогичное тождество выполнено и для невязки $\boldsymbol{r}_p$. Из последнего соотношения можно получить следующую форму итерационного метода:

$$\boldsymbol{u}_{p+1} = \boldsymbol{u}_p + \boldsymbol{\Delta}_p,$$

$$\boldsymbol{\Delta}_p = (-\boldsymbol{r}_p + \beta_{p-1}\boldsymbol{\Delta}_{p-1})/\gamma_{p-1},$$

где

$$\gamma_p = \frac{c}{2}\, T_{p+1}\left(\frac{d}{c}\right) \Big/ T_p\left(\frac{d}{c}\right) = -(\beta_{p-1} - d),$$

$$\beta_{p-1} = \frac{c}{2} T_{p-1}\left(\frac{d}{c}\right) \Big/ T_p\left(\frac{d}{c}\right) = \frac{c}{2}\frac{1}{\gamma_p}.$$

В качестве начальных значений берутся

$$\beta_{-1} = 0, \quad \beta_0 = \frac{c}{2}\frac{1}{\eta}, \quad \gamma_0 = d.$$

Осталось показать как выбираются параметры d и c. Согласно [3], они могут быть определены как

$$\min_{d,c} \max_{\lambda_p} r(\lambda_p) = \min_{d,c} \max_{\lambda_p} \left| \frac{(d-\lambda_p) + ((d-\lambda_p)^2 - c^2)^{1/2}}{d + ((d)^2 - c^2)^{1/2}} \right|, \tag{8}$$

где λ_p — собственные значения $\boldsymbol{A}$, а $r(z)$ определяется как

$$r(z) = \lim_{p\to\infty} |F_p(z)^{1/p}| = \left| \exp\left(\operatorname{acosh}\left(\frac{d-z}{c}\right) - \operatorname{acosh}\left(\frac{d}{c}\right)\right)\right|, \quad z \in \mathbb{C}.$$

В частном случае, когда все коэффициенты матрицы $\boldsymbol{A}$ действительны, данный минимум (8) можно искать среди действительных d и c^2, поскольку спектр $\boldsymbol{A}$ будет симметричен относительно действительной оси. А максимум, соответственно, можно искать среди собственных значений с неотрицательной мнимой частью.

Для решения системы уравнений для обоих вариантов метода Чебышева необходимо знать границы спектра матрицы. Поскольку вычисление каждого из значений спектра является задачей по сложности превосходящей решение самой системы, то на практике используются оценки для собственных значений $\boldsymbol{A}$, уточняемые в процессе решения системы. Данный метод входит в класс адаптивных чебышевских методов.

3 Оценки точности методов

В предыдущем разделе были коротко рассмотрены формулировки и принципы построения чебышевских итерационных методов для матриц с действительным и комплексным спектром. В данном разделе рассмотрим результаты, касающиеся оценки их точности.

В случае первого метода непосредственно из формулы для полинома ошибок следует, что

$$\sup_{\lambda\in[\lambda_{\min},\lambda_{\max}]} |F_p(\lambda)| = 1/\cosh(p\cdot \operatorname{acosh}(1/\rho_0)).$$

Если теперь обозначить через ϵ желаемую точность ($\|\boldsymbol{r}_p\| \leqslant \epsilon\|\boldsymbol{r}_0\|$), то получим оценку количества итераций (степени полинома Чебышева), необходимых для ее достижения:

$$p \approx \frac{\ln(\epsilon^{-1} + \sqrt{\epsilon^{-2} - 1})}{\ln\left[(1 + \sqrt{\eta})/(1 - \sqrt{\eta})\right]}, \tag{9}$$

где для вывода использовано тождество $\operatorname{acosh}(x) = \ln(x + \sqrt{x^2 - 1}), \quad x \geqslant 1$.

Для второго метода оценку точности при $d \neq 0$ и $\|c\|^2 < \|d\|^2$ можно найти в [6]. Для ее получения полином F_p представляется в виде

$$F_p(\lambda) = \left(\frac{\exp\left(\operatorname{acosh}((d-\lambda)/c)\right)}{\exp\left(\operatorname{acosh}(d/c)\right)}\right)^p \cdot \left(\frac{1 + \exp\left(\operatorname{acosh}(-2p(d-\lambda)/c)\right)}{1 + \exp\left(\operatorname{acosh}(-2pd/c)\right)}\right).$$

Обозначим

$$S(\lambda) = \frac{\exp(\operatorname{acosh}((d-\lambda)/c))}{\exp(\operatorname{acosh}(d/c))} = \frac{d - \lambda + ((d-\lambda)^2 - c^2)^{1/2}}{d + (d^2 - c^2)^{1/2}},$$

и

$$Q_p(\lambda) = \frac{1 + \exp(\operatorname{acosh}(-2p(d-\lambda)/c))}{1 + \exp(\operatorname{acosh}(-2pd/c))}.$$

Тогда имеем

$$F_p(\lambda) = S^p(\lambda) Q_p(\lambda).$$

При $(d-\lambda)/c$, не принадлежащем отрезку $[-1, 1] \subset \mathbb{C}$ комплексной плоскости, для Q_p оказывается верным следующее:

$$\lim_{p \to \infty} Q_p(\lambda) = 1.$$

Если же наоборот, $(d - \lambda)/c \in [-1, 1]$, то

$$0 \leqslant |Q_p(\lambda)| \leqslant 2/\left[1 - \exp(-p\operatorname{Re}(\operatorname{acosh}(d/c)))\right].$$

В [6] замечено, что в точках, где Q_p не стремится быстро к 1, $S^p(\lambda)$ убывает с наибольшей скоростью. Поэтому можно сказать, что для больших p сходимость метода определяется $S^p(\lambda)$ вместо $F_p(\lambda)$. Отсюда можно получить необходимое количество итераций для достижения заданной точности. В случае действительных c и d она будет совпадать с (9) в силу одинакового вида полиномов ошибки. Также в [6] приведен вид ошибки, возникающей в процессе итераций метода. Из разложения вектора ошибки $\boldsymbol{e}_0$ по векторам $\boldsymbol{w}_i$ инвариантных подпространств оператора $\boldsymbol{A}$ имеем

$$\boldsymbol{e}_0 = \sum_{i=1}^{m} \alpha_i \boldsymbol{w}_i,$$

где $\boldsymbol{w}_i$ — элемент инвариантного подпространства $\boldsymbol{A}$ размерности d_i и

$$\boldsymbol{w}_i = \sum_{j=1}^{d_i} \beta_i \boldsymbol{v}_{ij},$$

с соотношениями для базиса

$$(\boldsymbol{A} - \lambda_i)\boldsymbol{v}_{i1} = 0,$$

$$(\boldsymbol{A} - \lambda_i)\boldsymbol{v}_{ij} = \boldsymbol{v}_{ij-1},$$

где $\boldsymbol{v}_{ij}$ — вектор базиса инвариантного подпространства, отвечающего собственному значению λ_i. Отсюда имеем

$$\boldsymbol{e}_p = F_p(\boldsymbol{A})\boldsymbol{e}_0 = \sum_{i=1}^{m} \alpha_i F_p(\boldsymbol{A})\boldsymbol{w}_i,$$

$$F_p(\boldsymbol{A})\boldsymbol{w}_i = \sum_{j=1}^{d_i} \left(\sum_{k=0}^{d_i - j} \beta_{j+k} \frac{F_p^{(k)}(\lambda_i)}{k!} \right) \boldsymbol{v}_{ij}.$$

Далее используя

$$F_p^{(k)}(\lambda) = p^k S(\lambda)^n B_k(\lambda, n),$$

получим

$$\boldsymbol{e}_p = \sum_{i=1}^{m} \alpha_i S(\lambda_i)^p \left(\sum_{j=1}^{d_i} \left(\sum_{k=1}^{d_i - j} \beta_{j+k} \frac{p^k B_k(\lambda_i, p)}{k!} \right) \boldsymbol{v}_{ij} \right). \tag{10}$$

Отсюда видно, что вектор ошибок при больших p будет с хорошей точностью принадлежать некоторому собственному подпространству $\boldsymbol{A}$. Если же спектр матрицы $\boldsymbol{A}$ действительный, то ошибка будет приближенно пропорциональна некоторому собственному вектору, не обязательно отвечающему минимальному собственному значению. Используя вид данной ошибки для двух вариантов метода Чебышева, можно предложить варианты адаптации для ускорения сходимости, это делается в пункте 9 данной работы.

4 Процедура адаптации: случай действительного спектра

Для первого варианта метода процедура адаптации описана в работе [1] (также похожая идея рассмотрена в [7]). Кратко опишем тут, в чем заключается ее суть.

Как уже было отмечено выше, для применения чебышевского метода необходимо знать оценку максимального и минимального собственных значений. Для максимального собственного значения нужная оценка получается с помощью оценки Гершгорина [8]. Минимальное собственное число так оценить уже не получается. В результате применения итерационного метода, невязка $\boldsymbol{r}_p$ оказывается с хорошей точностью пропорциональной некоторому собственному значению, находящемуся в гиперболической области полинома F_p [1, 7]. Тогда имеются два простых варианта для оценки минимального собственного числа, отвечающего $\boldsymbol{r}_p$.

Первый дается неравенством Рэлея

$$\lambda_{\min} \leqslant \frac{\langle \boldsymbol{A}\boldsymbol{r}_p, \boldsymbol{r}_p \rangle}{\langle \boldsymbol{r}_p, \boldsymbol{r}_p \rangle}.$$

Второй следует из соотношения

$$F_p(\boldsymbol{A})\boldsymbol{r}_p \approx F_p(\lambda_p)\boldsymbol{r}_p = F_p(\lambda_p)F_p(\boldsymbol{A})\boldsymbol{r}_0.$$

В этом случае можно положить $\lambda_{\min} \approx \lambda^*$, где λ^* определяется из решения

$$\delta = \frac{\|r_p\|}{\|r_0\|} = |F_p(\lambda^*)|.$$

Метод адаптации, использующий второй способ оценки приведен в работе [6]. Приведем тут выражение для λ^* в явном виде. Используя соотношение для F_p, приведенное выше и соотношение $\operatorname{acosh} x = \ln(x + \sqrt{x^2 - 1})$, найдем

$$\begin{aligned} y_1 &= \ln(y_0 + \sqrt{y_0^2 - 1})/p, \\ y_0 &= T_p(1/\rho_0)\delta, \\ \lambda^* &= (1 - \rho_0 \cosh y_1)\lambda_{\text{ave}}. \end{aligned} \tag{11}$$

Более подробно данный вопрос адаптации рассмотрен в [1].

5 Процедура адаптации: случай комплексного спектра

В данном разделе для полноты изложения опишем основные идеи, применяемые в процедуре адаптации в случае комплексного спектра. В постановке метода Чебышева [6, 9] полином ошибок, как отмечалось выше, приближенно записывается в виде

$$F_p(\lambda) = S(\lambda)^p,$$

где

$$S(\lambda) = \frac{d - \lambda + \left[(d-\lambda)^2 - c^2\right]^{1/2}}{d + (d^2 - c^2)^{1/2}},$$

параметры d, c выбираются из условия

$$\min_{d,c} \max_{\lambda} |S(\lambda)|,$$

где λ принадлежат спектру $\boldsymbol{A}$. Рассмотрим выражение для ошибки (10). Допустим, что собственные значения, отвечающие наибольшему вкладу в ошибку, имеют кратность 1. Тогда невязку можно представить в виде линейной комбинации

$$\boldsymbol{r}_p = \beta_1 \boldsymbol{v}_1 + \bar{\beta}_1 \bar{\boldsymbol{v}}_1 + \beta_2 \boldsymbol{v}_2 + \bar{\beta}_2 \bar{\boldsymbol{v}}_2 + \boldsymbol{\epsilon}_p,$$

где собственные вектора $\boldsymbol{v}_1, \bar{\boldsymbol{v}}_1, \boldsymbol{v}_2, \bar{\boldsymbol{v}}_2$ соответствуют минимальным собственным значениям $\lambda_1, \bar{\lambda}_1, \lambda_2, \bar{\lambda}_2$, причем

$$|S(\lambda_1)| \geqslant |S(\lambda_2)| > |S(\lambda_i)|, \quad i > 2.$$

Отсюда получаем

$$\boldsymbol{A}^j \boldsymbol{r}_p = \lambda_1^j \beta_1 \boldsymbol{v}_1 + \bar{\lambda}_1^j \bar{\beta}_1 \bar{\boldsymbol{v}}_1 + \lambda_2^j \beta_2 \boldsymbol{v}_2 + \bar{\lambda}_2^j \bar{\beta}_2 \bar{\boldsymbol{v}}_2 + \boldsymbol{A}^j \boldsymbol{\epsilon}_p.$$

Пусть теперь полином $\rho_4(z) = z^4 + \rho_3 z^3 + \rho_2 z^2 + \rho_1 z + \rho_0$ имеет корни $\lambda_1, \bar{\lambda}_1, \lambda_2, \bar{\lambda}_2$, тогда имеем

$$\|\boldsymbol{A}^4 r_p + \rho_3 \boldsymbol{A}^3 r_p + \rho_2 \boldsymbol{A}^2 r_p + \rho_1 \boldsymbol{A} r_p + \rho_0 r_p\| \approx 0.$$

Из последнего выражения с помощью минимизации можно найти коэффициенты ρ_i, и, соответственно, корни λ_1 и λ_2, которые будут являться приближениями к собственным значениям матрицы $\boldsymbol{A}$. В такой формулировке для нахождения собственных значений потребуется вычисление рада скалярных произведений, впрочем в [6] показано, что от этой неприятной процедуры можно избавиться. Также можно обойтись вычислением только одного скалярного произведения, если воспользоваться рекуррентным соотношением для невязок (16), что даст еще один вариант для оценки собственных значений.

6 Численные ошибки: случай действительного спектра

При решении системы уравнений с помощью итерационного чебышевского метода может происходить накопление численных ошибок, в результате чего

вычислительный процесс теряет устойчивость. Для борьбы с этим требуется переупорядочить операторы $(\boldsymbol{I} - \tau_k \boldsymbol{A})$ определенным образом. В [2] приведен алгоритм переупорядочивания. Приведем тут другой его вариант, дающий тот же самый результат, только формулируемый несколько проще.

```
TAU-transpose(a, n)
 1 if n = 1 then
 2     a[0] = 0
 3     return a
 4 if n%2 = 0 then
 5     m = n/2
 6     TAU-transpose(a, m)
 7 else
 8     m = (n − 1)/2
 9     a[n − 1] = m
10     TAU-transpose(a, m)
11 i = m − 1
12 while i ⩾ 0 do
13     a[2i] = a[i]
14     a[2i + 1] = n − 1 − a[i]
15     i = i − 1
16 return a
```

Алгоритм 1: Переупорядочивание собственных значений

где a — массив целых чисел длины n.

7 Численные ошибки: случай комплексного спектра

Численные ошибки могут сказываться на сходимости метода при рекуррентном вычислении невязки $\boldsymbol{r}_p$, простейшим способом избавиться от данного недостатка является ее явное вычисление $\boldsymbol{r}_p = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{u}_p$.

В работе [9] дан хороший обзор различных способов организации итерационных процессов чебышеских методов, рассмотрены три пары различных вариантов метода. Также из приведенных вариантов видно, что явное вычисление невязки не приводит к увеличению вычислительной сложности алгоритма и к росту количества итераций.

8 Результаты расчетов

В данном разделе приведены результаты численных расчетов с помощью адаптивного метода Чебышева (в варианте для действительного спектра). Ниже приведен вид алгоритма адаптивного метода Чебышева.

```
ChebAdapt()
1 u := b
2 Оценить λ_min и λ_max по Рэлею и Гершгорину [8].
3 // циклы адаптации
4 while i ⩽ MaxIt do
5     Решить методом Чебышева Au = b.
6     if ||b − Au|| < ε then
7         return u
8     else
9         Оценить минимальное собственное значение по (11).
```

Алгоритм 2: Метод Чебышева с адаптацией

Количество итераций внутреннего метода Чебышева (шаг 5 алгоритма) выбиралось из требования убывания ошибки на 3 порядка, согласно (9). Расчеты проводились на регулярных кубических сетках, матрица линейной системы получена из конечно-разностной аппроксимации оператора Лапласа. Рассматривалась задача в кубической области с условиями Дирихле на границе.

В таблице 1 приведены время, количество итераций и оценка минимального собственного значения в зависимости от размеров сетки. Расчеты проводились на 28 ядрах с использованием MPI.

В таблице 2 приведено все то же самое, что и в таблице 1, только для других размерностей и на 560 ядрах. Количество циклов адаптации MaxIt в первом и втором случаях равно 3.

В таблице 3 приведены результаты расчетов на 560 ядрах для оператора Пуассона с кусочно-постоянными коэффициентами (область разделена на две половины, коэффициент теплопроводности отличается в 1000 раз). Количество циклов адаптации при этом равно 7.

Количество итераций, вообще говоря, зависит от диапазона, в котором находятся собственные значения. Соответственно для ускорения работы методов его следует уменьшить тем или иным способом. Особенно это важно при большом разбросе коэффициентов, поскольку в этом случае количество итераций сильно возрастает.

Размер сетки	Время счета (с)	Кол-во итераций	$\lambda_{\min}$
20^3	0.31918	152	0.0677024
40^3	1.53832	295	0.0179109
80^3	18.0568	584	0.00457295
160^3	274.755	1162	0.00115477

Таблица 1. Результаты расчетов для оператора Лапласа с постоянными коэффициентами на 560 ядрах.

Размер сетки	Время счета (с)	Кол-во итераций	$\lambda_{\min}$
320^3	222.05	2323	0.00028854
500^3	1661.66	3628	0.000118588
640^3	4132.10	4638	7.23E-05

Таблица 2. Результаты расчетов для оператора Лапласа с постоянными коэффициентами на 560 ядрах.

Размер сетки	Время счета (с)	Кол-во итераций
320^3	6008.19	80198

Таблица 3. Результаты расчетов для оператора Лапласа с постоянными коэффициентами на 28 ядрах.

9 Варианты уменьшения количества итераций

Методы Чебышева просты в реализации, но вместе с тем используют мало информации об особенности задачи. Поэтому возникает вопрос, нельзя ли используя больше информации, ускорить сходимость данного класса методов. В данном разделе проведем проверку ряда идей для ускорения сходимости чебышевских методов. Из формы ошибки ясно, что в обоих вариантах метода Чебышева, приведенных в разделах 2 и 3 количества итераций, необходимых для достижения заданной точности, должны быть равны, но из-за численных ошибок в варианте для действительного спектра количество итераций оказывается насколько меньшим. Из-за равенства единице полинома ошибок в нуле оказывается, что при наличии в окрестности нуля собственных значений матрицы $\boldsymbol{A}$ одновременно с наличием большого максимального собственного значения получаем быстрый рост числа итераций, согласно (9). Далее рассмотрим три варианта модификации метода Чебышева.

9.1 Вариант 1

Здесь используется идея адаптации полинома ошибок к нижней части спектра. Пример такой адаптации можно найти в [10], где спектр является сосредоточенным на двух отрезках и требуется найти полином наилучшего приближения на их объединении. В последней работе нужный полином ищется в виде произведения двух чебышевских полиномов

$$F_p(t) = P_q(t) \cdot Q_r(t), \tag{12}$$

где полиномы P_q и Q_r равны 1 в нуле и степень второго много больше степени первого, чтобы скомпенсировать ошибку вносимую первым полиномом на высоких гармониках. Тут появляется еще одна идея: если вариант чебышевского метода для комплексного спектра на каждой итерации дает ошибку, отвечающую чебышевскому полиному соответствующей степени на отрезке, то почему бы не составить из последовательности приближений к решению на итерациях новое, отвечающее полиному Чебышева для объединения отрезков, содержащихся в данном. Для этого потребуется, по сути, найти

$$F(t) = \underset{\sum \alpha_i = 1}{\arg\min} \left| \sum_{i=0}^{n} \alpha_i T_i(t)/T_i(0) \right|, \quad t \in [a_1, b_1] \cup [a_2, b_2], \tag{13}$$

где T_i — полиномы Чебышева на отрезке $[a_1, b_2]$. Можно упростить задачу и искать минимум в интегральной норме L_2. Тогда нужно минимизировать

$$F(t) = \underset{\sum \alpha_i = 1}{\arg\min} \sum_{i,j=0}^{n} \frac{\alpha_i \alpha_j}{T_i(0) T_j(0)} \int_{a_1}^{b_2} T_i(t) T_j(t) w(t) dt, \tag{14}$$

где $w(t)$ — некоторая весовая функция, например, характеристическая функция $\xi_X(t)$ для области $X = [a_1, b_1] \cup [a_2, b_2]$. На рисунках 1 и 2, изображен полином Чебышева $T_{10}(t)/T_{10}(0)$ и адаптированный полином $F_{10}(t)$ с весом

$$\xi_X(t)/\sqrt{(1 - ((t-d)/c)^2)},$$

где $d = 0.65$, $c = 0.217$ с $X = [0.3, 0.517] \cup [0.783, 1]$ (на 1 и 2 изображены одни и те же полиномы, только в разном диапазоне).

Тогда новое приближение к решению системы $\boldsymbol{Au} = \boldsymbol{b}$ представится в виде

$$\hat{\boldsymbol{u}} = \sum_{i=0}^{n} \alpha_i \boldsymbol{u}_i, \tag{15}$$

что также верно и для невязки $\hat{\boldsymbol{r}}$. Тут α_i находятся из условия минимизации (14).

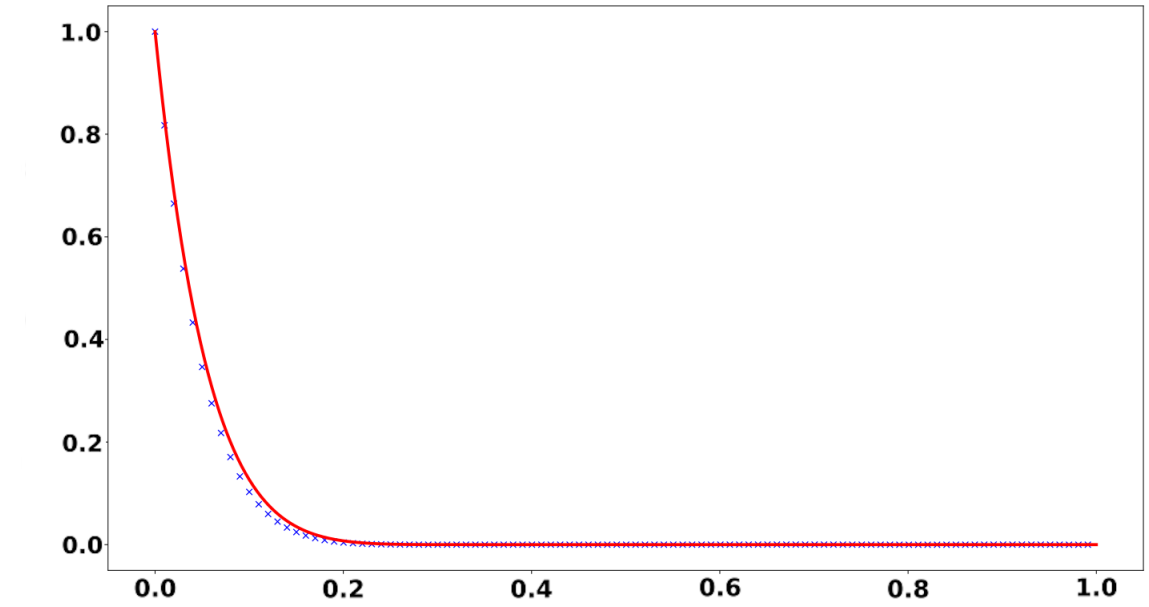

Рис. 1. Красным цветом обозначен полином Чебышева 10-й степени на отрезке $[0.3, 1]$, синим — адаптированный полином

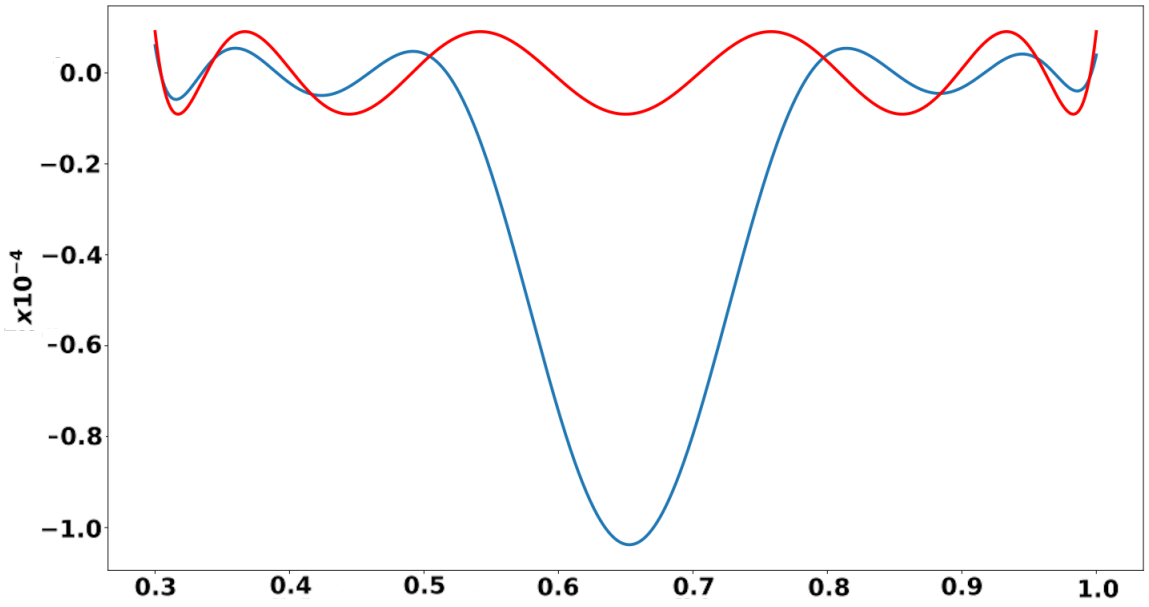

Рис. 2. Красным цветом обозначен полином Чебышева 10-й степени на отрезке $[0.3, 1]$, синим — адаптированный полином

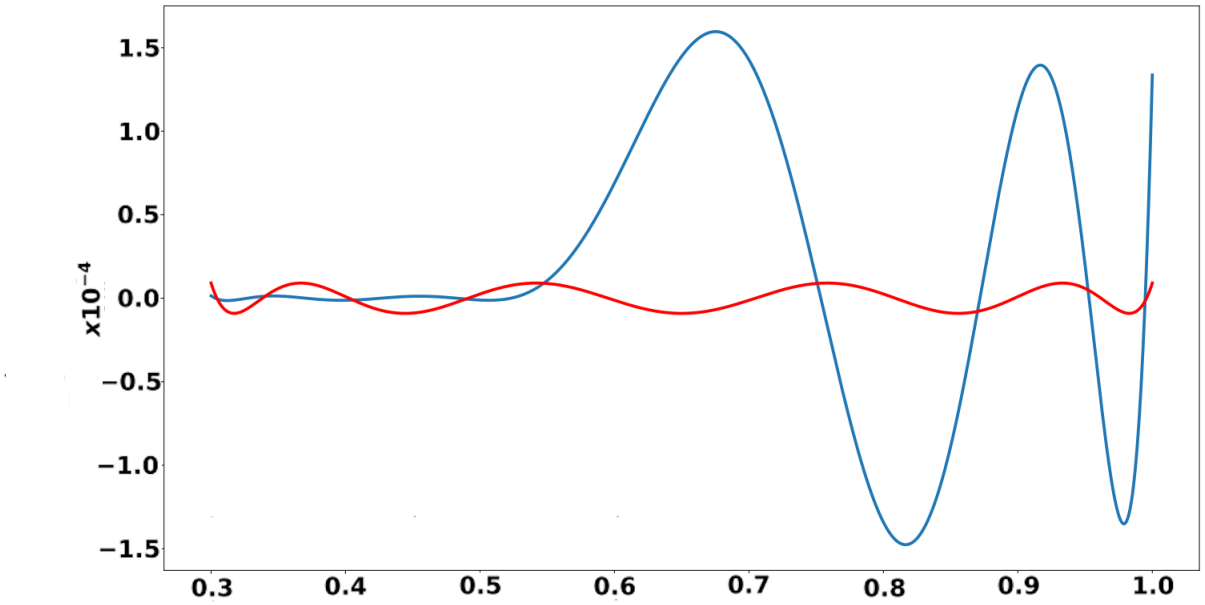

Рис. 3. Красным цветом обозначен полином Чебышева 10-й степени на отрезке $[0.3, 1]$, синим — адаптированный к нижней части спектра полином

К сожалению, в нашем случае (эллиптического $\boldsymbol{A}$) спектр явно не разделяется на отрезки. Тем не менее можно использовать данный результат. Например, адаптировать полином к нижней части спектра и далее решение, отвечающее данному полиному, использовать в качестве базисного элемента подпространства для поиска лучшего приближения. Например, полином, отвечающий весу $(10^4\xi_{[0.3,0.517]}(t) + \xi_{[0.3,1]}(t))/\sqrt{(1-((t-d)/c)^2)}$ приведен на рисунке 3.

В итоге при решении задачи методом Чебышева, складывая решения с соответствующими коэффициентами α_i, можно получать решения $\hat{\boldsymbol{u}}$ соответствующие различным полиномам ошибок. После этого решение ищется в подпространстве, образованном решением и невязкой из обычного метода Чебышева и невязкой, соответствующей полиному, адаптированному к определенной части спектра. Итогом данной процедуры является уменьшение количества итераций, необходимых для достижения заданной точности, как показано на рисунке 5.

Получается, что за один проход метода Чебышева можно получить два решения, соответствующие различным полиномам ошибок, комбинируя которые можно получить уменьшение количества итераций.

9.2 Вариант 2

Попробуем построить подпространство Крылова на векторах невязок и искать приближенное решение в данном подпространстве. Воспользовавшись соотношением

$$\boldsymbol{A}\boldsymbol{r}_p = (\boldsymbol{r}_p - \boldsymbol{r}_{p+1} - \beta_p(\boldsymbol{r}_{p-1} - \boldsymbol{r}_p))/\alpha_p, \tag{16}$$

где $\boldsymbol{r}_p$ — невязки чебышевского метода, получим последовательность векторов $\{\boldsymbol{r}_p, \boldsymbol{A}\boldsymbol{r}_p, \boldsymbol{A}^2\boldsymbol{r}_p, \boldsymbol{A}^3\boldsymbol{r}_p, ..., \boldsymbol{A}^k\boldsymbol{r}_p\}$. После можно произвести ортогонализацию с помощью метода Арнольди или Ланцоша (в симметричном случае) и получить ортогональную последовательность $\{\boldsymbol{\phi}_0, ..., \boldsymbol{\phi}_k\}$. Далее искать новое приближенное решение как

$$\hat{\boldsymbol{u}}_{p+k} = \boldsymbol{u}_{p+k} + \sum_{i=0}^{k} a_i \boldsymbol{\phi}_i, \tag{17}$$

удовлетворяющее например, критерию $\|\boldsymbol{b} - \boldsymbol{A}\hat{\boldsymbol{u}}_{p+k}\| \to \min$. Поскольку имеем из $\{\boldsymbol{r}_{p-k}, ..., \boldsymbol{r}_{p+k}\}$ сразу несколько семейств крыловских пространств, то последовательно проводим минимизацию по каждому.

9.3 Вариант 3

Воспользуемся соотношением (16) в форме

$$\boldsymbol{A}\boldsymbol{R}_{p-k,p} = \boldsymbol{R}_{p-k-1,p+1}\boldsymbol{T}, \tag{18}$$

где p — количество чебышевских итераций, k — размерность подпространства, в котором ищется решение, $\boldsymbol{R}_{p-k,p} = \{\boldsymbol{r}_{p-k}, \boldsymbol{r}_{p-k+1}, ..., \boldsymbol{r}_p\}$, а $\boldsymbol{T}$ — трехдиагональная матрица (16) со строками

$$-\beta_p/\alpha_p, 2\beta_p/\alpha_p, -\beta_p/\alpha_p.$$

Тогда, если рассмотреть $\hat{\boldsymbol{u}} = \boldsymbol{u}_p + \boldsymbol{R}_{p-k,p}\boldsymbol{y}$, то

$$\boldsymbol{b} - \boldsymbol{A}\hat{u} = \boldsymbol{r}_p - \boldsymbol{A}\boldsymbol{R}_{p-k,p}\boldsymbol{y} = \boldsymbol{r}_p - \boldsymbol{R}_{p-k-1,p+1}\boldsymbol{T}\boldsymbol{y} = \boldsymbol{R}_{p-k-1,p+1}(\boldsymbol{e}_{p+1} - \boldsymbol{T}\boldsymbol{y}). \tag{19}$$

Далее минимизируем данную невязку. К сожалению, векторы невязок в методе Чебышева не ортогональны и не получается избавится от $\boldsymbol{R}_{p-k-1,p+1}$ при переходе к задаче минимизации.

9.4 Результаты расчетов

На рисунках 4 ,5, 6 приведены результаты расчетов для различных степеней полиномов Чебышева, для трех рассмотренных вариантов методов. На

рисунке 7 изображены результаты расчетов по варианту 3 для различных размерностей подпространств. В данных экспериментах нижняя оценка границы спектра бралась на 2 порядка выше минимального собственного значения. Результаты, когда спектр заключен между оценками приведены на рисунке 8.

В итоге имеем, что для всех вариантов количество итераций получается меньшим такового для чебышевских методов, в конечном итоге скорость сходимости во всех вариантах становится равной таковой для стандартного метода. При известной нижней оценке минимального собственного значения имеем менее значительное сокращение количества итераций.

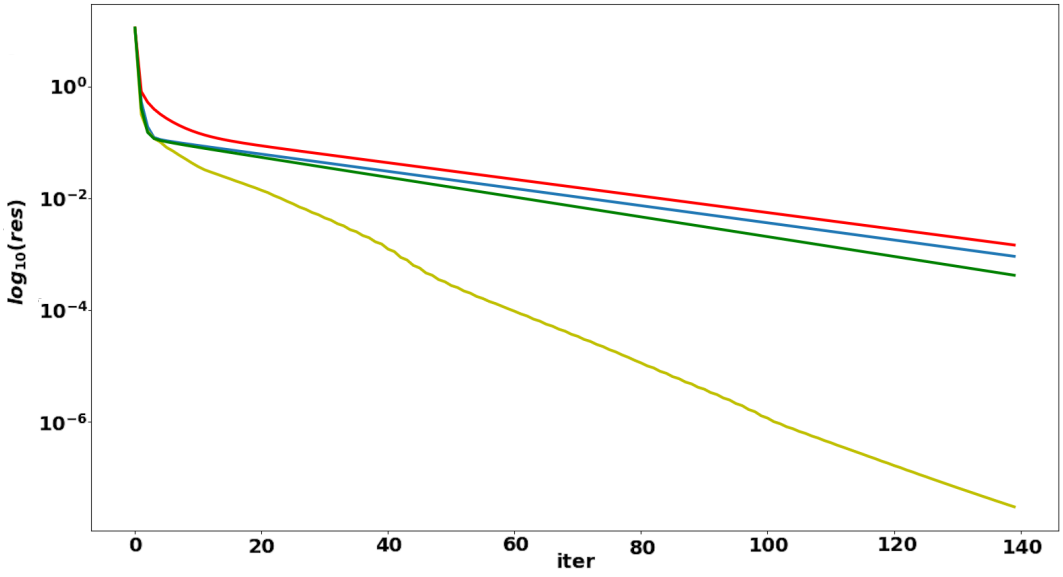

Рис. 4. Невязка решения в зависимости от количества итераций для полинома степени 250, красным цветом — стандартный метод, синим — метод с адаптацией к нижней части спектра, желтым — вариант 2, с $k = 10$, зеленым — вариант 3, с $k = 3$

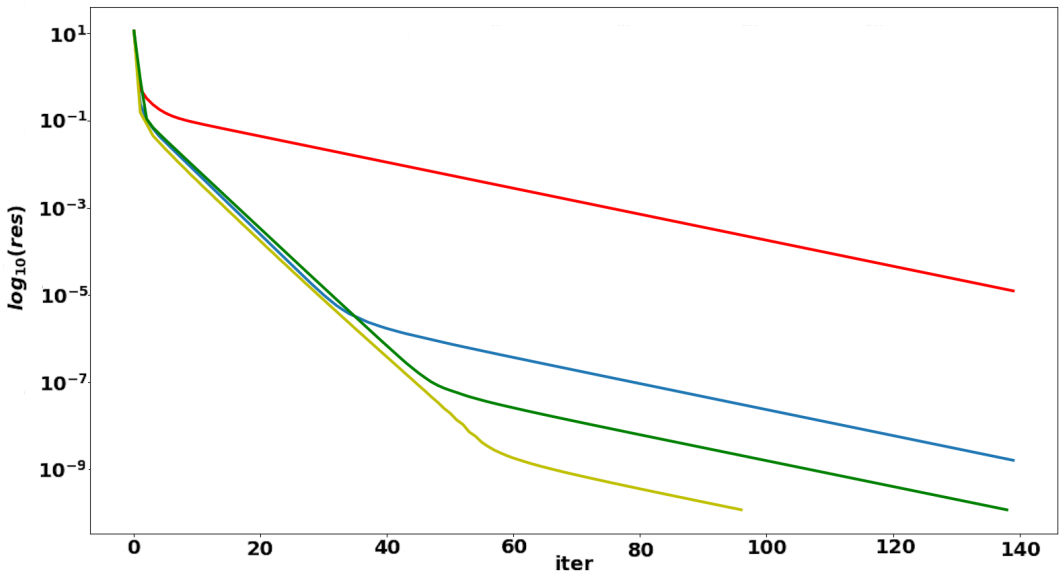

Рис. 5. Невязка решения в зависимости от количества итераций для полинома степени 500, красным цветом — стандартный метод, синим — метод с адаптацией к нижней части спектра, желтым — вариант 2, с $k = 10$, зеленым — вариант 3, с $k = 3$

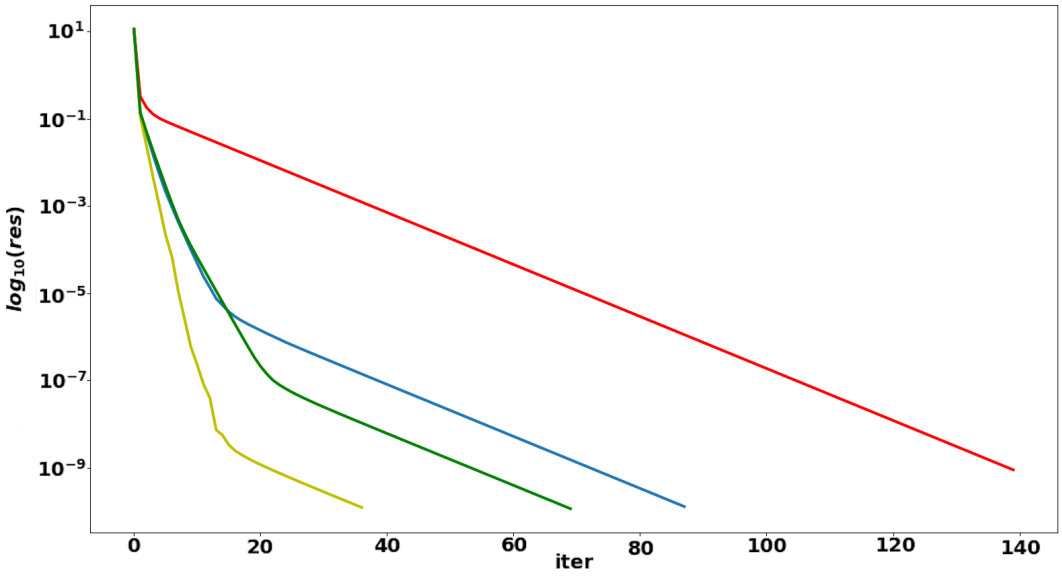

Рис. 6. Невязка решения в зависимости от количества итераций для полинома степени 1000, красным цветом — стандартный метод, синим — метод с адаптацией к нижней части спектра, желтым — вариант 2, с $k = 10$, зеленым — вариант 3, с $k = 3$

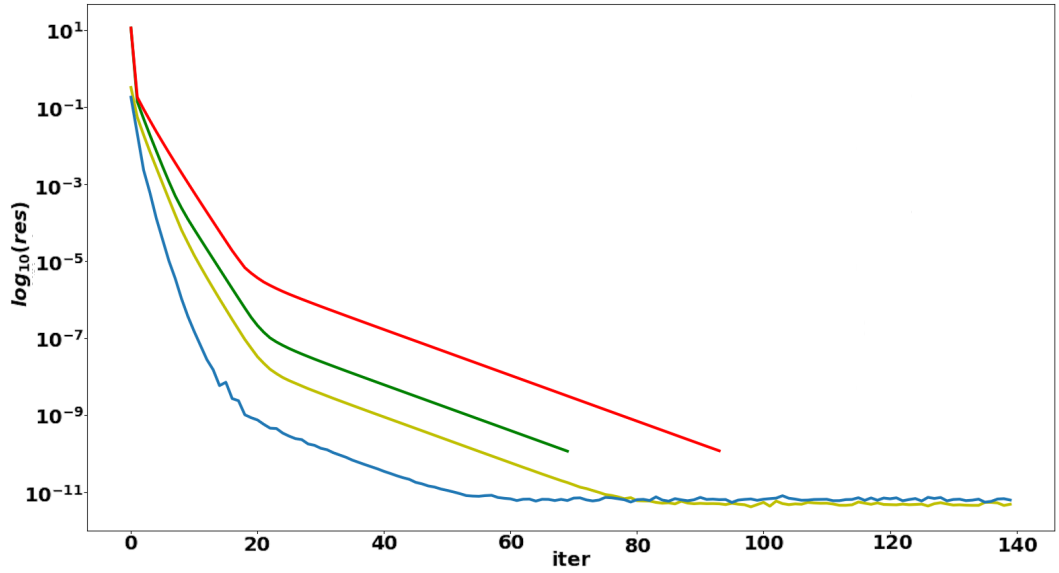

Рис. 7. Невязка решения в зависимости от количества итераций для полинома степени 1000 для варианта 3, красным цветом — $k = 1$, зеленым — $k = 3$, желтым — $k = 5$, синим — $k = 10$

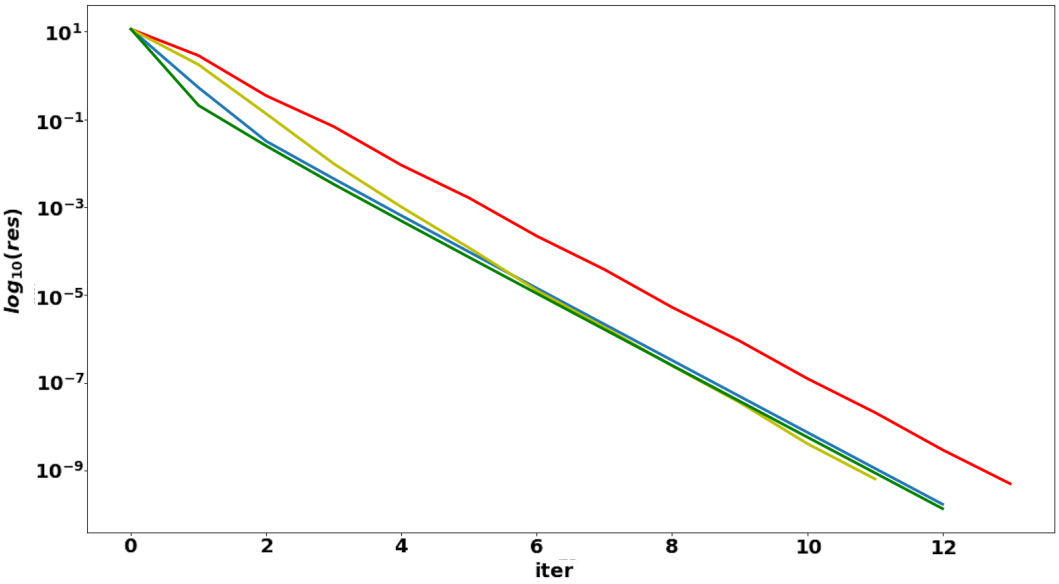

Рис. 8. Невязка решения в зависимости от количества итераций для полинома степени 1000 с точной оценкой границ спектра, красным цветом — стандартный метод, синим — метод с адаптацией к нижней части спектра, желтым — вариант 2, с $k = 10$, зеленым — вариант 3, с $k = 3$

10 Заключение

В работе рассмотрены различные типы итерационных методов Чебышева для решения сеточных аппроксимаций эллиптических уравнений на воксельных сетках. Предложена модификация алгоритма переупорядочивания корней полинома Чебышева, имеющая более простую форму и дающая результат аналогичный классическому. Рассмотрен ряд модификаций чебышевского итерационного метода, с уточнением нижней границы спектра. При использовании адаптированного варианта чебышевских итераций при решении неявных схем для гиперболических уравнений может быть получено значительное сокращение вычислительной работы. Метод адаптации к части спектра может быть использован при наличии существенно разделенного спектра у оператора, чего не наблюдается в наших примерах. Этот вопрос требует дальнейшего изучения. В заключение автор выражает благодарность В.Т.Жукову и Е.Б.Савенкову за неоценимую помощь при подготовке данной работы.

Список литературы

[1] В. Жуков, Н. Новикова, О. Феодоритова. “Чебышевские итерации с адаптивным уточнением нижней границы спектра матрицы”. *Препринты ИПМ*, Т. 172 (2018).

[2] А. Самарский, Е. Николаев. *Методы решения сеточных уравнений*. Наука, 1978.

[3] T. A. Manteuffel. “Numerische Mathematik The Tehebychev Iteration for Nonsymmetrie Linear Systems”. *Numer. Math*, Т. 28 (1977).

[4] H. E. Wrigley. “Accelerating the Jacobi Method for Solving Simultaneous Equations by Chebyshev Extrapolation When the Eigenvalues of the Iteration Matrix are Complex”. *The Computer Journal*, Т. 6 (2 1963). ISSN: 0010-4620. DOI: 10.1093/comjnl/6.2.169.

[5] G. Faber. “Über Tschebyscheffsche Polynome.” Т. 1920, № 150 (1920), с. 79—106. DOI: doi:10.1515/crll.1920.150.79. URL: https://doi.org/10.1515/crll.1920.150.79.

[6] T. A. Manteuffel. “Adaptive procedure for estimating parameters for the nonsymmetric Tchebychev iteration”. *Numerische Mathematik*, Т. 31 (2 1978). ISSN: 0029599X. DOI: 10.1007/BF01397475.

[7] K. C. Jea, D. M. Young. “ On the effectiveness of adaptive Chebyshev acceleration for solving systems of linear equations”. *Journal of Computational and Applied Mathematics*, Т. 24 (1-2 1988). ISSN: 03770427. DOI: 10.1016/0377-0427(88)90342-1.

[8] В. В. Воеводин. *Вычислительные основы линейной алгебры*. Наука, 1977.

[9] M. H. Gutknecht, S. Röllin. “The Chebyshev iteration revisited”. *Parallel Computing*, Т. 28 (2 2002). ISSN: 01678191. DOI: 10.1016/S0167-8191(01)00139-9.

[10] В. И. Лебедев. “Явные разностные схемы для решения жестких задач с комплексным или разделимым спектром”. *Ж. вычисл. матем. и матем. физ.*, Т. 40 (12 2000), с. 1801—1812.

Содержание